

Rethinking Learning with Noisy Labels: A Critical Review of Structured Supervision, Representation Dynamics, and Robust-Learning Pipelines

Wenxiao Fan

Beijing Institute of Technology

Kan Li

likan@bit.edu.cn

Beijing Institute of Technology

Research Article

Keywords: learning with noisy labels, noisy supervision, representation learning, memorization, robust learning, open-world learning

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10930944/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Rethinking Learning with Noisy Labels: A Critical Review of Structured Supervision, Representation Dynamics, and Robust-Learning Pipelines

Wenxiao Fan¹ and Kan Li^{1*}

^{1*}School of Computer Science, Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, Beijing, 100081, China.

*Corresponding author(s). E-mail(s): likan@bit.edu.cn;
Contributing authors: wenxiaofan@bit.edu.cn;

Abstract

Learning with noisy labels (LNL) is no longer well described as robust classification under random label flips. Modern learning systems draw supervision from crowdsourcing, web metadata, weak rules, cross-modal alignment, federated clients, large language models, and pre-trained representations. Their errors may be instance-dependent, semantically structured, open-set, or process-dependent. This critical narrative review organizes LNL around structured supervision mismatch, representation dynamics, and robust-learning pipelines. The central problem is not only to detect wrong labels. It is also to determine when reliability is observable, preserve useful feature geometry, refine uncertain targets, and evaluate robustness under realistic noise-generating processes. A focused diagnostic section examines the structural limits of small-loss selection and shows how optimization and representation statistics help explain when loss-based reliability signals are exposed or hidden. Across classical formulations, representative deep LNL methods, benchmark protocols, and recent 2023–2026 evidence from foundation-model and open-world settings, we connect noise assumptions with risk correction, identifiability, memorization, representation stability, pipeline design, and benchmark realism. Complementing prior reviews organized mainly around noise types, method families, or application domains, this review makes explicit how assumptions, training dynamics, algorithmic roles, and evaluation protocols constrain one another. We conclude by identifying cross-scenario pressure tests and open problems for noisy-supervision research in foundation-model and open-world ecosystems.

Keywords: learning with noisy labels, noisy supervision, representation learning, memorization, robust learning, open-world learning

1 Introduction

Reliable labels are a central assumption behind large-scale supervised learning, yet this assumption is increasingly difficult to satisfy in modern AI systems. Training labels may be produced by non-expert annotators, web crawlers, weak supervision rules, automatic cross-modal alignment, federated clients, outdated taxonomies, or large language models (LLMs). The resulting errors are rarely independent random mistakes. They often reflect semantic ambiguity, class similarity, sample difficulty, annotation bias, domain shift, and data-source heterogeneity. Learning with noisy labels (LNL) should therefore be studied as a broad noisy-supervision problem, not only as classification under random label flips.

This broader view matters because labels influence learning at more than one level. Statistically, they define the target distribution that a learner is expected to approximate. During optimization, they shape gradients, feature neighborhoods, class prototypes, pseudo-labels, and confidence estimates over many training iterations. A noisy label is therefore not only an incorrect entry in a dataset; it is a repeated supervisory signal that can change which features are learned and which samples are treated as reliable. This dual role explains why recent LNL research increasingly combines data diagnosis, training dynamics, and representation analysis.

The classical formulation of LNL assumes a clean pair (X, Y) and an observed noisy label $\tilde{Y}$ generated by a corruption mechanism, often summarized by a class-conditional transition matrix. This formulation has supported statistically grounded methods and remains useful for controlled analysis (Natarajan et al. 2013; Patrini et al. 2017; Zhang and Sabuncu 2018). However, a growing body of work shows that realistic noise is often instance-dependent, semantically structured, or open-set (Xia et al. 2020; Chen et al. 2021; Xia et al. 2023; Zhang et al. 2024). Once corruption depends on input features, unknown classes, or the annotation process, robustness can no longer be characterized only by estimating how one class flips into another.

This shift has also changed method design. Modern LNL methods rarely operate as a single robust loss or a one-shot correction module. Many successful approaches instead follow a staged process. They exploit early learning to isolate reliable signals, refine targets with confidence or temporal information, protect representation geometry through contrastive or structural constraints, and stabilize optimization through uncertainty, calibration, or dynamic schedules (Li et al. 2020, 2023; Zhang et al. 2023; Kim et al. 2024; Zhou et al. 2024). Recent work has therefore moved from asking only whether a label is wrong to asking which supervisory signals remain useful at each stage of training.

1.1 Positioning against prior surveys

Existing surveys have established the terminology, method families, and benchmark practices of LNL. Their scopes, however, are not interchangeable. Some focus on classical statistical formulations, some on deep robust-training methods or representation learning, and others on label cleaning, a particular application domain, or mislabeled-data detection. We therefore compare representative reviews by name in Table 1 and state the boundary of the present review relative to each one.

This review takes a different organizational stance. We use *problem formulation*, *training dynamics*, *pipeline roles*, *evaluation realism*, and *cross-domain pressure tests* as the main axes. This organization reflects changes in the literature. Open-set noise changes what counts as a valid training example, noisy pre-training changes what it means to inherit a representation, LLM-generated labels change the annotation process, and federated or graph settings change how noise propagates. A method list alone does not explain these dependencies.

The intended contribution is therefore not another exhaustive list of robust losses or sample selection rules. Instead, we ask what kind of supervision mismatch each setting creates, what evidence a learner can use before memorization occurs, how target refinement interacts with feature geometry, and whether the evaluation protocol actually tests the claimed robustness. This perspective lets us compare methods that look different architecturally but occupy similar roles in the learning process. Table 1 summarizes this complementary positioning.

The gap addressed here is therefore specific. We provide a critical narrative review that treats noisy labels as structured supervision mismatch, links observability to memorization and representation dynamics, maps representative methods to their assumptions and pipeline roles, and interprets realistic benchmarks as evidence about those assumptions. This framing is complementary to prior taxonomies rather than a claim of exhaustive coverage.

The review makes five contributions. First, it recasts noisy labels as structured supervision mismatch and clarifies the boundaries among class-conditional, instance-dependent, semantic, open-set, open-world, partial-label, and benign noise. Second, it connects statistical assumptions with training-time phenomena by synthesizing risk correction, identifiability, causal assumptions, memorization, and optimization dynamics. Third, it provides a focused diagnostic analysis of small-loss selection. The analysis uses optimization and representation evidence to show why loss-only reliability signals can collapse under structured semantic corruption. Fourth, the review organizes methods into a unified robust-learning pipeline rather than parallel method families. This organization makes explicit where sample selection, target refinement, representation learning, and uncertainty control depend on one another. Fifth, it uses benchmark realism and cross-scenario evidence as part of the main argument, highlighting where current methods still depend on closed-set or synthetic-noise assumptions.

1.2 Research questions and organization

We organize the discussion around five research questions rather than around a flat list of algorithms. These questions connect the proposed framework in Fig. 1 to the sections that follow.

2 Review Scope and Synthesis Protocol

This article is a critical narrative review. It synthesizes classical LNL studies, recent work from major artificial intelligence, machine learning, computer vision, natural language processing, data mining, and information retrieval venues, and domain-specific evidence from medical imaging, federated learning, graph learning, and multimodal

Table 1 Named comparison with representative prior surveys and adjacent reviews. The comparison concerns organizing emphasis rather than overall quality or comprehensiveness.

Prior review	Primary scope	Organizing emphasis	Boundary relative to this review
Frénay and Verleysen (Frenay and Verleysen 2014)	Classical classification with mislabeled training instances	Noise sources, consequences, noise-tolerant learning, label cleansing, and experimental design	Predates the current deep-learning emphasis on memorization, representation dynamics, and foundation-model supervision
Cordeiro and Carneiro (Cordeiro and Carneiro 2020)	Deep learning with noisy labels	Taxonomy of robust training strategies and benchmark datasets	Provides a method-family overview; the present review instead connects assumptions, training signals, pipeline dependencies, and evaluation realism
Han et al. (Han et al. 2020)	Label-noise representation learning	Historical and theoretical perspectives on learning representations under noise	Closest to our representation emphasis, but not organized around realistic noise provenance and cross-scenario pressure tests
Song et al. (Song et al. 2023)	Deep neural networks trained with noisy labels	Five robust-training groups, comparison properties, noise-rate estimation, datasets, and metrics	Offers broad method coverage; our focus is the dependency between formulation, dynamics, pipeline role, and evidence supplied by a benchmark
Shin et al. (Shin et al. 2024)	Label cleaning	Identification and correction of erroneous labels	Concentrates on data cleaning rather than the full robust-training and representation-learning pipeline
Shi et al. (Shi et al. 2024)	Medical image analysis	Domain-specific noise sources, methods, and clinical evaluation issues	Provides medical depth; our review uses medical evidence as one pressure test within a cross-domain framework
Song et al. (Song et al. 2025)	Deep LNL and application settings	Method taxonomy and application-oriented overview	Extends application coverage; our synthesis centers on reliability signals, stage interactions, and assumption failures
Nazaretyan et al. (Nazaretyan et al. 2025)	Identification of mislabeled tabular data	Empirical comparison of label-error filters and data-dependent recommendations	Benchmarks data-quality detection rather than end-to-end deep LNL, representation dynamics, or target refinement

learning. The literature search was updated through August 2026 using prior surveys, proceedings and publisher pages for the principal venues represented in the review, targeted journal searches, and backward and forward citation tracing. We prioritize papers that introduce a distinct noise formulation, explain a training or representation mechanism, define an evaluation protocol, or substantially change how noisy supervision is used in modern learning systems.

Table 2 Research questions and manuscript organization.

ID	Research question	Main sections
RQ1	What forms of noisy supervision are not captured by classical random-flip or class-conditional formulations?	Sections 3 and 6
RQ2	When do memorization and representation dynamics make label reliability observable, and when do they hide it?	Sections 4, 5, and 6
RQ3	How do robust-learning methods occupy pipeline roles, and what assumptions support each role?	Section 7
RQ4	Which benchmark protocols and metrics actually test the robustness claims made by LNL methods?	Section 8
RQ5	How do foundation-model, open-world, graph, federated, and preference-label settings pressure-test LNL assumptions?	Sections 9 and 10

The search combined broad terms such as “learning with noisy labels,” “label noise,” and “noisy supervision” with targeted concepts including class-conditional and instance-dependent noise, open-set and open-world noise, memorization, representation learning, calibration, federated and graph learning, foundation models, and LLM-generated labels. Foundational papers were traced through earlier surveys and the reference lists of representative methods; recent work was checked against its formal venue or publisher record when available. Papers concerned only with generic weak supervision, domain adaptation, or data cleaning were treated as adjacent unless they posed an explicit label-reliability question. Literature was added until further papers no longer changed the mismatch taxonomy, the observable-signal account, the five pipeline roles, or the evaluation boundaries developed in the review. This stopping rule supports a focused synthesis, not a claim of exhaustive retrieval.

The review follows three inclusion principles. First, a paper is treated as central when it clarifies an assumption that many later methods depend on, such as class-conditional noise, instance-dependent noise, open-set contamination, identifiability, or early memorization. Second, a recent method is discussed in the main text when it contributes a reusable pipeline role, including reliability estimation, target refinement, representation preservation, calibration, or optimization control. Third, an application paper is included when it exposes a structural mismatch that closed-set image-classification benchmarks cannot represent, such as client-level noise in federated learning, topological propagation in graphs, localization noise in detection, or prompt-dependent annotation noise from LLMs.

This scope deliberately avoids an exhaustive method catalogue. Instead, it aims to give AI researchers and practitioners an assumption-aware map of the field. The tables are therefore used as organizing devices rather than complete bibliographies. They identify representative mechanisms, the supervision mismatch each mechanism

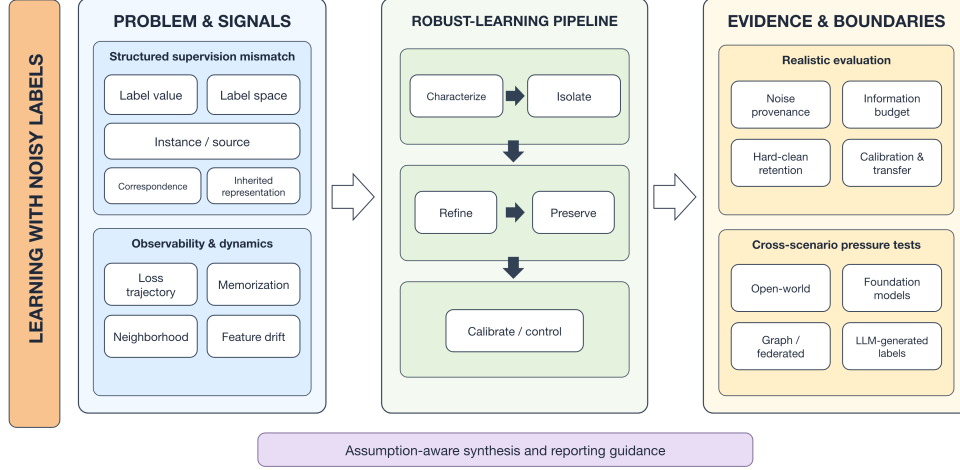


Fig. 1 Taxonomic landscape of this critical narrative review. The framework links structured supervision mismatch and observable training signals to five robust-learning pipeline roles, then interprets realistic evaluation and cross-scenario settings as evidence about method assumptions and boundaries.

addresses, and the evaluation risks that should be checked before a method is transferred to a new domain. Because some 2026 papers remain close to publication or preprint release, we treat them as directional evidence of emerging pressure points rather than as settled consensus. ArXiv-only entries are explicitly separated from formally published work.

Table 3 clarifies how different literature streams are used in the synthesis. The table describes the review’s scope and selection logic; it does not imply exhaustive retrieval or formal systematic-review coverage.

3 From Wrong Labels to Structured Supervision Mismatch

3.1 Basic formulation

Let $\mathcal{X}$ denote an input space and $\mathcal{Y} = \{1, \dots, C\}$ a closed label space. In the clean setting, training examples are drawn from $P(X, Y)$. In LNL, the learner observes $(X, \tilde{Y})$, where $\tilde{Y}$ is produced by a noise process. The classical class-conditional model assumes

$$T_{ij} = P(\tilde{Y} = j \mid Y = i), \quad (1)$$

where T does not depend on X . Symmetric and asymmetric synthetic protocols instantiate this assumption by fixing the probability of random or semantically biased class flips. These protocols are analytically useful because they isolate a controlled corruption mechanism and make results comparable across methods.

Table 3 Scope and synthesis protocol for the critical narrative review.

Literature stream	Selection emphasis	Role in the synthesis	Boundary
Foundational and theoretical work	Risk correction, transition matrices, identifiability, causal assumptions, and memorization	Establish the assumptions and information limits reused by later methods	Theory without a clear connection to noisy supervision is outside the main scope
Representative deep-LNL methods	Methods that introduce a reusable reliability signal or pipeline operation	Compare assumptions, observables, target refinement, representation preservation, and optimization control	Incremental variants are represented selectively rather than exhaustively
Benchmarks and evaluation studies	Datasets or protocols that alter the source, structure, or observability of noise	Connect empirical evidence to the noise model and claimed capability	Generic accuracy testbeds are not included solely for completeness
Cross-domain and emerging settings	Medical, graph, federated, multimodal, pre-training, preference, and LLM-generated labels	Pressure-test whether the framework survives changes in modality, label space, and unit of noise	Emerging work is treated as directional evidence when consensus is not established

However, the same formulation becomes restrictive when the corruption depends on the input:

$$T_{ij}(x) = P(\tilde{Y} = j \mid Y = i, X = x). \quad (2)$$

Instance-dependent noise is harder because reliability is tied to sample geometry, class boundaries, visual ambiguity, or annotator behavior (Xia et al. 2020; Chen et al. 2021). The transition process itself may become difficult to identify; in generic instance-dependent settings, additional noisy labels or structural assumptions may be necessary (Liu et al. 2023). Moving beyond class-conditional noise is therefore not merely an empirical complication. It changes what can be estimated and which assumptions a method must rely on.

This distinction also separates two forms of uncertainty that are often conflated. Aleatoric ambiguity arises when the input admits multiple plausible semantic interpretations, such as visually similar fine-grained classes or medical images with expert disagreement. Epistemic uncertainty arises because the learner has limited information about the noise process, the annotator, or the existence of unseen classes. A

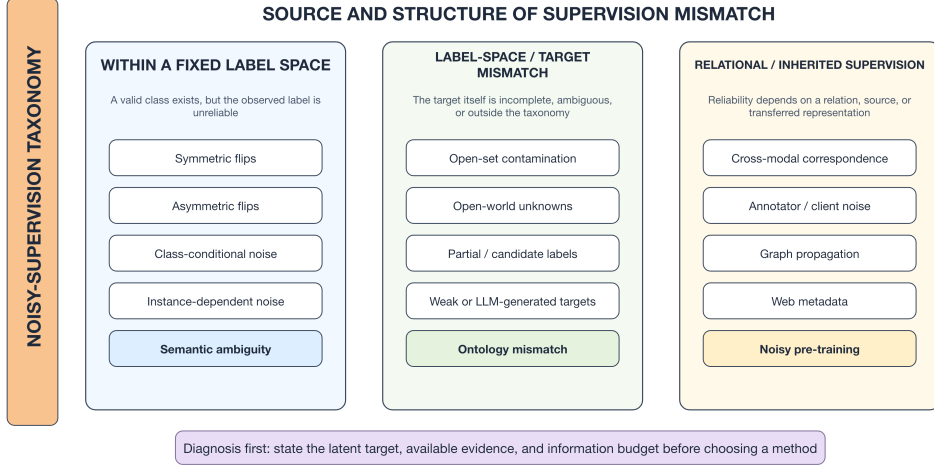


Fig. 2 Taxonomy of structured noisy supervision. The three groups distinguish corruption within a fixed label space, mismatch in the label space or target construction, and supervision whose reliability depends on correspondences, clients, graph structure, or pre-training. The groups are not mutually exclusive; practical datasets may combine several mechanisms.

transition-matrix method, a small-loss selector, and an open-world detector make different assumptions about which uncertainty is dominant. Making these assumptions explicit is essential for fair comparison.

3.2 A taxonomy for structured noisy supervision

Modern LNL problems are better described by the source and structure of supervision mismatch. Figure 2 separates errors within a fixed label space from label-space or target mismatch and from relational or inherited supervision. The central distinction concerns where the mismatch occurs. The observed label may be wrong within a fixed label space, the example may not belong to that space, or the target may be incomplete, ambiguous, process-generated, relationally propagated, or inherited from pre-training.

Instance-dependent and semantic noise are especially important because they connect label reliability to representation learning. Boundary samples, long-tailed classes, fine-grained categories, and visually ambiguous images are more likely to be mislabeled. Recent work on fine-grained and long-tailed LNL therefore does not merely add imbalance to a standard noise protocol; it changes the relation between sample difficulty, representation geometry, and label reliability (Wei et al. 2023b; Lu et al. 2023).

Open-set and open-world noise further challenge the closed-label-space assumption. In open-set LNL, some examples are assigned target labels even though their true categories lie outside the target taxonomy. In open-world noisy data, unknown classes may appear during both training and evaluation, making it necessary to distinguish

clean in-distribution examples, closed-set noisy examples, and out-of-distribution contaminants (Xia et al. 2023; Pan et al. 2025; Zhang et al. 2025b). The practical difficulty is that unknown examples can resemble hard clean examples during early training, while hard clean examples can be mistaken for outliers.

The taxonomy should not be read as a set of mutually exclusive bins. A web-scale dataset may contain class-conditional mistakes, visually ambiguous instances, out-of-taxonomy examples, and machine-generated weak labels at the same time. The value of the taxonomy is diagnostic: it asks which part of the supervision channel is unreliable and what evidence is available to detect that unreliability. This is also where the present review differs from method-family taxonomies. The same algorithmic component, such as pseudo-labeling or contrastive regularization, can have different meanings depending on whether the main failure mode is instance-dependent confusion, open-set contamination, or partial supervision.

A useful diagnostic distinction is between *annotation error*, *semantic ambiguity*, *ontology mismatch*, and *out-of-label-space contamination*. Annotation error assumes that a valid clean class exists and the observed label is simply wrong. Semantic ambiguity means that several labels may be defensible because the sample lies near a class boundary or reflects expert disagreement. Ontology mismatch means that the available label set is too coarse, outdated, or inconsistent with the data-generating process. Out-of-label-space contamination means that the sample should not be forced into any target class. These cases require different responses: correction is natural for annotation error, calibrated uncertainty is more appropriate for ambiguity, taxonomy revision may be needed for ontology mismatch, and rejection or open-world discovery is needed for out-of-label-space samples.

3.3 Structured supervision mismatch

We use the term *structured supervision mismatch* to describe this broader setting. The mismatch may lie between an instance and its observed label, between a dataset and the assumed closed label space, between weak labels and semantic targets, or between a pre-trained representation and downstream annotations. The objective is therefore not always to recover a hidden clean label. In some settings, a robust learner should preserve task-useful representations, recover calibrated uncertainty, or exploit partial semantic information even when the observed label remains imperfect.

This formulation also clarifies the role of classical synthetic noise. Synthetic corruption remains useful as a mechanism testbed because it offers control and reproducibility. Its limitation is external validity: success under fixed random flips does not necessarily imply robustness to instance-dependent, open-world, or annotator-dependent noise. The next section explains why this gap appears at the level of training dynamics and representation learning.

The mismatch view further changes how we interpret failure. A method may select the wrong samples or rely on a representation that is already biased. Its target-refinement step may also turn uncertain labels into overconfident pseudo-labels, or its benchmark may hide out-of-distribution contamination inside a closed-set metric. These failure modes are difficult to see when noisy labels are treated only as corrupted

symbols. They become clearer when supervision, representation, and evaluation are analyzed together.

4 Representation and Memorization Dynamics under Label Noise

The previous section framed noisy labels as structured supervision mismatch. This section explains why that mismatch cannot be understood only from the dataset snapshot. During training, the model repeatedly converts labels into gradients, gradients into features, features into confidence estimates, and confidence estimates back into sample weights or pseudo-labels. Label noise is therefore a dynamic influence on representation formation, not merely a static annotation error. This view matters because many LNL methods work only when training exposes a temporary reliability signal before the model memorizes corrupted supervision or distorts the feature space used to detect it.

4.1 Memorization as a training-time opportunity

Deep networks can fit random labels, but they do not typically memorize all examples at the same rate. Empirical studies on memorization show that networks tend to learn simple, shared, and repeatedly supported patterns before fitting irregular or corrupted labels (Arpit et al. 2017). This empirical asymmetry motivates small-loss selection, co-training, and early-learning methods (Han et al. 2018; Li et al. 2020). The key insight is not simply that memorization is harmful. It is that the period before full memorization can make label reliability partially observable.

The early-learning window should therefore be interpreted as an observability condition rather than as a universal law. It is useful when clean examples tend to be easier to fit than corrupted examples, but this ordering can break under class imbalance, fine-grained ambiguity, open-set contamination, or instance-dependent noise. Tail-class samples may remain high-loss even when correctly labeled, whereas mislabeled but visually typical samples can quickly become low-loss. Boundary samples and ambiguous examples may also look unreliable not because their labels are wrong, but because the current representation has not yet separated the relevant semantics. This creates a central tension in LNL: early training provides the best opportunity for noise diagnosis, yet the diagnostic signals available at that time are already biased by difficulty, class frequency, and representation quality.

This tension explains why a single global small-loss threshold is often too crude. The same loss value can have different meanings across classes, epochs, and local neighborhoods. A high loss early in training may indicate a corrupted label, but it may also indicate a rare class, a hard clean example, or an example whose relevant features are not yet learned. A low loss late in training may indicate a clean easy example, but it may also indicate that the network has already memorized the observed label. Robust methods therefore need to ask not only whether an example is currently easy, but also how its reliability signal changes over time and relative to comparable examples.

Recent methods increasingly exploit this window in more fine-grained ways. DISC models instance-specific memorization trajectories rather than using a single global

threshold (Li et al. 2023). Logit clipping controls the effective loss bound at the logit level to mitigate later-stage memorization (Wei et al. 2023a). Early Cutting and instance-level dynamic optimization further show that easy-looking mislabeled examples and hard clean examples require different training responses (Yuan et al. 2025; Zhang et al. 2025a). Taken together, these works shift the role of memorization from a generic motivation for sample selection to a concrete design constraint: reliable-signal isolation should be time-aware, difficulty-aware, and class-aware.

4.2 From scalar loss to representation geometry

Loss values provide an accessible reliability signal, but they are a one-dimensional projection of a richer training state. A loss value depends jointly on the observed label, the classifier head, the current feature extractor, the class prior, and the optimization schedule. Low loss can indicate a clean easy example, a memorized mislabeled example, or a biased classifier decision. Conversely, high loss can correspond to a corrupted label, a hard but informative clean example, a tail sample, or an out-of-distribution input. This ambiguity motivates a move from scalar ranking toward representation geometry.

The geometric view asks whether an example is consistent with its neighbors, prototypes, ranks, or external semantic priors. Methods such as RankMatch, CSOT, and structural-label learning use confidence, feature ranks, optimal transport, reverse nearest neighbors, or local distribution structure to identify reliable supervision (Zhang et al. 2023; Chang et al. 2023; Kim et al. 2024). These signals are not just additional heuristics. They reflect a different assumption about what remains trustworthy when the observed label is unreliable: not the label itself, but the relational structure among samples and classes.

This view also explains why label noise damages models beyond the classifier head. Corrupted labels can distort local neighborhoods, pull features toward the wrong class prototype, collapse semantically related classes, or create shortcuts that are aligned with annotation artifacts rather than task semantics. Once the representation space is contaminated, later label correction may reinforce the error rather than repair it: the model uses a biased feature space to select reliable examples, then uses the selected examples to further shape the feature space. This feedback loop is one reason contrastive learning, neighborhood consistency, structural labels, and representation-classifier decoupling have become prominent in recent LNL research (Huang et al. 2023; Wei et al. 2023b; Kim et al. 2024).

Representation geometry provides a language for comparing methods that otherwise appear unrelated. A contrastive objective attempts to preserve neighborhood structure; a structural-label method builds alternative targets from reverse nearest-neighbor relations; a vision-language-model detector imports semantic alignment from a pre-trained model; and an optimal transport filter imposes global matching constraints. All of them reduce reliance on the raw observed label by asking whether an example is consistent with a broader geometry. The practical challenge is that this geometry is not fixed: it evolves during training and can itself be corrupted by the labels it is meant to diagnose.

Table 4 Training-dynamics diagnostics for noisy supervision.

Diagnostic	What it can reveal	Typical false signal	Useful reporting target
Loss trajectory	Whether an example becomes easy before or after memorization	Tail, boundary, or rare clean samples may remain high-loss	Class-aware clean/noisy separability over epochs
Forgetting or transition events	Whether predictions are unstable under repeated training updates	Ambiguous examples may be unstable without being mislabeled	Frequency and timing of clean-to-noisy role changes
Neighborhood consistency	Whether local feature geometry supports the observed label	Biased representations can make wrong neighborhoods look coherent	kNN purity, reverse-neighbor support, and hard-clean retention
Prototype or centroid drift	Whether class anchors move toward corrupted supervision	Long-tailed classes can drift because of sparse support, not only noise	Per-class prototype stability and tail-class behavior
Gradient agreement	Whether samples reinforce or conflict with shared optimization directions	Easy mislabeled examples can agree with dominant shortcuts	Agreement by class, difficulty, and suspected-noise group
Calibration or entropy drift	Whether confidence tracks uncertainty rather than memorized labels	Overconfident pseudo-labels can mimic reliable targets	Expected calibration error, coverage, and uncertainty on suspected noise

The implication for review organization is that representation should not be treated as one method family among others. It is the substrate on which many other components depend. Sample selection uses features to decide which examples are reliable; target refinement uses predictions or neighborhoods derived from the representation; calibration depends on whether uncertainty reflects semantic ambiguity or memorized noise; and open-set detection depends on whether unknown samples occupy separable regions. A method can therefore fail even when its label-correction rule is reasonable, simply because the representation used to support that rule has already absorbed the wrong supervision.

Table 4 turns this discussion into a set of diagnostics. The point is not that every paper must report all of them, but that each diagnostic answers a different question about when label reliability is observable and when it is being hidden by optimization, representation drift, or class structure.

4.3 Benign noise and noisy pre-training

Not all deviations from clean labels are equally harmful. Some noise can act as a form of regularization or can be absorbed by overparameterized models under specific regimes. Recent work on noisy pre-training provides a concrete example: slight label noise during pre-training can improve in-domain transfer while hurting out-of-domain generalization, suggesting that label noise reshapes the feature space in a distribution-dependent manner (Chen et al. 2024). Related observations on flatness-aware optimization and noisy supervision also suggest that the effect of noise depends on how optimization translates label perturbations into representation changes (Baek et al. 2024). These findings caution against treating clean-label recovery as the only objective of LNL.

The benign-noise perspective should be used carefully. Noise is not intrinsically benign; it becomes benign only relative to a data distribution, model class, optimization regime, and evaluation target. A perturbation that improves local invariance or in-domain transfer may still damage minority classes, calibration, or out-of-domain generalization. Similarly, a pre-trained representation may appear robust on a downstream benchmark because it has absorbed useful invariances, but it may also carry hidden taxonomy mismatches or annotation biases from the pre-training corpus. The important question is therefore not whether noise is good or bad in the abstract, but which aspects of the representation it preserves, distorts, or transfers.

This perspective also changes how LNL methods should be evaluated. If noise affects the feature extractor, then final accuracy is not enough. A robust method should be assessed by representation stability, neighborhood consistency, class-wise behavior, calibration, clean/noisy separability over time, and transfer under distribution shift. These diagnostics can reveal whether a method truly preserves useful structure or simply suppresses difficult examples. They also help distinguish harmful memorization from useful adaptation, a distinction that becomes increasingly important in foundation-model and open-world settings.

This section motivates a closer look at the field’s most influential reliability proxy. If noisy labels affect models through time-varying loss dynamics, feature geometry, and representation transfer, then small-loss selection should not be treated as a universal clean-sample detector. It should be treated as an observability proxy whose reliability depends on whether training makes clean and noisy samples separable in the first place.

5 Diagnostic Case Study: Structural Limits of Small-Loss Selection

Small-loss selection is a useful lens because it exposes the central tension in noisy-label learning. Many successful LNL methods assume that clean samples become easier to fit before noisy samples, so low-loss examples can be selected, reweighted, or used as anchors for semi-supervised training. This assumption is practical and often effective, but it is not a law of label correctness. It is a statement about observability: a loss-based selector can work only when the training process creates a sufficiently separable signal between clean and corrupted supervision.

This section is a diagnostic case study, not a proposal for a new LNL algorithm or a ranked comparison among methods. Its role is to expose one assumption that recurs throughout the literature: loss-based reliability is informative only when optimization and representation geometry make clean and corrupted supervision distinguishable. We therefore interpret the controlled analyses below as mechanism probes and avoid using them to claim universal superiority of one corruption model, learner, or diagnostic statistic.

5.1 Small loss as an observability proxy

The small-loss criterion should be interpreted as a proxy for early structural learnability. A clean sample often has a lower early loss because its visual or semantic pattern is supported by nearby samples and by the class decision boundary being learned. A corrupted sample often has a higher early loss because its observed target conflicts with those patterns. However, the loss value itself does not reveal which of these causes is present. A high-loss example can be a mislabeled sample, a tail-class example, a boundary example, or a sample whose useful features have not yet been learned. A low-loss example can be a genuinely clean easy sample, but it can also be an easy mislabeled sample or a memorized corrupted target.

This interpretation connects small-loss selection to the broader mismatch view. Once noise is semantic, instance-dependent, or class-imbalanced, the ordering between clean and noisy losses can be distorted by sample difficulty and representation quality. Loss is therefore not a direct observation of label correctness. It is a one-dimensional projection of the current model state, and that state is jointly determined by the label, the feature extractor, the classifier, class frequency, and the optimization schedule. The question is not simply whether small-loss works, but which structural conditions make it informative.

5.2 Optimization and representation diagnostics

A diagnostic way to inspect these conditions is to measure how different groups of samples affect optimization and representation geometry. We use two simple statistics for this purpose. Let θ_t be the vector of all trainable parameters at epoch t . Parameter update magnitude is operationalized as $\text{PUM}_t = \|\theta_t - \theta_{t-1}\|_2$, so smaller values indicate less change in the model parameters. Following the feature-spectrum diagnostic used in noisy pre-training analysis (Chen et al. 2024), the largest singular value ratio is $\text{LSVR}_t = -\log[\sigma_1(Z_t)/\sum_i \sigma_i(Z_t)]$. Here, $Z_t \in \mathbb{R}^{N \times 512}$ contains the 512-dimensional penultimate representations extracted after global average pooling and before the final linear classifier of ResNet-18. LSVR is therefore a transformed spectrum-share diagnostic rather than a probability. Smaller values indicate stronger dominance by the leading singular direction, whereas larger values indicate a less concentrated spectrum.

The case study uses CIFAR-100 for the main diagnostics and fixes the nominal corruption rate at 40% for the separation summary in Table 5. Symmetric, instance-dependent, and SILN corruption provide three controlled structures; cross-entropy and RoLR supply contrasting training objectives; the loss distributions are read at epoch 50. Supplementary Appendix A isolates the feature-space intervention on CIFAR-10

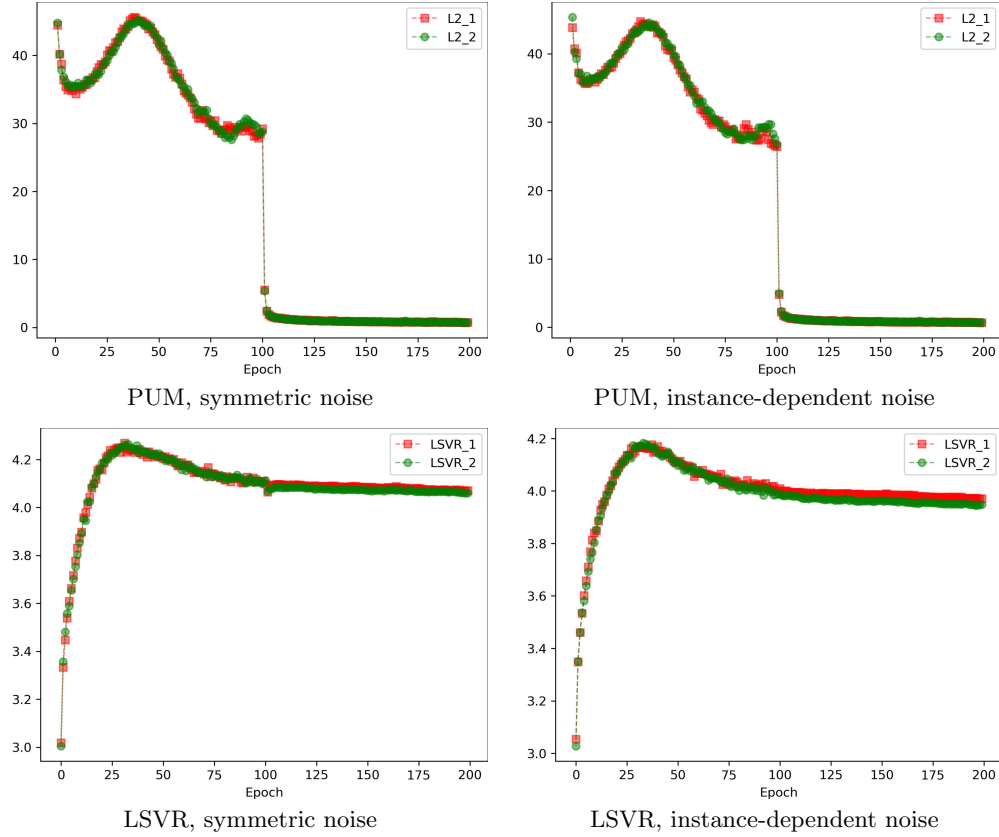


Fig. 3 Optimization and representation diagnostics for small-loss analysis on CIFAR-100. PUM tracks parameter update magnitude, while LSVR tracks feature-spectrum concentration. Reporting both symmetric and instance-dependent corruption makes the diagnostic point more explicit: loss separability is tied to optimization and representation structure rather than label correctness alone.

and reports confidence- and architecture-transfer checks. These panels are interpreted only as within-protocol mechanism probes. They are not pooled with results from different implementations or training budgets, and they do not support claims of cross-run statistical significance.

The diagnostic implementation uses ResNet-18 with random seeds 0 and 42. Its optimization and input processing follow the public RoLR implementation on which the experiments were built (Chen et al. 2023). SGD uses an initial learning rate of 0.03, momentum 0.9, weight decay 5×10^{-4} , and an effective main-training batch size of 448. Training starts with 15 warm-up epochs and continues to epoch 500; the learning rate is reduced by a factor of ten during the final 100 epochs. The basic CIFAR-100 transform applies a 32×32 random crop with four-pixel padding, random horizontal flipping, tensor conversion, and channel-wise normalization with mean (0.507, 0.487, 0.441) and standard deviation (0.267, 0.256, 0.276). The strong view additionally applies the implementation’s modified RandAugment with three sampled operations and Cutout. PUM is computed over all trainable parameters.

Figure 3 illustrates the diagnostic role of these statistics. The important message is not that PUM or LSVR should replace loss in every algorithm. Rather, they show why small-loss can emerge: clean samples are often supported by coherent structure, so the model can organize them with smoother updates and more stable representations. This shifts the explanation from “clean labels have small loss” to a more cautious statement: clean samples tend to become small-loss when they are structurally easier for the current model to fit.

Supplementary Fig. 7 isolates the same point with a minimal intervention. The toy setting is not meant to define a realistic annotation model. It asks a narrower question: if noisy samples become structurally easy to fit, does the loss ranking still separate them from clean samples? The observed overlap suggests that small-loss is fragile whenever the feature space makes corrupted targets look coherent.

5.3 Structured noise can collapse loss separability

The same structural interpretation also explains a failure mode. If corrupted labels are attached to samples in a semantically coherent way, then noisy samples may no longer look irregular to the model. They can become easy to fit, occupy plausible neighborhoods, and overlap with clean samples in the loss distribution. This is precisely the regime in which a loss-only selector has little information left to use.

To make this failure mode visible, we use Structural-Induced Latent Noise (SILN) as a controlled stress test rather than as a universal model of real annotation behavior. SILN constructs semantically structured corruptions by using superclass-level ambiguity and model-predicted proximity to choose plausible wrong labels. The purpose is diagnostic: it intentionally narrows the structural gap that conventional random or class-conditional synthetic noise often leaves too wide. A useful stress test need not describe every real annotation process; it should expose an assumption that would otherwise remain hidden.

Supplementary Fig. 8 shows why SILN should be read as a structural stress test. Symmetric noise spreads corruptions broadly, while SILN concentrates transitions around more structured alternatives. This difference matters because a corruption process that preserves semantic plausibility can be easier for a model to absorb than a random flip, even when the nominal noise rate is lower.

Figure 4 shows the consequence for clean/noisy separability. Under conventional synthetic corruption, clean and noisy samples can remain visibly separated in loss space. Under SILN, the distributions overlap much more heavily. This does not mean that SILN is the only hard noise process, nor that all semantic noise behaves identically. It means that structured corruptions can remove the very gap that small-loss methods implicitly exploit.

5.4 Failure of loss-only splitting

The overlap view can be made explicit through a thresholding argument. Suppose clean and noisy losses follow equal-variance Gaussian distributions with equal priors, means separated by Δ , and common standard deviation σ_0 . The Bayes-optimal threshold then lies midway between the means, with error $P_e = \Phi[-\Delta/(2\sigma_0)]$, where Φ is the

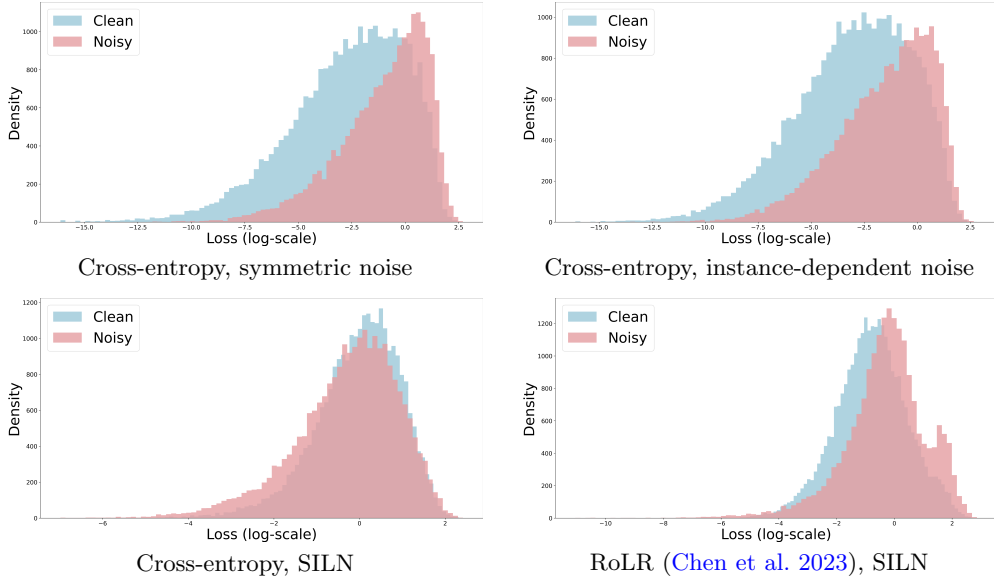


Fig. 4 Clean/noisy loss distributions on CIFAR-100 at epoch 50 under different corruption processes and learners. Compared with conventional synthetic noise, SILN produces stronger overlap between clean and noisy samples, including under a competitive small-loss learner.

Table 5 Normalized separation Δ/σ_0 between clean and noisy losses on CIFAR-100 at 40% noise. Lower values indicate stronger overlap and a harder clean/noisy splitting problem for loss-based selectors.

Training method	Symmetric noise	Instance-dependent noise	SILN
Cross-entropy	3.78	3.32	2.56
RoLR	5.12	4.98	2.74

standard normal cumulative distribution function. The normalized separation Δ/σ_0 therefore controls the best error achievable from loss alone. When this value is small, even an optimal threshold has a non-negligible clean/noisy classification error. The point is mechanism-agnostic: any corruption process that collapses the loss gap can impose an error floor on loss-only sample selection.

Table 5 gives a compact example of this effect. The normalized separation is much smaller under SILN than under symmetric or instance-dependent corruption, even for RoLR (Chen et al. 2023), a learner designed around robust loss behavior. In this regime, a two-component loss model such as the Gaussian mixture split used by many small-loss pipelines can mistakenly assign hard clean samples to the noisy component and easy corrupted samples to the clean component. Tuning the threshold may shift the trade-off, but it cannot create information that is absent from the loss distribution.

Supplementary Fig. 9 directly evaluates the splitter used by many small-loss pipelines. The clean-sample panel is especially important for evaluation: a method

Table 6 Diagnostic evidence from the small-loss analysis and its role in the review argument.

Evidence	What it shows	Review implication
PUM and LSVR diagnostics	Early loss separability is linked to functional perturbation and feature-spectrum structure	Small-loss should be treated as a representation-dependent proxy, not a direct label-correctness oracle
Toy structural intervention	Making noisy samples structurally easy increases clean/noisy loss overlap	The proxy can fail when noisy targets are coherent rather than random
SILN transition matrices	Structured semantic corruptions differ from uniform random flips	Benchmark design should test corruption structure, not only noise rate
Loss distribution and Δ/σ_0	SILN reduces normalized clean/noisy loss separation under CE and RoLR	Loss-only selectors face an information bottleneck when distributions overlap
GMM splitter accuracy	Clean samples in the overlap region are often assigned to the noisy component	Evaluations should report hard-clean retention and clean/noisy detection behavior
Confidence and architecture checks	The overlap persists beyond one posterior model or one training architecture	Stress tests should check proxy robustness across diagnostics and model choices

can appear aggressive against noisy labels while discarding hard clean samples that should remain useful. This is why hard-clean retention should be reported alongside noisy-sample detection.

Supplementary Fig. 10 adds two sanity checks. First, the difficulty is not limited to a GMM posterior over losses; confidence-based filtering also faces overlap. Second, the overlap persists when the model used for downstream training differs from the model used to construct the noisy labels. These checks do not make SILN universally architecture-independent, but they strengthen its role as a diagnostic stress test.

5.5 Implications for taxonomy and evaluation

The expanded evidence in Table 6 clarifies why this review treats representation dynamics and benchmark realism as core axes. Loss-based reliability is not wrong; it is conditional. It works best when clean examples are structurally easier than corrupted examples, when class imbalance does not make hard clean samples look noisy, and when the benchmark preserves a meaningful gap between the two groups. Once these conditions fail, robust learning requires additional evidence from temporal trajectories, neighborhoods, prototypes, calibration, external semantic priors, or task-specific structure.

This conclusion changes how LNL papers should be compared. A method that improves final accuracy under symmetric noise may have learned to exploit an unusually favorable loss gap rather than a generally transferable noise model. Conversely, a method that reports clean/noisy detection precision, hard-clean retention, class-aware separability, neighborhood consistency, and representation stability gives stronger evidence about why it works. The small-loss analysis therefore acts as a bridge between memorization dynamics and the method/evaluation sections that follow: it explains why reliable-signal isolation should be reported as a measurable condition, not assumed as an automatic consequence of early learning.

The next section turns this diagnostic lesson into a broader theoretical question. If a reliability signal is only conditionally observable, then LNL methods must state which assumptions make correction possible, when clean/noisy separation is identifiable, how uncertainty should be encoded, and how optimization should avoid turning temporary confidence into persistent error.

6 Theoretical Perspectives and Assumption Boundaries

6.1 Risk correction under class-conditional noise

Classical LNL theory starts from the question of when learning from $(X, \tilde{Y})$ can recover the predictor that would have been learned from (X, Y) . Under class-conditional noise, the noise process is summarized by the transition matrix $T_{ij} = P(\tilde{Y} = j \mid Y = i)$. If T is known, well conditioned, and independent of X , the noisy risk can be related to the clean risk by correcting the loss or the prediction probabilities. This line of work supports unbiased risk estimators, backward correction, forward correction, and robust-loss variants (Natarajan et al. 2013; Patrini et al. 2017; Zhang and Sabuncu 2018).

The theoretical value of this setting is clarity. It separates the statistical problem of recovering clean supervision from the optimization problem of fitting deep networks. It also explains why transition-matrix methods remain important: they provide a concrete answer to the question of what information is needed to undo label corruption. However, the same clarity depends on restrictive assumptions. If T is poorly estimated, non-invertible, feature-dependent, or incompatible with open-set contamination, correction can amplify errors rather than remove them.

Robust losses provide a related but different theoretical lens. Instead of explicitly estimating T , they change the loss geometry so that corrupted labels have bounded or reduced influence. Generalized cross entropy, logit clipping, active negative losses, joint asymmetric losses, regroup median losses, and Rockafellian relaxation all reflect this idea in different forms (Zhang and Sabuncu 2018; Wei et al. 2023a; Ye et al. 2023; Wang et al. 2025; Li et al. 2025a; Chen et al. 2026). The common theoretical question is not only whether a loss is robust in isolation, but also which noise model, class structure, and optimization regime make its robustness meaningful.

6.2 Identifiability beyond class-conditional noise

The main theoretical difficulty of modern LNL is that realistic noise is often not identifiable from noisy labels alone. When the corruption process becomes instance-dependent, $T_{ij}(x) = P(\tilde{Y} = j \mid Y = i, X = x)$, the learner must infer both the clean decision structure and the feature-dependent reliability pattern. Generic identifiability may require additional information, such as multiple noisy labels, clean anchor examples, structural constraints, or assumptions on how labels are generated (Liu et al. 2023).

This limitation changes how transition-based methods should be interpreted. A method may appear to estimate a transition process, but the estimate is meaningful only under its information budget. Part-dependent and extended-transition formulations reduce the difficulty by imposing structure on the noise process or by expanding the label space to represent open-set contamination (Xia et al. 2020, 2023). These approaches are not merely engineering refinements; they are attempts to restore identifiability by adding structure that the raw noisy-label problem does not provide.

Identifiability also matters for semantic, partial-label, and open-world settings. If an example is ambiguous, outside the target taxonomy, or only weakly labeled, there may be no single clean class that a method can recover without extra semantic information. This is why structured supervision mismatch is a useful theoretical abstraction. It asks which latent variables are recoverable: the true class, the in-distribution status, the annotator reliability, the semantic neighborhood, or only a calibrated uncertainty distribution.

Table 7 summarizes this recoverability view. It separates the object a method may hope to recover from the extra information needed to make that recovery meaningful. This distinction is important because many empirical disagreements in LNL are not only disagreements about algorithms; they are disagreements about which latent object the benchmark makes observable.

6.3 Causal assumptions and method choice

Different LNL method families rely on different assumptions about the data-generating process. Semi-supervised approaches assume that reliable examples can seed a representation or pseudo-labeling process, and that the remaining examples can be treated as unlabeled or weakly labeled. Explicit noise-modeling approaches assume that the corruption process has a structure that can be estimated or constrained. Causal analyses show that these two strategies are not universally interchangeable; their success depends on how features, clean labels, and noisy labels are generated (Yao et al. 2023).

This perspective explains why empirical rankings of method families are often unstable across benchmarks. A semi-supervised method may work well when clean and noisy examples are separable in the early representation space, but it can fail when hard clean samples resemble noisy ones. A transition-modeling method may work well when noise is close to class-conditional, but it can fail when reliability depends on local geometry or when open-set samples violate the closed-label-space assumption. The theoretical task is therefore to specify an assumption profile before comparing methods.

Table 7 Recoverability under common noisy-supervision assumptions.

Assumption profile	Potentially recoverable object	Extra information or structure needed	Failure if treated too strongly
Class-conditional closed-set noise	Clean class posterior or corrected risk	Known, estimable, or well-conditioned transition process; clean anchors or validation data	Incorrect transition estimates amplify errors
Instance-dependent noise	Local reliability pattern and sample-specific clean target	Repeated labels, clean subsets, structural constraints, temporal dynamics, or neighborhood assumptions	Hard clean samples are identified as noise
Semantic ambiguity	Uncertainty over plausible labels or class neighborhoods	Class hierarchy, expert disagreement, semantic embeddings, or calibrated soft targets	Ambiguity is collapsed into overconfident correction
Open-set or open-world contamination	In-distribution membership and unknown-class evidence	OOD signals, prototype distance, open-world discovery cues, or mixed-label-space metadata	Unknown samples are forced into known classes
Weak, partial, or LLM-generated labels	Candidate-set reliability or annotator/process reliability	Label provenance, repeated prompts, agreement signals, or process-specific uncertainty models	Generated labels become pseudo-ground truth
Federated or graph noise	Client-level, neighborhood-level, or propagation-level reliability	Client metadata, topology, privacy-aware aggregation signals, or graph-specific corruption models	Noise is confused with heterogeneity or relational structure

An assumption profile should include the label space, the availability of clean validation data, the number of labels per instance, the noise rate or transition prior, the presence of open-set examples, the class imbalance regime, and whether external representations are used. These details determine whether a method has enough information to identify the target it claims to recover. Without them, robustness claims risk becoming benchmark-specific observations rather than transferable principles.

Taken together, these theoretical perspectives define an information boundary for robust learning. Identifiability determines what can be recovered from the available supervision, while the dynamics reviewed in Section 4 determine when the required reliability signals become observable or disappear. The unified pipeline in the next section can therefore be read as an algorithmic response to two constraints: what information is available and how training transforms that information over time.

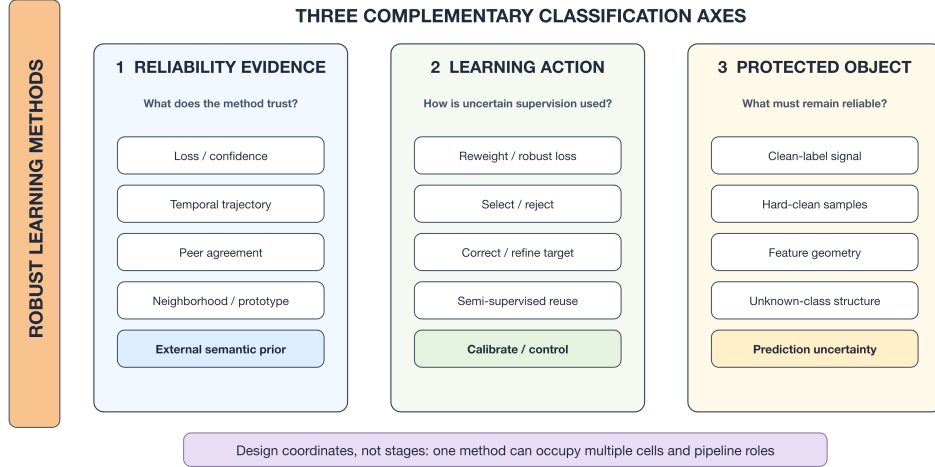


Fig. 5 Three complementary axes for classifying robust LNL methods. A method can be characterized by the reliability evidence it trusts, the learning action it applies to uncertain supervision, and the object it aims to protect. These axes are not sequential stages, and a single method may occupy multiple categories.

7 A Unified Pipeline for Robust Learning

The method literature is often described through families such as robust loss, sample selection, label correction, semi-supervised learning, contrastive learning, meta-learning, open-set detection, and calibration. These categories are useful for indexing papers, but they can obscure how components depend on one another during training. A sample-selection rule depends on the current representation, and a target-refinement rule depends on the quality of the selected examples. A robust loss changes the loss landscape that later selection rules observe, while an external foundation-model prior may help detection but import a different taxonomy. We therefore organize methods as a pipeline with five recurring roles: noise characterization, reliable-signal isolation, target refinement, representation-preserving learning, and optimization or uncertainty control. Figure 5 provides a complementary method taxonomy, while Table 8 maps representative works to their assumptions, roles, resource budgets, and boundaries.

The three axes in Fig. 5 and the five pipeline roles answer different questions. The axes describe a method’s design coordinates: what evidence it trusts, what action it takes, and what object it seeks to protect. The pipeline describes where those choices function in a robust-learning system. Reliability evidence primarily supports characterization and isolation; learning actions such as reweighting, selection, correction, and semi-supervised reuse implement isolation and refinement; protected objects such as labels, representations, calibration, and unknown-class structure determine whether preservation or uncertainty control is required. The mapping is many-to-many rather than one-to-one: a prototype-consistency method, for example, uses neighborhood evidence, performs both selection and refinement, and protects representation geometry.

This distinction prevents the taxonomy from being mistaken for a second chronological pipeline.

The pipeline does not imply that every method explicitly implements all five stages or that the stages always occur in a fixed chronological order. It is an analytical decomposition for reading methods. Co-teaching and DivideMix, for example, are often grouped as sample-selection or semi-supervised methods, yet their behavior also depends on memorization dynamics, pseudo-label calibration, and augmentation consistency (Han et al. 2018; Li et al. 2020). Transition correction and robust-loss methods appear more statistical, but their practical success still depends on whether the assumed noise process is identifiable and whether optimization amplifies correction errors (Natarajan et al. 2013; Patrini et al. 2017; Zhang and Sabuncu 2018; Liu et al. 2023). Recent representation, open-world, graph, and foundation-model methods make these dependencies even more explicit (Kim et al. 2024; Pan et al. 2025; Kim et al. 2025; Wei et al. 2024; Chen et al. 2025).

Historically, this pipeline grows out of several complementary 2018–2022 design choices rather than replacing them. MentorNet learned a data-driven curriculum for sample weighting, while gradient-based meta-reweighting used a small clean validation set to define which training examples should matter (Jiang et al. 2018; Ren et al. 2018). Co-teaching exchanged small-loss examples between peer networks, and JoCoR coupled peer agreement with small-loss joint training (Han et al. 2018; Wei et al. 2020). Early-Learning Regularization protected early predictions from later memorization, while DivideMix converted the estimated clean/noisy partition into a semi-supervised target-refinement problem (Liu et al. 2020; Li et al. 2020). Recent trajectory, neighborhood, prototype, open-world, and foundation-model approaches can be read as richer answers to the same three questions—which reliability evidence is available, how uncertain supervision should be used, and what information should be preserved—under less favorable noise and information assumptions.

This decomposition helps explain why component combinations can improve performance in one setting and fail in another. A robust loss mainly acts at the optimization stage, but it may implicitly affect reliable-signal isolation by changing which examples obtain low loss. A semi-supervised method mainly uses sample selection and target refinement, but its success depends on the quality of the representation used to form pseudo-labels. A prototype or structural-label method mainly protects representation geometry, but it can also become a selection rule when prototypes define which examples are trusted. The practical question is therefore not only which method family is strongest, but which stage provides the bottleneck under a given noise model.

Table 8 Representative LNL methods mapped to assumptions, observable signals, pipeline roles, resource budgets, and practical boundaries.

Method	Main assumption and observable signal	Principal pipeline role	Information or resource budget	Main boundary
Forward correction (Patrini et al. 2017)	Class-conditional corruption represented by an estimable transition matrix	Noise characterization and loss correction	Known or estimated transition probabilities	Sensitive to transition estimation and instance-dependent or open-set noise
GCE (Zhang and Sabuncu 2018)	Bounded influence of high-loss examples improves tolerance to mislabeled targets	Optimization control through a robust loss	No clean subset required; loss hyperparameter must be chosen	Does not explicitly identify or repair noisy labels
Co-teaching (Han et al. 2018)	Clean examples tend to have smaller early losses; peer disagreement limits self-confirmation	Reliable-signal isolation	Two networks and a keep-rate schedule, commonly linked to a noise-rate estimate	Hard clean examples may be discarded; both networks can share bias
DivideMix (Li et al. 2020)	Per-sample loss mixtures can separate mostly clean from noisy examples	Isolation, semi-supervised reuse, and target refinement	Two networks, mixture fitting, augmentation, and iterative pseudo-labels	Separation can collapse under structured hard noise or poorly calibrated confidence
DISC (Li et al. 2023)	Instance-specific temporal prediction consistency reveals changing reliability	Dynamic selection and target correction	History of predictions and instance-wise thresholds	More state and schedule sensitivity; evidence is protocol dependent
Structural labels (Kim et al. 2024)	Neighborhood structure carries supervision that is less brittle than one-hot labels	Representation preservation and target refinement	Reliable feature neighborhoods and structural-label construction	Unreliable neighborhoods can propagate early representation errors
VLM-LNL (Wei et al. 2024)	A pre-trained vision-language model provides external semantic evidence about label compatibility	Noise detection and representation-guided correction	Access to a suitable foundation model and its label vocabulary	Can inherit pre-training bias, domain mismatch, and taxonomy mismatch
Open-world prototype methods (Pan et al. 2025; Zhang et al. 2025b)	Prototype distance and cluster structure distinguish clean, mislabeled, and unknown samples	Noise characterization, rejection, and representation preservation	Feature clustering or prototype estimation with an open-world decision rule	Unknown discovery is sensitive to representation quality and class overlap

7.1 Noise characterization

The first stage of the pipeline asks what kind of supervision mismatch is present before deciding which examples can be trusted. Classical transition-based formulations model label corruption through a class-conditional matrix and motivate loss correction when the matrix is known or estimable (Natarajan et al. 2013; Patrini et al. 2017; Zhang and Sabuncu 2018). Later work relaxes this assumption by modeling part-dependent or instance-dependent noise, showing that reliability may vary with input features and local geometry (Xia et al. 2020; Chen et al. 2021; Liu et al. 2023). Open-set and open-world formulations add another diagnostic step: the learner must decide whether an example belongs to the target label space before deciding how to repair its

observed label (Xia et al. 2023; Wan et al. 2024; Pan et al. 2025; Zhang et al. 2025b). This characterization step is often implicit, but it determines whether the rest of the pipeline is solving correction, rejection, calibration, or representation recovery.

7.2 Reliable-signal isolation

Reliable-signal isolation then asks which training signals can guide learning at the current stage of optimization. Classical small-loss and disagreement-based methods assume that clean examples are easier to fit in early epochs and that peer models can reduce confirmation bias (Arpit et al. 2017; Han et al. 2018). DivideMix turns the clean/noisy split into a semi-supervised learning problem, treating uncertain samples as unlabeled rather than discarding them (Li et al. 2020). More recent approaches make the reliability signal richer: DISC tracks instance-specific selection and correction dynamics, RankMatch uses confidence and rank consistency, OT-Filter uses optimal transport, and CrossSplit uses data splitting to reduce memorization feedback loops (Li et al. 2023; Zhang et al. 2023; Feng et al. 2023; Kim et al. 2023). These methods share the idea that reliability is not a fixed property of an example; it is observed through loss, ranking, agreement, neighborhood structure, or temporal behavior.

The trend is toward reliability estimation that is both structured and economical. ANNE combines loss-based and nearest-neighbor structure so that high- and low-confidence regions can be handled differently (Cordeiro and Carneiro 2025). Surrogate selection models try to learn reusable reliability estimators rather than relying on a single hand-designed score (Liang et al. 2025). Early Cutting focuses on mislabeled easy examples, which are dangerous because they can pass small-loss filters before conventional noisy samples are memorized (Yuan et al. 2025). Jump-teaching uses temporal disagreement to reduce sample-selection bias without the full cost of dual-network training (Ji et al. 2026). Vision-language-model detectors add external semantic priors to the isolation stage, but later VLM work also shows that such priors can introduce endogenous confirmation bias if the detector and corrector reinforce each other (Wei et al. 2024; Ning and Chen 2026).

Reliable-signal isolation therefore determines the data-utilization trade-off. Conservative selection reduces the risk of using corrupted labels but may discard hard clean examples, tail classes, or boundary samples. Aggressive selection preserves more data but increases confirmation bias. Recent methods soften this binary choice by weighting examples, delaying correction, using multiple views of reliability, or moving low-confidence samples into unlabeled or weakly labeled roles (Li et al. 2020; Zhou et al. 2024; Cordeiro and Carneiro 2025; Zhang et al. 2025a). For this review, the important question is not whether a method selects samples, but what evidence it treats as reliable, how expensive that evidence is, and how the evidence changes as the representation evolves.

7.3 Target refinement and correction

After useful signals have been isolated, the next question is how unreliable targets should be represented. Target refinement includes hard relabeling, soft targets, teacher predictions, bootstrapped objectives, probabilistic correction, and role reassignment

between labeled, unlabeled, and ignored sets. DivideMix is a canonical example because it couples a clean/noisy split with MixMatch-style semi-supervised learning, thereby converting label uncertainty into a target-refinement problem (Li et al. 2020). DISC and L2B show a different pattern: selection and correction are intertwined, and the relative trust placed in observed labels and model predictions is adjusted over training (Li et al. 2023; Zhou et al. 2024). The central risk is error amplification. If a model’s early representation is biased, corrected labels can become more confident versions of the same bias.

Modern refinement methods address this risk by representing uncertainty more explicitly. Dirichlet-based prediction calibration treats calibration as part of noisy-label training rather than as a post-processing step (Zong et al. 2024). Variational rectification inference formulates adaptive loss rectification as amortized inference, making correction itself an uncertainty-aware object (Sun et al. 2025). Detect-and-Correct separates detection from selective correction, which is useful when only part of the noisy set should be relabeled (Grinberg et al. 2025). Instance-level dynamic optimization refines the training response according to sample difficulty, avoiding the assumption that all high-confidence corrections are equally useful (Zhang et al. 2025a). These methods indicate that correction is less about replacing a wrong symbol once than about maintaining a calibrated target distribution over training.

The target to be refined also changes across supervision regimes. In partial-label, multi-view, or multi-label settings, the learner may need to refine a candidate set or a missing view rather than a single class label (Li et al. 2025b). In LLM-generated label settings, the annotation process is prompt- and model-dependent, so refinement may rely on neighborhood label distributions, simplex diffusion, or repeated agreement patterns rather than a fixed clean/noisy split (Ye et al. 2025). In noisy preference learning and alignment, the noisy target may be a preference comparison rather than a class label, motivating semi-supervised or robust policy-optimization views of preference data (Cao et al. 2026; Liu et al. 2026). These cases broaden target refinement from label correction to supervision-shape correction.

A useful way to compare target-refinement methods is to ask what is being refined. Some methods refine labels, some refine sample weights, some refine prediction distributions, and some refine the role of an example by moving it between labeled and unlabeled sets. These choices are not equivalent. Hard relabeling can be effective when classes are well separated, but it is brittle under semantic ambiguity. Soft labels better express uncertainty, but they can dilute minority-class signals. Calibration-based approaches improve confidence or uncertainty estimates (Zong et al. 2024; Graham-Knight et al. 2025; Shi et al. 2026). Such improvements do not by themselves demonstrate that representation geometry has been preserved. The safest target-refinement designs therefore make explicit which uncertainty is epistemic, which is semantic ambiguity, and which is likely caused by corrupted supervision.

7.4 Representation-preserving learning

Representation-preserving learning is the part of the pipeline that most clearly separates recent LNL from a purely label-cleaning view. Fine-grained LNL, long-tailed

noisy data, and structural-label learning show that robustness often requires preserving or reconstructing class geometry under unreliable labels (Wei et al. 2023b; Lu et al. 2023; Kim et al. 2024). Contrastive methods attempt to prevent noisy labels from collapsing local neighborhoods; structural-label methods use reverse nearest-neighbor or related geometric information to create alternative targets; and optimal-transport methods impose global matching constraints so that selection is not purely local (Huang et al. 2023; Chang et al. 2023; Feng et al. 2023; Kim et al. 2024). In such settings, correcting labels without protecting the representation can leave the model with unstable neighborhoods and unreliable prototypes.

Open-set and open-world LNL make representation preservation even more central. Extended transition models expand the label space to handle mixed closed-set and open-set noise, but practical open-world methods must also separate clean in-distribution samples, closed-set noisy samples, and out-of-distribution contaminants in feature space (Xia et al. 2023; Wan et al. 2024; Pan et al. 2025). Multi-prototype modeling addresses this by representing each class with multiple feature anchors, which helps distinguish hard clean samples from unknown-class contaminants (Zhang et al. 2025b). Prototype-guided graph supervision similarly replaces direct noisy labels with class-level prototype signals under sparse or unreliable node labels (Li et al. 2026a). These examples show that prototypes and neighborhoods are not only evaluation tools; they are supervision carriers.

This stage also connects LNL with multimodal and foundation-model methods. Noisy correspondence learning treats mismatched image-text pairs as supervision mismatch in cross-modal alignment rather than ordinary class-label corruption (Han et al. 2023). Vision-language-model noisy-label detectors use external semantic alignment to separate reliable from unreliable examples (Wei et al. 2024). Noisy pre-training studies further show that label noise can be embedded in representations before downstream fine-tuning begins, affecting in-domain transfer and out-of-domain generalization differently (Chen et al. 2024, 2025). These methods point to a broader principle: when labels are unreliable, robust learning can borrow structure from representations, modalities, or pre-trained models.

The risk is that externally supplied structure can become another source of bias. A pre-trained model may encode a taxonomy that differs from the downstream task, a cross-modal similarity score may be unreliable for rare concepts, and neighborhood constraints may amplify majority-class structure. Recent foundation-model and domain-generalization evidence suggests that external priors can confuse label noise with domain shift or create confirmation loops when the same representation is used for detection and correction (Ning and Chen 2026; Wang et al. 2026). Representation-preserving learning therefore requires a balance between using external or geometric priors and preserving task-specific distinctions. This balance is particularly important in fine-grained, long-tailed, and open-world settings, where the distinction between a hard clean example and an out-of-distribution example may be subtle (Lu et al. 2023; Pan et al. 2025; Sosea and Caragea 2026).

7.5 Optimization and uncertainty control

Optimization details are not secondary implementation choices in LNL. Warm-up length, confidence thresholds, teacher refresh frequency, loss bounds, pseudo-label schedules, and uncertainty calibration can determine whether a method exploits early learning or memorizes noise. Classical robust objectives such as generalized cross entropy reduce the influence of corrupted labels by changing the loss geometry (Zhang and Sabuncu 2018). LogitClip controls memorization by bounding model predictions at the logit level, while active negative losses redesign the passive component of the objective (Wei et al. 2023a; Ye et al. 2023). Joint asymmetric loss, RML++, and Rockafellian relaxation continue this line by combining robustness conditions with practical reweighting or relaxation frameworks (Wang et al. 2025; Li et al. 2025a; Chen et al. 2026). The common point is that optimization is not merely the executor of a denoising rule; it is one of the places where robustness is created or lost.

Schedule design interacts strongly with sample selection and target refinement. If a warm-up period is too short, the representation may not expose meaningful reliability signals; if it is too long, the model may already have memorized easy noisy examples. Early Cutting and instance-level dynamic optimization explicitly treat this timing problem as part of method design (Yuan et al. 2025; Zhang et al. 2025a). CrossSplit and Jump-teaching address related confirmation-bias issues by changing how temporal information or data partitions are used during training (Kim et al. 2023; Ji et al. 2026). In object detection, early-learning distillation shows that the same timing issue appears when labels contain both classification and localization noise (Choi et al. 2025). These examples suggest that a method’s schedule should be reported as a core component, not as an implementation detail.

Flatness and uncertainty provide two additional views of optimization control. SAM-style analysis suggests that sharpness-aware optimization can be robust to label noise under certain regimes, while noise-compensated SAM variants attempt to adapt this idea more directly to noisy-label settings (Baek et al. 2024; Xu and Pang 2026). Conformal-in-the-loop and conformal margin risk minimization tie reliability control to coverage or margin guarantees rather than only to heuristic confidence thresholds (Graham-Knight et al. 2025; Shi et al. 2026). Theoretical work on label-noise SGD further suggests that noisy supervision can shape the implicit bias and trajectory of optimization, not just the final empirical risk (Zhang et al. 2026). These threads make optimization a bridge between statistical assumptions and training-time behavior.

The choice between semi-supervised isolation and explicit noise modeling also depends on assumptions about the data-generating process. Causal analyses of method choice show that semi-supervised strategies and noise-modeling strategies are not universally interchangeable; their success depends on how features and labels are generated (Yao et al. 2023). This is why a pipeline view is useful: it makes explicit which assumption supports each stage and where a method may fail when the noise structure changes. In practice, a method that improves closed-set synthetic-noise accuracy by aggressive pseudo-labeling may be less suitable when the main risk is open-set contamination, domain shift, or annotator-process noise (Wan et al. 2024; Wang et al. 2026; Ye et al. 2025).

Table 9 Assumption-aware method-selection guide.

Diagnosis	Observable evidence to check	Pipeline emphasis	Metrics and reporting	Failure if ignored
Closed-set class-conditional noise	Credible noise-rate prior, clean anchors, or validation labels; fixed label space	Transition correction, robust losses, and calibrated reweighting	State noise-rate/anchor assumptions; report sensitivity to T	Incorrect transition priors dominate
Instance-dependent or boundary noise	Loss trajectories, feature neighborhoods, repeated labels, or clean validation signals	Time-aware isolation plus target refinement and geometry checks	Report detection precision/recall and hard-clean retention	Hard or tail samples are removed as noise
Open-set or open-world contamination	OOD scores, prototype distance, unknown-class evidence, or mixed-label-space metadata	Membership detection before label correction; multi-prototype separation	Report clean/ID-noisy/OOD separation, AUROC, and rejection behavior	Unknown samples are forced into known classes
Fine-grained or long-tailed data	Class hierarchy, class frequency, semantic similarity, and tail-class loss patterns	Class-aware selection, contrastive/prototype preservation, and calibration	Report per-class, tail, and ambiguity-sensitive metrics	Subtle distinctions are erased
Weak, LLM-generated, or preference labels	Prompt/model provenance, repeated agreement, annotator reliability, or preference uncertainty	Process-specific uncertainty modeling and selective refinement	Report provenance, calibration, and uncertainty over targets	Pseudo-labels become overconfident
Graph, federated, or foundation-model settings	Client-level reliability, neighborhood propagation, pre-training source, or domain shift	Multi-level reliability: sample, client, neighborhood, model prior, and task	Report client/graph/transfer metrics and pipeline ablations	Noise is confused with structure or shift

7.6 Synthesis: assumption-aware method selection

The pipeline also suggests a practical way to choose among methods. If the main uncertainty is class-conditional and the label space is closed, transition modeling or robust losses may be appropriate, especially when clean anchors, validation data, or credible transition priors are available (Natarajan et al. 2013; Patrini et al. 2017; Zhang and Sabuncu 2018). If the main difficulty is instance-dependent reliability, methods that use trajectories, neighborhoods, or clean validation signals become more relevant (Xia et al. 2020; Chen et al. 2021; Li et al. 2023; Cordeiro and Carneiro 2025). If open-set contamination is present, the learner must first detect whether an example belongs to the target label space before deciding how to refine its label (Xia et al. 2023; Wan et al. 2024; Pan et al. 2025; Zhang et al. 2025b). If LLM-generated labels or weak labels are used, uncertainty over the annotation process should be modeled rather than hidden behind a single noise rate (Wu et al. 2023b; Ye et al. 2025; Sosea and Caragea 2026).

Assumption-aware selection also has to account for domain structure. Fine-grained and long-tailed settings require class-aware reliability estimation because hard clean examples may systematically look noisy (Wei et al. 2023b; Lu et al. 2023). Federated settings add client-level heterogeneity, privacy constraints, and class imbalance, so robust learning must reason about both sample reliability and client reliability (Wu et al. 2023a; Huang et al. 2026). Graph settings add relational propagation: a wrong label can influence neighbors through message passing (Kim et al. 2025). A correct but sparse node label may instead need prototype-level support (Li et al. 2026a). Foundation-model settings add pre-training mismatch, semantic priors, and domain generalization concerns (Chen et al. 2024, 2025; Wang et al. 2026). These examples show why a single leaderboard ranking is rarely enough to choose a method.

This assumption-aware view is important for reproducibility. Many reported gains depend on auxiliary conditions such as access to a clean subset, known noise rates,

Clean sample (ID)	Sym	Asym	IDN	SILN	CIFAR-10N Worst	
 airplane · 35873	 airplane	 airplane	 bird	 ship	 airplane	Worst = Clean
 automobile · 43195	 ship	 automobile	 truck	 frog	 automobile	
 truck · 7556	 ship	 automobile	 automobile	 automobile	 automobile	Worst ≠ Clean
 cat · 9357	 ship	 dog	 bird	 dog	 dog	
 clean-label retention  semantically plausible direction  other synthetic target  distant class						

Fig. 6 Same-sample schematic audit comparing representative Sym, Asym, IDN, and SILN targets with CIFAR-10N Worst labels. Each row fixes one CIFAR-10 training image and its sample ID; the rightmost column reports the aligned human-annotated CIFAR-10N Worst label. Colors distinguish clean-label retention, semantically plausible directions, other synthetic targets, and distant classes. The two row blocks separate cases where Worst keeps the clean label from cases where it changes the label (Wei et al. 2022).

balanced classes, strong augmentations, repeated labels, or a pre-trained model with compatible semantics. A method that is robust under one assumption profile may be fragile under another. Method comparison should therefore report both accuracy and the information budget used at each pipeline stage. It should state whether clean data are used, the noise rate is known, open-set samples are present, external models provide semantic priors, and calibration or uncertainty is evaluated (Liu et al. 2023; Yao et al. 2023; Shi et al. 2026). Without this information, robustness claims risk becoming benchmark-specific observations rather than transferable guidance.

Table 9 serves as the method-choice map: each row links a diagnosed supervision mismatch to observable evidence, pipeline emphasis, reporting requirements, and the main failure risk that should be checked before comparing methods.

8 Benchmarking and Evaluation under Realistic Noise

8.1 Synthetic, controlled, and real-world protocols

Evaluation is a central methodological issue in LNL because the apparent difficulty of a method often depends on how noise is generated. Synthetic noise protocols remain useful because they provide controlled stress tests and reproducible comparisons. However, they often simplify the structure of real annotation errors. Random class flips may make clean and noisy examples easier to separate than in realistic settings, and conclusions under CIFAR-style synthetic noise may not transfer to human annotation noise, open-set contamination, long-tailed noisy data, or LLM-generated labels.

Figure 6 uses four aligned training examples as label-level probes. The point is not to rank protocols from a few images, but to show what a noise protocol actually fixes: which samples are corrupted and which target classes they receive. The top

Table 10 Controlled quantitative snapshot under 40% instance-dependent noise (IDN). Test accuracy (% , mean $\pm$ standard deviation) is transcribed from Table 1 of DISC (Li et al. 2023) for its CIFAR protocol. Asterisks mark results reimplemented by the DISC authors; the remaining baseline results were taken by that paper from its cited comparison source. This table documents one protocol and is not a universal method ranking.

Method	Principal mechanism	CIFAR-10	CIFAR-100
CE*	Standard cross-entropy baseline	67.64 $\pm$ 0.26	43.17 $\pm$ 0.15
Forward T (Patrini et al. 2017)	Transition-matrix loss correction	79.37 $\pm$ 2.72	42.80 $\pm$ 1.01
GCE* (Zhang and Sabuncu 2018)	Robust generalized cross entropy	78.95 $\pm$ 0.15	45.69 $\pm$ 0.14
Co-teaching (Han et al. 2018)	Peer small-loss sample exchange	73.00 $\pm$ 1.24	23.21 $\pm$ 0.57
DivideMix (Li et al. 2020)	Loss-mixture separation and semi-supervised reuse	95.07 $\pm$ 0.11	76.08 $\pm$ 0.35
DISC (Li et al. 2023)	Dynamic instance-specific selection and correction	95.94 $\pm$ 0.04	78.44 $\pm$ 0.19

block shows a failure mode of synthetic evaluation that is easy to miss in aggregate statistics: CIFAR-10N Worst keeps the clean label for these images, yet some synthetic protocols still assign a different target. The bottom block shows the complementary case, where real annotator noise changes the label in semantically plausible directions such as truck-to-automobile and cat-to-dog. Some synthetic protocols reproduce these directions on particular rows, but others either keep the clean label or send the image to distant classes. These row-level mismatches illustrate why reporting only a global noise rate is insufficient: two benchmarks with similar rates may still differ in sample selection, semantic direction, and target-label generation.

Table 10 complements this qualitative audit with one deliberately narrow numerical comparison. It holds the source paper, datasets, IDN generator, and nominal noise rate fixed. The substantial changes in relative performance across CIFAR-10 and CIFAR-100 illustrate why a review should not combine results from different backbones, data splits, noise generators, or training budgets into a single ranking.

Recent benchmarks and analyses make this limitation explicit. CIFAR-N introduced real human annotations for controlled study of real-world label noise (Wei et al. 2022). BadLabel shows that even synthetic corruption can be constructed to make clean and noisy loss distributions difficult to distinguish, exposing the fragility of many small-loss assumptions (Zhang et al. 2024). NoisywikiHow extends the discussion to NLP by constructing a benchmark with heterogeneous real-world-inspired noise sources (Wu et al. 2023b). The BeGIN benchmark further shows that graph learning under instance-dependent noise requires node-specific and topology-aware evaluation rather than a direct transfer of image-classification protocols (Kim et al. 2025). These works show that benchmark realism is not a binary distinction between synthetic and

Table 11 Evaluation axes for realistic noisy-supervision benchmarks.

Axis	Key question	Typical confound	Reporting expectation
Noise source	Who or what generated the label?	Human ambiguity, web collection, weak rules, LLM prompts	Describe annotation or generation process
Noise structure	How does noise depend on data?	Class imbalance, instance difficulty, open-set samples	Report beyond a single global noise rate
Observability	What clean information is available?	Clean validation data, repeated labels, known priors	State information budget explicitly
Evaluation unit	What must be robust?	Accuracy, detection, calibration, representation, transfer	Use metrics aligned with the claimed problem
Transfer setting	Where is the model evaluated?	In-domain versus out-of-domain shift	Separate robustness from domain adaptation

real. It depends on the difficulty, source, semantic structure, relational context, and observability of the noise process.

Existing real-noise benchmarks should not be treated as interchangeable. They differ in how labels are produced, whether clean labels are available for diagnosis, whether the noise is human, web-derived, semantic, open-set, or relational, and whether the benchmark is intended for classification accuracy, noise detection, calibration, or representation transfer. Table 12 summarizes commonly used real-noise or real-source noisy-label datasets and the evaluation caveat associated with each.

This map suggests three practical reporting rules. First, papers should distinguish datasets with instance-level clean references, such as CIFAR-N, from datasets where clean labels are available mainly for validation or testing, such as Clothing1M and WebVision. Second, web-derived datasets should report whether the main difficulty is class confusion, search-engine bias, open-set contamination, or domain shift, because these factors support different method choices. Third, cross-domain datasets such as NoisywikiHow and BeGIN should be used not only to report extra accuracy numbers, but also to test whether assumptions developed on image benchmarks survive changes in modality and data structure. The final column incorporates the case-study reporting lessons that would otherwise require a second, partly repetitive table.

A complementary line of evaluation treats mislabeled-data identification as a benchmarkable data-quality problem. The Artificial Intelligence Review benchmark by Nazaretyan et al. surveys label-error detection methods for tabular data and empirically compares representative filters across many datasets (Nazaretyan et al. 2025). Its focus differs from deep representation learning, but its lesson is relevant to LNL. It

separates suspicious-label identification from downstream training, reports when classical data-cleaning filters are competitive, and shows that recommendations depend on dataset size, feature type, and error structure. A robustness claim should therefore specify whether a method addresses label-error detection, robust prediction under uncorrected noise, representation preservation, or all three.

8.2 Case-study readings of representative benchmarks

The benchmark map becomes more useful when representative datasets are read as case studies rather than as interchangeable rows in an accuracy table. CIFAR-N is a controlled human-noise case: the noisy labels come from human annotators while the original CIFAR labels remain available as a clean reference (Wei et al. 2022). This makes it valuable for asking whether a method detects genuinely mislabeled training examples or merely exploits synthetic-noise artifacts. It is less suitable for claims about web-scale open-set noise, because the label space is fixed, the images are low resolution, and the main uncertainty is human disagreement within a familiar closed-set taxonomy. A convincing CIFAR-N result should therefore report not only accuracy, but also whether the method preserves hard clean examples and whether its clean/noisy separation matches the available reference labels.

Web-derived benchmarks form a different case study because the noise is entangled with collection bias. Clothing1M, WebVision, Food-101N, and fine-grained web datasets derive supervision from shopping metadata, image search, recipe pages, or web category names rather than from a controlled annotation interface (Xiao et al. 2015; Li et al. 2017; Lee et al. 2018; Sun et al. 2021). Their labels can be wrong because the visual content is ambiguous, because the surrounding text is misleading, because a query retrieves an out-of-scope image, or because the web domain differs from the clean validation/test domain. For these datasets, a method that uses clean validation data, strong pre-training, or aggressive sample rejection should state that information budget explicitly. The useful question is not simply whether the method wins on Clothing1M or WebVision, but whether it is solving label correction, domain adaptation, open-set rejection, or class-imbalance control.

Text and LLM-generated labels provide a third case study in which the target itself may be process-dependent. NoisywikiHow introduces heterogeneous, instance-dependent noise in an NLP intent-classification setting, while SiDyP studies labels generated by LLMs and refines them through neighborhood and diffusion-style target updates (Wu et al. 2023b; Ye et al. 2025). In these settings, the source of noise is not only an annotator’s mistaken class choice. It can come from ambiguous instructions, prompt sensitivity, model-version changes, or a weak semantic match between an input and a generated label. Evaluation should therefore record label provenance and calibration behavior, not only final classification accuracy. A robust method should show that it handles uncertain supervision without turning LLM-generated targets into overconfident pseudo-ground truth.

Graph and tabular benchmarks highlight still another reading. BeGIN shows that noisy labels in graphs can propagate through neighborhoods and message passing, so the unit of failure is not always an isolated sample (Kim et al. 2025). The Artificial Intelligence Review tabular benchmark shows the complementary data-quality view:

suspicious-label detection can be benchmarked directly, and the best filter may depend on feature type, dataset size, and error structure (Nazaretyan et al. 2025). Together, these cases warn against treating LNL as only a deep image-classification problem. A graph method should report whether it controls propagation error and preserves useful topology; a tabular or data-quality method should clarify whether it is detecting label errors for human review or improving downstream robust prediction.

8.3 Metrics beyond final accuracy

Final test accuracy is necessary but insufficient. A realistic evaluation should also consider clean/noisy detection precision, performance on hard clean samples, calibration, representation stability, robustness across noise rates, data utilization, computation cost, and sensitivity to class imbalance or open-set contamination. For open-world noise, AUROC for unknown-sample detection and the quality of clean/noisy/out-of-distribution separation are important. For noisy pre-training and foundation-model learning, in-domain and out-of-domain transfer should be evaluated separately because noise may help local adaptation while degrading broader generalization (Chen et al. 2024, 2025).

Evaluation should also include pipeline-level ablations. Many methods combine sample selection, target correction, consistency regularization, and dynamic schedules. Reporting only final accuracy makes it difficult to determine whether a gain comes from better signal isolation, better target refinement, better representation geometry, or better optimization stability. This matters for reproducibility and for transferring methods across domains.

Metrics should also expose negative cases. A method that improves average accuracy by ignoring difficult samples may reduce class-level robustness or degrade tail-class recognition. A detector that removes many noisy samples may also remove informative clean samples near decision boundaries. A calibration or conformal-control method may improve confidence or coverage without reporting whether representation quality changes (Graham-Knight et al. 2025). These trade-offs are important for interpreting the review because different papers often optimize different implicit objectives while reporting a common top-line accuracy.

These metrics should be paired with the evaluation metadata in Table 11. In particular, detection metrics are meaningful only when instance-level clean references exist, while OOD metrics require an explicit unknown-sample construction. Hard-clean retention and retained-data fraction are useful for sample-selection methods because they distinguish genuine denoising from simply removing difficult supervision. Calibration metrics should be applied to the probability object actually used by the method—test predictions, selection confidence, or refined targets—rather than reported without that distinction.

8.4 Benchmark realism as an organizing axis

Prior surveys often include datasets as a later reference section. We argue that benchmark realism should be part of the main narrative because it determines what a method is actually shown to solve. A method that performs well on symmetric noise

may still fail under fine-grained ambiguity, long-tailed noisy data, open-set contamination, or LLM-generated labels. Conversely, a method designed for a domain-specific setting may rely on inductive biases unavailable elsewhere. Evaluation therefore bridges method design and cross-domain pressure tests.

This bridge is essential for interpreting recent work. As LNL moves from curated image classification to open-world data, multimodal alignment, federated clients, and LLM-generated annotations, a benchmark no longer measures only whether an algorithm tolerates wrong labels. It measures whether the algorithm can preserve useful information under multiple interacting sources of supervision mismatch. The next section uses recent conference, journal, and emerging preprint studies as pressure tests for this broader interpretation.

Table 12 Representative real-noise or real-source benchmarks, their diagnostic value, and comparison caveats.

Benchmark and task	Noise source and clean signal	Diagnostic value	Caveat and minimum reporting
CIFAR-N: image classification (Wei et al. 2022)	Human annotations with original CIFAR labels retained as a reference	Tests whether human label errors can be identified without discarding hard clean samples	Low-resolution closed set; report accuracy, detection quality, and hard-clean retention
Clothing1M: clothing recognition (Xiao et al. 2015)	Shopping metadata with clean train, validation, and test subsets	Tests high-volume web supervision and class imbalance	State exactly which clean subsets and pre-training are used
WebVision: large-scale image classification (Li et al. 2017)	Search-query labels with curated validation and test sets	Tests web-scale supervision, transfer, and dataset bias	Separate label-noise robustness from web-to-curated domain shift
Food-101N: food recognition (Lee et al. 2018)	Web recipe images with partially verified labels	Tests taxonomy noise under limited label verification	Report verification coverage and treatment of irrelevant images
ANIMAL-10N: animal recognition (Song et al. 2019)	Human labels for visually similar animal classes with a clean test set	Tests fine-grained semantic confusion	Narrow domain; report class-wise effects rather than only mean accuracy
Controlled web labels: image recognition (Jiang et al. 2020)	Professionally relabeled web subsets with controlled noise levels	Tests web noise while retaining an estimate of corruption rate	More curated than unconstrained web collection; document subset construction
WebFG and WebiNat: fine-grained recognition (Sun et al. 2021)	Web searches over fine-grained categories and hierarchies	Tests semantic hierarchy, imbalance, and irrelevant retrievals	No complete instance-level clean reference; state rejection and verification rules
NoisywikiHow: intent classification (Wu et al. 2023b)	Heterogeneous instance-dependent text noise	Tests whether image-derived assumptions transfer to text	Constructed benchmark; report noise provenance and calibration
BeGIN: graph node classification (Kim et al. 2025)	Algorithmic and LLM-based corruptions embedded in graph structure	Tests node-specific noise and propagation through topology	Graph-specific; report topology effects separately from node accuracy

Table 13 Metrics beyond final accuracy for LNL evaluation. No single row is mandatory for every study; the metric should match the claimed capability and the available reference information.

Metric or metric family	What it measures	Recommended reporting	Interpretation boundary
Balanced accuracy, macro-F1, and worst-class accuracy	Robustness under class imbalance and heterogeneous class-wise corruption	Report with overall accuracy and per-class noise statistics when available	Improvements may reflect class reweighting rather than better noise handling
Noise-detection precision, recall, and AUPRC	Whether selected or rejected samples correspond to known label errors	Define the positive class and report the threshold or precision-recall curve	Requires instance-level clean reference labels; AUROC can obscure rare noisy cases
Hard-clean retention	Fraction of difficult but correctly labeled samples retained as trusted supervision	Predefine difficulty independently of the evaluated selector and report retention at a fixed selection budget	Otherwise “hard” can be defined circularly by the method’s own loss
Unknown/OOD AUROC and FPR@95TPR	Separation of in-distribution mislabeled samples from examples outside the target label space	Report the unknown class construction and threshold-selection protocol	High OOD separation does not imply correct within-class relabeling
ECE, NLL, and Brier score	Calibration and probabilistic quality of predictions or refined targets	Report binning for ECE and distinguish test predictions from training pseudo-labels	ECE is bin-sensitive and can improve without better representation quality
Retained-data fraction and supervision-use breakdown	How much data is trusted, relabeled, treated as unlabeled, or rejected	Report fractions over time and per class, together with final performance	Similar accuracy may hide very different information and compute efficiency
Representation diagnostics	Preservation of semantic neighborhoods and transferable features	Use linear probes, k NN purity, prototype drift, or in-/out-of-domain transfer under a fixed protocol	A diagnostic correlation is not by itself a mechanism proof
Robustness profile and cost	Sensitivity across noise rates/seeds and the resources needed to obtain it	Report mean and variation, training time, model count, memory, and external pre-training access	A stronger average may rely on a larger information or compute budget

9 Cross-Scenario Pressure Tests

Studies from vision, language, multimodal learning, federated learning, graph learning, and domain-specific applications show that LNL is no longer confined to closed-set image classification. We organize this evidence by the assumptions it stresses rather than by publication year. Chronological evidence-status tables are retained in Supplementary Appendix B for provenance and updateability, while the main text focuses on what changes when the unit of noise, the source of reliability, or the evaluation target moves beyond a single class label.

The main pattern is that cross-scenario evidence stresses assumptions behind classical LNL. Some settings challenge the independence between label error and sample difficulty; others change the label space, move noisy supervision into pre-training or multimodal correspondence, or make annotation depend on an LLM, client, graph neighborhood, or preference pipeline. Taken together, these studies serve as pressure tests of the proposed framework rather than as a second method catalogue.

Read together, the cross-scenario evidence changes three objects that older LNL surveys often kept fixed. First, the unit of noise is no longer always a sample label; it can be a client, graph neighborhood, modality pair, prompt, preference comparison, or pre-trained representation. Second, the source of reliability is no longer only the model’s own loss; it may come from temporal trajectories, structural neighborhoods, prototypes, external semantic priors, or provenance metadata. Third, the evaluation target is no longer only closed-set accuracy; many settings require rejection, calibration, transfer, topology preservation, or preference consistency. These shifts explain why the studies below are used as assumption stressors rather than as a longer method catalogue.

9.1 Vision, long-tailed data, and detection

Recent vision papers show that noisy-label robustness must account for task structure. Fine-grained classification introduces semantic ambiguity, where classes are visually close and label errors are often meaningful rather than random (Wei et al. 2023b). Long-tailed noisy data adds another complication: tail-class samples are naturally hard, so their losses and predictions may resemble noisy examples (Lu et al. 2023). Object detection adds localization noise in addition to categorization noise, and ELDET shows that early-learning behavior can differ across these two noise types (Choi et al. 2025). Emerging 3D perception work pushes this pressure further: OccNL studies semantic occupancy under asymmetric voxel noise and dynamic trailing artifacts, where 2D image-classification denoising strategies can collapse in sparse 3D spaces (Li et al. 2026b). These settings pressure-test the assumption that sample-level loss is a reliable proxy for label correctness.

They also show why representation preservation is not optional. In fine-grained classification, an overly aggressive correction step may erase subtle inter-class distinctions. In long-tailed data, a selector that favors high-confidence samples may overfit head classes and underutilize tail classes. In detection, the model must reason about both what object is present and where it is located, so a label-noise method designed only for image-level targets may miss localization-specific corruption. These

cases illustrate how task structure determines which part of the pipeline becomes the bottleneck.

9.2 Foundation models and multimodal noisy supervision

Foundation models change where noisy supervision enters the system. Noisy labels can contaminate pre-training and therefore affect downstream representations even when downstream labels are clean (Chen et al. 2024, 2025). Vision-language models can also serve as semantic priors for detecting noisy labels during downstream adaptation (Wei et al. 2024). Cross-modal retrieval introduces noisy correspondence: the supervision is not a class label but a pairwise alignment between modalities (Han et al. 2023). These works extend LNL from label correction to representation and alignment reliability.

LLM-generated labels create another emerging setting. Their noise is prompt-dependent, context-dependent, and often different from synthetic flips or human crowd errors. SiDyP explicitly studies this problem and uses iterative diffusion-based refinement to calibrate pre-trained language classifiers trained on LLM-generated labels (Ye et al. 2025). Preference optimization makes the same issue visible in alignment pipelines: RE-PO treats preference-label correctness and annotator reliability as latent variables (Cao et al. 2026), while Semi-DPO reframes noisy preference pairs for diffusion models as a semi-supervised learning problem (Liu et al. 2026). This line of work suggests that future LNL must model annotators that are themselves learned systems or preference-aggregation pipelines.

This shift creates a feedback loop between noisy data and noisy models. A foundation model trained on weak web supervision may later be used to clean labels, generate pseudo-labels, or judge semantic consistency. Its prior can be useful, but it can also transfer hidden biases or taxonomy mismatches into downstream LNL. Future evaluations should therefore distinguish between using a foundation model as an external oracle and studying noise already embedded in that model.

9.3 Federated, graph, and domain-specific settings

Federated LNL illustrates how noise can be distributed across clients rather than examples. FedNoRo jointly handles class imbalance and label-noise heterogeneity, showing that client-level reliability can be as important as sample-level reliability (Wu et al. 2023a). FedRNC extends this view to federated class-incremental learning, where unseen classes may be mislabeled as known classes and later reappear on other clients or tasks (Huang et al. 2026). Graph and network settings introduce topological propagation: noisy labels or pseudo-labels can spread through message passing and amplify errors. BeGIN shows that graph-specific instance-dependent corruption, including LLM-based corruption, can expose failures hidden by class-dependent graph noise protocols (Kim et al. 2025). Prototype-guided graph supervision responds to this issue by replacing direct supervision from noisy sparse node labels with class-level prototype anchors (Li et al. 2026a). Medical LNL highlights ambiguity, expert disagreement, and risk-sensitive evaluation (Shi et al. 2024). These settings are not peripheral applications. They expose which parts of the unified pipeline are robust

Table 14 Boundary domains that expose additional noisy-supervision assumptions.

Domain	Typical supervision mismatch	What would need to be added	Status in this review
Tabular and data-quality filtering	Label errors mixed with feature outliers, missing values, and small-sample effects	Stronger connection between data-cleaning filters and downstream deep LNL objectives	Discussed through benchmarking evidence (Nazaretyan et al. 2025)
Remote sensing and geospatial data	Weak labels from maps, temporal drift, mixed pixels, and spatial autocorrelation	Spatially aware noise models and transfer evaluation across regions	Boundary / future work
Speech, audio, and time series	Alignment errors, segment-level labels, temporal ambiguity, and weak event tags	Sequence-aware reliability and metrics beyond image-level accuracy	Boundary / future work
Industrial quality control and fault diagnosis	Rare faults, noisy sensors, and partially verified labels	Integration of anomaly detection, class imbalance, and label reliability	Boundary / future work
Recommendation and click logs	Implicit feedback, exposure bias, and preference labels that are not direct ground truth	Joint modeling of selection bias, preference noise, and calibration	Connected to preference-noise discussion

to changes in data structure and which assumptions are inherited from closed-set image-classification benchmarks.

The common theme is that the unit of noise changes. In federated learning, a client can be unreliable even if some of its samples are clean. In graph learning, a neighborhood can propagate errors even when an individual label appears plausible. In medical imaging, a label may encode clinical uncertainty rather than a simple annotation mistake. These settings make it difficult to define robustness as recovering a single hidden clean label for each sample. They instead require reasoning about supervision reliability at the levels of samples, clients, neighborhoods, experts, and tasks.

9.4 Boundary domains and under-covered settings

The cross-scenario discussion is intentionally selective. It focuses on settings that currently have strong links to deep LNL, representation learning, and recent benchmark design. Other domains also create noisy-supervision problems, but the evidence is less evenly connected to the main LNL pipeline. Table 14 records these boundary areas so that the scope is explicit rather than implicit.

Table 15 Synthesis of the five research questions.

RQ	Main answer	Primary evidence	Boundary or caveat
RQ1	Modern noisy supervision includes semantic ambiguity, ontology mismatch, open-set contamination, weak targets, and process-dependent labels that cannot be reduced to random flips.	Taxonomy, structured supervision mismatch, and recoverability analysis	Different mismatches imply different latent targets; a single hidden clean label may not exist.
RQ2	Reliability becomes observable when optimization exposes stable differences in difficulty, trajectories, or representation geometry; it disappears when hard clean and coherent noisy samples overlap.	Memorization dynamics, representation diagnostics, and the small-loss case study	Observability depends on class balance, semantic structure, model state, and information budget.
RQ3	Robust methods can be compared through five interacting roles: characterization, isolation, refinement, representation preservation, and optimization or uncertainty control.	Unified pipeline and assumption-aware method guide	A component’s role and effectiveness can change with the noise process and available priors.
RQ4	A benchmark supports a robustness claim only when its noise-generating process, available clean information, metrics, and stress tests match the claimed setting.	Protocol comparison, benchmark case studies, and reporting checklist	Final accuracy alone cannot reveal hard-clean rejection, calibration failure, or representation damage.
RQ5	Emerging settings change the unit of noise, the source of reliability, and the evaluation target, thereby stress-testing assumptions inherited from closed-set classification.	Vision, foundation-model, federated, graph, and domain-specific evidence	Recent and preprint evidence should be treated as directional unless supported across settings.

10 Synthesis and Open Problems

10.1 Answers to the research questions

The five research questions in Table 2 can now be answered through a common principle: robustness depends on whether the supervision process, the information budget, the training dynamics, and the evaluation protocol make the claimed target recoverable. Table 15 records the main answer to each question together with the evidence base and the boundary that prevents overgeneralization.

These answers also define the research agenda. The open problems below are not independent topic lists; each addresses a missing formalism, observable, reporting standard, or provenance mechanism exposed by the RQ synthesis.

10.2 Formalizing semantic and open-world noise

The field still lacks a unified formalism for semantic ambiguity, open-world contamination, and partial supervisory targets. Current work often treats these as separate settings, but real data can contain all of them simultaneously. A useful future direction is to develop models that represent label reliability, class membership, and semantic proximity jointly while remaining identifiable under realistic supervision budgets.

Such a formalism should also separate invalid labels from incomplete labels. An out-of-distribution sample assigned to an in-distribution class is different from a partial-label example whose candidate set contains the correct class, and both differ from a fine-grained sample with legitimate semantic ambiguity. Treating all three as high-noise examples hides the information that each setting provides. A more useful theory would specify which latent variables are recoverable under which observations, such as repeated labels, clean anchors, class hierarchy, external semantic embeddings, or open-world discovery signals.

10.3 Identifiability and assumption-aware design

Identifiability remains a central theoretical and practical issue. Under instance-dependent noise, the transition process may not be recoverable without additional information (Liu et al. 2023). Under open-world noise, the learner must decide whether a failure is due to a wrong in-distribution label or an example outside the target label space. Under LLM-generated labels, the annotation mechanism may change with prompt, context, or model version. Future methods should therefore state which part of the supervision process is identifiable and which part is controlled by assumptions or priors.

This direction also changes how algorithms should be reported. Rather than presenting a method as generally noise-robust, authors should specify the assumption profile: closed or open label space, available clean data, repeated annotations, known noise rates, external priors, and expected class imbalance. Such reporting would make theoretical guarantees and empirical comparisons more compatible.

10.4 Representation diagnostics as first-class tools

Many methods implicitly rely on representation stability, but benchmarks rarely measure it directly. Future work should use diagnostics such as neighborhood consistency, prototype stability, calibration, feature drift, class-separation quality, and clean/noisy separability over time. These diagnostics can clarify when a method protects useful geometry and when it only improves final accuracy under a narrow protocol.

A promising direction is to treat representation diagnostics as training monitors rather than post-hoc visualizations. If neighborhood structure becomes unstable, a method may reduce the influence of aggressive pseudo-labeling. If prototype drift is

concentrated in tail classes, it may adjust sampling or calibration. If a foundation-model prior disagrees with learned features, the system may flag uncertain supervision instead of forcing a label correction. Such diagnostics would make representation analysis actionable within the pipeline.

10.5 Pipeline-level comparison and assumption reporting

The unified pipeline suggests a practical reporting standard. Authors should state what signal is used for reliability estimation, whether clean validation data or noise priors are assumed, how targets are refined, how representation collapse is prevented, and which schedule controls memorization. Such reporting would make it easier to compare semi-supervised isolation, transition modeling, external-prior methods, and dynamic optimization under shared assumptions.

Pipeline-level comparison also requires stronger ablations. For a hybrid method, removing only one module may not reveal whether the remaining modules compensate for it. More informative ablations would test each stage under controlled failures: noisy reliability signals, biased pseudo-labels, corrupted representation neighborhoods, and mismatched optimization schedules. This would help identify whether a method is robust because of a specific component or because the benchmark does not stress its weakest link.

10.6 Noisy supervision in foundation-model ecosystems

Foundation-model ecosystems introduce new forms of noisy supervision: noisy pre-training labels, web-scale image-text mismatches, LLM-generated annotations, preference noise, and noisy synthetic data. These settings blur the boundary between noisy data and noisy models. Future LNL research should therefore evaluate not only how to train a classifier on a noisy dataset, but also how label noise is stored, transferred, amplified, or mitigated through pre-training, adaptation, and inference.

The practical challenge is governance of supervision at scale. Future datasets may be labeled by mixtures of humans, weak rules, retrieval systems, vision-language models, and LLMs. Robust learning will need provenance-aware benchmarks that record how labels were produced, when they were revised, and which model or annotator generated them. Without such metadata, it will be difficult to determine whether a method learns from noisy labels, learns from noisy annotators, or inherits noise from a pre-trained system.

Table 16 Concrete research artifacts implied by the open problems.

Open problem	Concrete artifact needed	What would count as progress
Semantic and open-world noise	A formal model that jointly represents label validity, class membership, semantic proximity, and unknown-class contamination	The model distinguishes correction, calibrated ambiguity, rejection, and discovery under explicit observation budgets
Identifiability and assumptions	A recoverability checklist tied to clean anchors, repeated labels, priors, external representations, and label-space metadata	Papers state which latent object is identifiable and report sensitivity when the required information is removed
Representation diagnostics	Shared diagnostics for neighborhood stability, prototype drift, calibration, class separation, and clean/noisy separability over time	Benchmarks report when a method protects geometry rather than only improving top-line accuracy
Pipeline-level comparison	Ablation protocols that perturb isolation, refinement, representation, and schedule stages separately	Hybrid methods can identify the stage that creates robustness and the stage that fails under realistic noise
Foundation-model supervision	Provenance-aware datasets and audits for LLM labels, VLM priors, preference data, and noisy pre-training sources	Studies separate learning from noisy labels, learning from noisy annotators, and inheriting noise from pre-trained models

11 Limitations of This Review

This review has several limitations. First, it is a critical narrative review rather than a systematic review. The scope and synthesis protocol in Table 3 explain the selection logic but do not claim exhaustive coverage of every LNL paper or application domain. Second, the literature from 2025–2026 is moving quickly. We use recent peer-reviewed papers and preprints to identify emerging pressure points, but some bibliographic details and empirical conclusions may change before final publication.

Third, the review is biased toward English-language literature and papers from major AI, machine learning, computer vision, natural language processing, data mining, information retrieval, and related journal venues. Domain-specific work in medicine, remote sensing, speech, and industrial quality control may expose additional forms of noisy supervision that are not discussed in depth. Fourth, cross-paper empirical comparison remains difficult because reported results often differ in architectures, augmentation policies, clean validation access, noise-rate assumptions, and training

budgets. The tables in this review should therefore be read as assumption maps and representative evidence, not as a ranked benchmark leaderboard.

12 Conclusion

A central lesson from this review is that modern LNL is better understood as learning under structured supervision mismatch than as robust classification under random label flips. The relevant mismatch may involve an incorrect class, semantic ambiguity, an incomplete or changing ontology, open-set contamination, a noisy annotator or model, a corrupted correspondence, or a representation inherited from weak pre-training. These settings do not always contain a single recoverable clean label, so robustness must be defined relative to the latent target and information budget available in each problem.

The literature also shows that reliability is conditional rather than intrinsic to an example. Memorization dynamics, local geometry, class frequency, optimization schedules, and external priors determine whether clean and corrupted supervision become distinguishable. The small-loss case study makes this boundary concrete: loss is useful when training exposes structural separability, but it cannot serve as a universal correctness oracle. Robust methods should therefore be read through five interacting roles—noise characterization, reliable-signal isolation, target refinement, representation preservation, and optimization or uncertainty control—with their assumptions reported explicitly.

Finally, robustness claims are only as strong as the protocols used to test them. Realistic evaluation must describe how noise is generated, what clean information or external priors are available, which hard-clean or unknown examples are retained, and whether calibration, representation stability, transfer, or propagation effects matter in addition to final accuracy. The research agenda therefore extends beyond improving accuracy under a fixed synthetic noise rate. It requires identifiable formulations, observable diagnostics, provenance-aware benchmarks, and pipeline-level evaluations that explain when noisy supervision is harmful, when it remains informative, and how its effects propagate through models and domains.

Statements and Declarations

Funding This research was supported by Beijing Natural Science Foundation of China (no. 4262065, 4222037, L181010).

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Data availability No new dataset was created for this review. The datasets and published results discussed in the article are documented in the cited sources.

Author contributions Wenxiao Fan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. Kan

Li: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – review and editing.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the authors used OpenAI’s ChatGPT and Codex for language editing, organizational suggestions, citation checking, and preliminary figure concepts. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. All scholarly claims, literature interpretations, citations, data, and final figures were verified by the authors. No generative-AI-produced image is included in the final manuscript.

ORCID iDs Wenxiao Fan: 0000-0001-5200-9670; Kan Li: 0000-0003-3528-4739.

References

- Arpit D, Jastrzebski S, Ballas N, et al (2017) A closer look at memorization in deep networks. In: Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 70. PMLR, pp 233–242, URL <https://proceedings.mlr.press/v70/arpit17a.html>
- Baek C, Kolter JZ, Raghunathan A (2024) Why is SAM robust to label noise? In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=3aZCPl3ZvR>
- Cao X, Xu Z, Guang M, et al (2026) RE-PO: Robust enhanced policy optimization as a general framework for LLM alignment. In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=jDKpOvTCM8>
- Chang W, Shi Y, Wang J (2023) CSOT: Curriculum and structure-aware optimal transport for learning with noisy labels. In: Advances in Neural Information Processing Systems, pp 8528–8541, <https://doi.org/10.52202/075280-0373>, URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b0da24d136f46bfaee78e8da907127e-Abstract-Conference.html
- Chen H, Wang J, Shah AP, et al (2024) Understanding and mitigating the label noise in pre-training on downstream tasks. In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=TjhUtloBZU>
- Chen H, Wang Z, Tao R, et al (2025) Impact of noisy supervision in foundation model learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(7):5690–5707. <https://doi.org/10.1109/TPAMI.2025.3552309>, URL <https://ieeexplore.ieee.org/document/10934976>
- Chen L, Chern B, Eckstrand E, et al (2026) Enhancing learning with noisy labels via rockafellian relaxation. In: International Conference on Learning Representations, URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7e810b2c75d69be186cadd2fe3febeab-Abstract-Conference.html

- Chen M, Cheng H, Du Y, et al (2023) Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 14765–14773, <https://doi.org/10.1609/aaai.v37i12.26725>
- Chen P, Ye J, Chen G, et al (2021) Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11442–11450, <https://doi.org/10.1609/aaai.v35i13.17363>
- Choi D, Lee S, Yun E, et al (2025) ELDET: Early-learning distillation with noisy labels for object detection. In: Advances in Neural Information Processing Systems, pp 69345–69372, <https://doi.org/10.52202/085713-2331>, URL https://papers.neurips.cc/paper_files/paper/2025/hash/6460e378f24da3a79f20ac2640732a00-Abstract-Conference.html
- Cordeiro FR, Carneiro G (2020) A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, pp 9–16, <https://doi.org/10.1109/SIBGRAPI51738.2020.00010>
- Cordeiro FR, Carneiro G (2025) ANNE: Adaptive nearest neighbours and eigenvector-based sample selection for robust learning with noisy labels. Pattern Recognition 159:111132. <https://doi.org/10.1016/j.patcog.2024.111132>, URL <https://www.sciencedirect.com/science/article/pii/S0031320324008835>
- Feng C, Ren Y, Xie X (2023) OT-Filter: An optimal transport filter for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16164–16174, <https://doi.org/10.1109/CVPR52729.2023.01551>, URL https://openaccess.thecvf.com/content/CVPR2023/html/Feng_OT-Filter_An_Optimal_Transport_Filter_for_Learning_With_Noisy_Labels_CVPR_2023_paper.html
- Frenay B, Verleysen M (2014) Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems 25(5):845–869. <https://doi.org/10.1109/TNNLS.2013.2292894>
- Graham-Knight JB, Fayyad J, Bayasi N, et al (2025) Conformal-in-the-loop for learning with imbalanced noisy data. <https://doi.org/10.48550/arXiv.2411.02281>, URL <https://arxiv.org/abs/2411.02281>, arXiv:2411.02281
- Grinberg Y, Harel N, Goldberger J, et al (2025) Detect and correct: A selective noise correction method for learning with noisy labels. Computer Science Research Notes 3501(2025):173–181. <https://doi.org/10.24132/CSRN.2025-19>, URL <https://doi.org/10.24132/CSRN.2025-19>, 33rd International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG 2025)

- Han B, Yao Q, Yu X, et al (2018) Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems
- Han B, Yao Q, Liu T, et al (2020) A survey of label-noise representation learning: Past, present and future. <https://doi.org/10.48550/arXiv.2011.04406>, URL <https://arxiv.org/abs/2011.04406>, arXiv:2011.04406
- Han H, Miao K, Zheng Q, et al (2023) Noisy correspondence learning with meta similarity correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7517–7526, <https://doi.org/10.1109/CVPR52729.2023.00726>, URL https://openaccess.thecvf.com/content/CVPR2023/html/Han_Noisy_Correspondence_Learning_With_Meta_Similarity_Correction_CVPR_2023_paper.html
- Huang X, Sun Z, Shi J, et al (2026) FedRNC: Addressing spatio-temporal label misalignment in federated noisy class-incremental learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 22048–22056, <https://doi.org/10.1609/aaai.v40i26.39359>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/39359>
- Huang Z, Zhang J, Shan H (2023) Twin contrastive learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11661–11670, <https://doi.org/10.1109/CVPR52729.2023.01122>, URL https://openaccess.thecvf.com/content/CVPR2023/html/Huang_Twin_Contrastive_Learning_With_Noisy_Labels_CVPR_2023_paper.html
- Ji K, Cheng F, Wang Z, et al (2026) Jump-teaching: Combating sample selection bias via temporal disagreement. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 22191–22199, <https://doi.org/10.1609/aaai.v40i26.39375>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/39375>
- Jiang L, Zhou Z, Leung T, et al (2018) MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 80. PMLR, pp 2304–2313, URL <https://proceedings.mlr.press/v80/jiang18c.html>
- Jiang L, Huang D, Liu M, et al (2020) Beyond synthetic noise: Deep learning on controlled noisy labels. In: Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 119. PMLR, pp 4804–4815, URL <https://proceedings.mlr.press/v119/jiang20c.html>
- Kim J, Baratin A, Zhang Y, et al (2023) CrossSplit: Mitigating label noise memorization through data splitting. In: Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 202. PMLR, pp 16377–16392, URL <https://proceedings.mlr.press/v202/kim23a.html>

- Kim Nr, Lee JS, Lee JH (2024) Learning with structural labels for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 27610–27620, <https://doi.org/10.1109/CVPR52733.2024.02607>, URL https://openaccess.thecvf.com/content/CVPR2024/html/Kim_Learning_with_Structural_Labels_for_Learning_with_Noisy_Labels_CVPR_2024_paper.html
- Kim S, Kang SK, Kim D, et al (2025) Delving into instance-dependent label noise in graph data: A comprehensive study and benchmark. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, pp 5539–5550, <https://doi.org/10.1145/3711896.3737376>, URL <https://doi.org/10.1145/3711896.3737376>
- Lee KH, He X, Zhang L, et al (2018) CleanNet: Transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5447–5456, <https://doi.org/10.1109/CVPR.2018.00571>, URL https://openaccess.thecvf.com/content_cvpr_2018/html/Lee_CleanNet_Transfer_Learning_CVPR_2018_paper.html
- Li F, Li K, Wang Q, et al (2025a) RML++: Regroup median loss for combating label noise. International Journal of Computer Vision 133(9):6400–6421. <https://doi.org/10.1007/s11263-025-02494-4>, URL <https://doi.org/10.1007/s11263-025-02494-4>
- Li J, Socher R, Hoi SCH (2020) DivideMix: Learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=HJgExaVtwr>
- Li Q, Luo T, Liao J (2025b) Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 20706–20715, <https://doi.org/10.1109/CVPR52734.2025.01928>, URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Theory-Inspired_Deep_Multi-View_Multi-Label_Learning_with_Incomplete_Views_and_Noisy_CVPR_2025_paper.html
- Li Q, Li X, Li D, et al (2026a) Prototype-guided supervision for graph learning with noisy and sparse labels. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 23105–23113, <https://doi.org/10.1609/aaai.v40i27.39477>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/39477>
- Li W, Wang L, Li W, et al (2017) Webvision database: Visual learning and understanding from web data. [arXiv:1708.02862](https://arxiv.org/abs/1708.02862)
- Li W, Peng K, Wen D, et al (2026b) Can we trust unreliable voxels? exploring 3d semantic occupancy prediction under label noise. <https://doi.org/10.48550/arXiv.2603.06279>, URL <https://arxiv.org/abs/2603.06279>, accepted to IROS 2026, [arXiv:2603.06279](https://arxiv.org/abs/2603.06279)

- Li Y, Han H, Shan S, et al (2023) DISC: Learning from noisy labels via dynamic instance-specific selection and correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 24070–24079, <https://doi.org/10.1109/CVPR52729.2023.02305>, URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_DISC_Learning_From_Noisy_Labels_via_Dynamic_Instance-Specific_Selection_and_CVPR_2023_paper.html
- Liang C, Zhu L, Shi H, et al (2025) Combating label noise with a general surrogate model for sample selection. *International Journal of Computer Vision* 133(6):3166–3179. <https://doi.org/10.1007/s11263-024-02324-z>, URL <https://doi.org/10.1007/s11263-024-02324-z>
- Liu S, Niles-Weed J, Razavian N, et al (2020) Early-learning regularization prevents memorization of noisy labels. In: *Advances in Neural Information Processing Systems*, pp 20331–20342, URL <https://proceedings.neurips.cc/paper/2020/hash/ea89621bee7c88b2c5be6681c8ef4906-Abstract.html>
- Liu X, Li M, Lyu Z, et al (2026) Learning from noisy preferences: A semi-supervised learning approach to direct preference optimization. In: *International Conference on Learning Representations*, pp 13690–13709, URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/17061a94c3c7fda5fa24bbdd1832fa99-Abstract-Conference.html
- Liu Y, Cheng H, Zhang K (2023) Identifiability of label noise transition matrix. In: *Proceedings of the 40th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol 202. PMLR, pp 21475–21496, URL <https://proceedings.mlr.press/v202/liu23g.html>
- Lu Y, Zhang Y, Han B, et al (2023) Label-noise learning with intrinsically long-tailed data. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp 1369–1378, <https://doi.org/10.1109/ICCV51070.2023.00132>, URL https://openaccess.thecvf.com/content/ICCV2023/html/Lu_Label-Noise_Learning_with_Intrinsically_Long-Tailed_Data_ICCV_2023_paper.html
- Natarajan N, Dhillon IS, Ravikumar PK, et al (2013) Learning with noisy labels. In: *Advances in Neural Information Processing Systems*, pp 1196–1204, URL <https://proceedings.neurips.cc/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html>
- Nazaretyan L, Leser U, Kircher M (2025) Benchmarking machine learning methods for the identification of mislabeled data. *Artificial Intelligence Review* 58(10):310. <https://doi.org/10.1007/s10462-025-11293-9>, URL <https://link.springer.com/article/10.1007/s10462-025-11293-9>
- Ning F, Chen X (2026) Mitigating endogenous confirmation bias in noisy label learning for vision-language models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 24576–24584, <https://doi.org/10.1609/aaai.v40i29.39641>, URL <https://doi.org/10.1609/aaai.v40i29.39641>

<https://ojs.aaai.org/index.php/AAAI/article/view/39641>

- Pan L, Gao C, Zhou J, et al (2025) Learning with open-world noisy data via class-independent margin in dual representation space. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 6290–6298, <https://doi.org/10.1609/aaai.v39i6.32673>
- Patrini G, Rozza A, Menon AK, et al (2017) Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2233–2241, <https://doi.org/10.1109/CVPR.2017.240>, URL <https://doi.org/10.1109/CVPR.2017.240>
- Ren M, Zeng W, Yang B, et al (2018) Learning to reweight examples for robust deep learning. In: Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 80. PMLR, pp 4334–4343, URL <https://proceedings.mlr.press/v80/ren18a.html>
- Shi J, Zhang K, Guo C, et al (2024) A survey of label-noise deep learning for medical image analysis. Medical Image Analysis 95:103166. <https://doi.org/10.1016/j.media.2024.103166>
- Shi Y, Li P, Zhang Z, et al (2026) Conformal margin risk minimization: An envelope framework for robust learning under label noise. In: Proceedings of The 29th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 300. PMLR, pp 4276–4284, URL <https://proceedings.mlr.press/v300/shi26b.html>, [arXiv:2604.06468](https://arxiv.org/abs/2604.06468)
- Shin J, Won J, Lee HS, et al (2024) A review on label cleaning techniques for learning with noisy labels. ICT Express 10(6):1315–1330. <https://doi.org/10.1016/j.icte.2024.09.007>
- Song B, Zhao S, Dang L, et al (2025) A survey on learning from data with label noise via deep neural networks. Systems Science & Control Engineering 13(1):2488120. <https://doi.org/10.1080/21642583.2025.2488120>
- Song H, Kim M, Lee JG (2019) SELFIE: Refurbishing unclean samples for robust deep learning. In: Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 97. PMLR, pp 5907–5915
- Song H, Kim M, Park D, et al (2023) Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems 34(11):8135–8153. <https://doi.org/10.1109/TNNLS.2022.3152527>
- Sosea T, Caragea C (2026) LANE: Label-aware noise elimination for fine-grained text classification. In: International Conference on Learning Representations, URL <https://openreview.net/forum?id=PuayLKdrFz>

- Sun H, Wei Q, Feng L, et al (2025) Variational rectification inference for learning with noisy labels. *International Journal of Computer Vision* 133(2):652–671. <https://doi.org/10.1007/s11263-024-02205-5>, URL <https://doi.org/10.1007/s11263-024-02205-5>
- Sun Z, Yao Y, Wei XS, et al (2021) Webly supervised fine-grained recognition: Benchmark datasets and an approach. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp 10602–10611, <https://doi.org/10.1109/ICCV48922.2021.01043>, URL https://openaccess.thecvf.com/content/ICCV2021/html/Sun_Webly_Supervised_Fine-Grained_Recognition_Benchmark_Datasets_and_an_Approach_ICCV_2021_paper.html
- Wan W, Wang X, Xie MK, et al (2024) Unlocking the power of open set: A new perspective for open-set noisy label learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 15438–15446, <https://doi.org/10.1609/aaai.v38i14.29469>
- Wang J, Liu X, Zhou X, et al (2025) Joint asymmetric loss for learning with noisy labels. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp 1947–1956, <https://doi.org/10.1109/ICCV51701.2025.00189>, URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Joint_Asymmetric_Loss_for_Learning_with_Noisy_Labels_ICCV_2025_paper.html
- Wang S, Liu A, Plummer BA (2026) Noise-aware generalization: Robustness to in-domain noise and out-of-domain generalization. In: *International Conference on Learning Representations*, URL <https://openreview.net/forum?id=wb83wO41QT>
- Wei H, Feng L, Chen X, et al (2020) Combating noisy labels by agreement: A joint training method with co-regularization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 13726–13735, <https://doi.org/10.1109/CVPR42600.2020.01374>, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wei_Combating_Noisy_Labels_by_Agreement_A_Joint_Training_Method_with_CVPR_2020_paper.html
- Wei H, Zhuang H, Xie R, et al (2023a) Mitigating memorization of noisy labels by clipping the model prediction. In: *Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol 202. PMLR, pp 36868–36886, URL <https://proceedings.mlr.press/v202/wei23e.html>
- Wei J, Zhu Z, Cheng H, et al (2022) Learning with noisy labels revisited: A study using real-world human annotations. In: *International Conference on Learning Representations*, URL <https://openreview.net/forum?id=TBWA6PLJZQm>
- Wei Q, Feng L, Sun H, et al (2023b) Fine-grained classification with noisy labels. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 11651–11660, <https://doi.org/10.1109/CVPR52729.2023.01121>, URL https://openaccess.thecvf.com/content/CVPR2023/html/Wei_Fine-Grained_Classification_With_Noisy_Labels_CVPR_2023_paper.html

- Wei T, Li HT, Li CS, et al (2024) Vision-language models are strong noisy label detectors. In: Advances in Neural Information Processing Systems, pp 58154–58173, <https://doi.org/10.52202/079017-1854>, URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6af08ba9468f0daca4b8dd388cb95824-Abstract-Conference.html
- Wu N, Yu L, Jiang X, et al (2023a) FedNoRo: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, pp 4424–4432, <https://doi.org/10.24963/ijcai.2023/492>, URL <https://www.ijcai.org/proceedings/2023/492>
- Wu T, Ding X, Tang M, et al (2023b) NoisywikiHow: A benchmark for learning with real-world noisy labels in natural language processing. In: Findings of the Association for Computational Linguistics: ACL 2023, pp 4856–4873, <https://doi.org/10.18653/v1/2023.findings-acl.299>, URL <https://aclanthology.org/2023.findings-acl.299/>
- Xia X, Liu T, Han B, et al (2020) Part-dependent label noise: Towards instance-dependent label noise. In: Advances in Neural Information Processing Systems, pp 7597–7610, URL <https://proceedings.neurips.cc/paper/2020/hash/5607fe8879e4fd269e88387e8cb30b7e-Abstract.html>
- Xia X, Han B, Wang N, et al (2023) Extended T: Learning with mixed closed-set and open-set noisy labels. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(3):3047–3058. <https://doi.org/10.1109/TPAMI.2022.3180545>
- Xiao T, Xia T, Yang Y, et al (2015) Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2691–2699, <https://doi.org/10.1109/CVPR.2015.7298885>
- Xu J, Pang J (2026) NCSAM noise-compensated sharpness-aware minimization for noisy label learning. <https://doi.org/10.48550/arXiv.2601.19947>, URL <https://arxiv.org/abs/2601.19947>, arXiv:2601.19947
- Yao Y, Gong M, Du Y, et al (2023) Which is better for learning with noisy labels: The semi-supervised method or modeling label noise? In: Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 202. PMLR, pp 39660–39673, URL <https://proceedings.mlr.press/v202/yao23a.html>
- Ye L, Shah A, Zhang C, et al (2025) Calibrating pre-trained language classifiers on LLM-generated noisy labels via iterative refinement. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. Association for Computing Machinery, pp 3598–3609, <https://doi.org/10.1145/3711896.3736871>

- Ye X, Li X, Dai S, et al (2023) Active negative loss functions for learning with noisy labels. In: Advances in Neural Information Processing Systems, pp 6917–6940, <https://doi.org/10.52202/075280-0303>, URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/15f4cefb0e143c7ad9d40e879b0a9d0c-Abstract-Conference.html
- Yuan S, Feng L, Han B, et al (2025) Enhancing sample selection against label noise by cutting mislabeled easy examples. In: Advances in Neural Information Processing Systems, pp 49163–49191, <https://doi.org/10.52202/085713-1466>, URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3e826f682178d9830a3b704141b4989e-Abstract-Conference.html
- Zhang J, Song B, Wang H, et al (2024) BadLabel: A robust perspective on evaluating and enhancing label-noise learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(6):4398–4409. <https://doi.org/10.1109/TPAMI.2024.3355425>
- Zhang K, Chai C, Xu J, et al (2025a) Handling label noise via instance-level difficulty modeling and dynamic optimization. In: Advances in Neural Information Processing Systems, <https://doi.org/10.52202/085713-1553>, URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/429e7b31625a8b7839f9e4d6e2aa9bb9-Abstract-Conference.html
- Zhang T, Zhou Z, Wang M, et al (2026) On the learning dynamics of two-layer linear networks with label noise SGD. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 28400–28408, <https://doi.org/10.1609/aaai.v40i33.40069>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/40069>
- Zhang Y, Chen Y, Fang C, et al (2025b) Learning from open-set noisy labels based on multi-prototype modeling. Pattern Recognition 157:110902. <https://doi.org/10.1016/j.patcog.2024.110902>, URL <https://www.sciencedirect.com/science/article/pii/S0031320324006538>
- Zhang Z, Sabuncu MR (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in Neural Information Processing Systems, pp 8778–8788, URL <https://proceedings.neurips.cc/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html>
- Zhang Z, Chen W, Fang C, et al (2023) RankMatch: Fostering confidence and consistency in learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1644–1654, <https://doi.org/10.1109/ICCV51070.2023.00158>, URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_RankMatch_Fostering_Confidence_and_Consistency_in_Learning_with_Noisy_Labels_ICCV_2023_paper.html
- Zhou Y, Li X, Liu F, et al (2024) L2B: Learning to bootstrap robust models for combating label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 23523–23533,

<https://doi.org/10.1109/CVPR52733.2024.02220>, URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_L2B_Learning_to_Bootstrap_Robust_Models_for_Combating_Label_Noise_CVPR_2024_paper.html

Zong CC, Wang YW, Xie MK, et al (2024) Dirichlet-based prediction calibration for learning with noisy labels. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 17254–17262, <https://doi.org/10.1609/aaai.v38i15.29672>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table18.tex](#)
- [Table17.tex](#)
- [Table16.tex](#)
- [Table15.tex](#)
- [Table07.tex](#)
- [Table20.tex](#)
- [Table12.tex](#)
- [Table13.tex](#)
- [Table08.tex](#)
- [Table10.tex](#)
- [Table11.tex](#)
- [Table06.tex](#)
- [Table19.tex](#)
- [Table09.tex](#)
- [Table01.tex](#)
- [Table04.tex](#)
- [Table03.tex](#)
- [SupplementaryMaterial.pdf](#)