Novelty Assessment Report

Paper: Scalable Oversight via Partitioned Human Supervision

PDF URL: https://openreview.net/pdf?id=IoxBlRhANV

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

As artificial intelligence (AI) systems approach and surpass expert human performance across a broad range of tasks, obtaining high-quality human supervision for evaluation and training becomes increasingly challenging. Our focus is on tasks that require deep knowledge and skills of multiple domains. Unfortunately, even the best human experts are knowledgeable only in a single narrow area, and will not be able to evaluate the correctness of advanced AI systems on such superhuman tasks. However, based on their narrow expertise, humans may provide a weak signal, i.e., a complementary label indicating an option that is incorrect. For example, a cardiologist could state that ``this is not related to cardiology,'' even if they cannot identify the true disease. Based on this weak signal, we propose a scalable oversight framework that enables us to evaluate frontier AI systems without the need to prepare the ground truth. We derive an unbiased estimator of top-1 accuracy from complementary labels and quantify how many complementary labels are needed to match the variance of ordinary labels. We further introduce two estimators to combine scarce ordinary labels with abundant complementary labels. We provide finite-sample deviation guarantees for both complementary-only and the mixed estimators. Empirically, we show that we can evaluate the output of large language models without the ground truth, if we have complementary labels. We further show that we can train an AI system with such weak signals: we show how we can design an agentic AI system automatically that can perform better by these partitioned human supervision.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Evaluating and Training AI Systems Using Complementary Labels from Domain Experts**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Complementary-Label Learning Frameworks and Theory**
- **Medical Imaging and Diagnostic AI Evaluation**
- **Clinical Decision Support and Language Model Evaluation**
- **General AI Evaluation and Annotation Quality**
- **Domain-Specific Applications and Specialized Tasks**

Complete Taxonomy Tree

- Evaluating and Training AI Systems Using Complementary Labels from Domain Experts Survey Taxonomy
- Complementary-Label Learning Frameworks and Theory
  - Scalable Oversight and Unbiased Estimation ★ (1 papers)
  - [0] Scalable Oversight via Partitioned Human Supervision (Anon et al., 2026) View paper
  - Complementary-Label Datasets and Benchmarks (1 papers)
  - [44] CLImage: Human-Annotated Datasets for Complementary-Label Learning (Wang, 2023) View paper
  - Weak and Multi-Source Supervision Integration (2 papers)
  - [39] Recycling weak labels for multiclass classification (Miquel PerellÃ³-Nieto, 2020) View paper
  - [48] Augmenting Document-level Relation Extraction with Efficient Multi-Supervision (Lin Xiangyu, 2024) View paper
- Medical Imaging and Diagnostic AI Evaluation
  - Radiology and Chest X-Ray Analysis (4 papers)
  - [17] A Generative Foundation Model for Chest Radiography (JI YUANFENG, 2025) View paper
  - [29] CheXseg: combining expert annotations with DNN-generated saliency maps for X-ray segmentation (Gadgil, 2021) View paper
  - [35] Reasoning Visual Language Model for Chest X-Ray Analysis (Myronenko, 2025) View paper
  - [40] Impact of hybrid supervision approaches on the performance of artificial intelligence for the classification of chest radiographs (Ryan Ellis, 2020) View paper
  - Pathology and Histology Image Analysis (2 papers)
  - [9] A study of criteria for grading follicular lymphoma using a cell type classifier from pathology images based on complementary-label learning (R. Koga, 2024) View paper
  - [46] Machine-learning for quantitative histopathology of piglet intestinal tissues: challenges with limited training data (Cecilie Brandt Becker, 2025) View paper
  - Ophthalmology and Specialized Imaging (2 papers)
  - [4] Automated ultrasound system ARTHUR V.2.0 with AI analysis DIANA V.2.0 matches expert rheumatologist in hand joint assessment of rheumatoid arthritis patients (Bill Aplin Frederiksen, 2025) View paper
  - [41] Synergistic AI-resident approach achieves superior diagnostic accuracy in tertiary ophthalmic care for glaucoma and retinal disease (Dalia Camacho-GarcÃa-FormentÃ, 2025) View paper
  - Multimodal Medical Imaging and Foundation Models (3 papers)
  - [18] Improving medical multi-modal contrastive learning with expert annotations (Kumar Yogesh, 2024) View paper

  - [33] Semi-Supervised Learning in Prostate MRI Tumor Segmentation Approaches Fully-Supervised Performance on External Validation (E. Pooch, 2025) View paper
  - [42] A semi-automated quality assurance tool for cardiovascular magnetic resonance imaging: application to outlier detection, artificial intelligence evaluation and trainee â□¦ (T Hadler, 2025) View paper
  - Expert-Guided Explainability and Attention Supervision (1 papers)
  - [16] Expert-guided explainable few-shot learning for medical image diagnosis (Wang Long-wei, 2025) View paper
- Clinical Decision Support and Language Model Evaluation
  - Medical Question Answering and Exam Performance (2 papers)
  - [3] Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders (Olena Bolgova, 2025) View paper
  - [11] Expert of experts verification and alignment (eval) framework for large language models safety in gastroenterology (Mauro GiuffrÃ¨, 2025) View paper
  - Clinical Task Annotation and Real-World Complexity (2 papers)
  - [8] Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare (Lamparth, 2025) View paper
  - [38] Unlocking new frontiers in leukemia diagnostics through large language Modelâ□□Driven report generation (Vivian Wuerf, 2025) View paper
  - Safety and Harm Assessment in Clinical AI (2 papers)
  - [37] MATRIX: Multi-Agent simulaTion fRamework for safe Interactions and conteXtual clinical conversational evaluation (Lim, 2025) View paper
  - [50] First, do NOHARM: towards clinically safe large language models (David Wu, 2025) View paper
  - Human-AI Collaborative Decision Making (1 papers)
  - [13] A human-ai collaborative approach for clinical decision making on rehabilitation assessment (Min Hun Lee, 2021) View paper
- General AI Evaluation and Annotation Quality
  - Expert Annotation Consistency and Reliability (3 papers)
  - [5] Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what (Sarah Lebovitz, 2021) View paper
  - [14] A case for moving beyond â□□gold dataâ□□ in AI safety evaluation (D Wang, 2024) View paper
  - [26] The impact of inconsistent human annotations on AI driven clinical decision making (Aneeta Sylolypavan, 2023) View paper
  - Multi-Annotator Learning and Expertise Modeling (2 papers)
  - [22] Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis (Khiem H. Le, 2023) View paper
  - [36] Modeling annotator expertise: Learning when everybody knows a bit of something (YAN Yan, 2010) View paper
  - AI-Assisted Annotation and Human-in-the-Loop (4 papers)
  - [19] Agent-in-the-loop to distill expert knowledge into artificial intelligence models: a survey (Jiayuan Gao, 2025) View paper
  - [20] Human supervision is key to achieving accurate AI-assisted wildlife identifications in camera trap images (Sarah E. Huebner, 2024) View paper
  - [24] Man and the machine: Effects of AI-assisted human labeling on interactive annotation of real-time video streams (Marko Radeta, 2024) View paper
  - [25] Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists (Adrian Krenzer, 2022) View paper
  - Benchmark Development and Evaluation Frameworks (4 papers)
  - [12] FindTheFlaws: Annotated Errors for Detecting Flawed Reasoning and Scalable Oversight Research (Recchia, 2025) View paper
  - [15] 11plus-bench: Demystifying multimodal llm spatial reasoning with cognitive-inspired analysis (Li, 2025) View paper
  - [30] Advancing AI Factuality via Comprehensive Evaluation (Song, 2025) View paper
  - [34] 2025 Expert consensus on retrospective evaluation of large language model applications in clinical scenarios (Zhenchang Wang, 2025) View paper
  - Expert Definition and Role Characterization (2 papers)
  - [27] What Makes An Expert? Reviewing How ML Researchers Define" Expert" (Mark Diaz, 2024) View paper
  - [28] Enabling Domain Expert Evaluation of Emerging AI Technologies in Healthcare Settings (Dhar, 2024) View paper
- Domain-Specific Applications and Specialized Tasks
  - Multilingual and Cross-Lingual Text Annotation (2 papers)
  - [2] Machine learning tools match physician accuracy in multilingual text annotation (Marta Zielonka, 2025) View paper
  - [23] Building Foundations for Inclusiveness through Expert-Annotated Data. (M La Quatra, 2024) View paper
  - Computer Vision for Non-Medical Domains (2 papers)
  - [7] Weakly supervised visible-infrared person re-identification via heterogeneous expert collaborative consistency learning (Zhang Yafei, 2025) View paper
  - [43] Exploring Diagnostic Prompting Approach for Multimodal LLM-based Visual Complexity Assessment: A Case Study of Amazon Search Result Pages (Divendar Murtadak, 2025) View paper
  - Surgical Skill Assessment and Simulation (1 papers)
  - [10] Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation (A. Winkler-Schwartz, 2019) View paper
  - Measurement Error Correction and Robustness (2 papers)
  - [32] Correcting the Measurement Errors of AI-Assisted Labeling in Image Analysis Using Design-Based Supervised Learning (Alessandra Rister Portinari Maranca, 2025) View paper
  - [45] Robustness of human vs. AI measurements under progressive image degradation (J. Jevsikov, 2025) View paper
  - Prompt Engineering and LLM Evaluation Methodology (2 papers)
  - [47] A Multi-faceted Analysis of Cognitive Abilities: Evaluating Prompt Methods with Large Language Models on the CONSORT Checklist (Lee Hyung Chul, 2025) View paper
  - [49] Human Experts vs. Large Language Models: Evaluating Annotation Scheme and Guidelines Development for Clinical (ALC Fernandes, 2025) View paper
  - Cross-Domain AI Foundations and Supervision Paradigms (4 papers)
  - [1] Scientific discovery in the age of artificial intelligence (Hanchen Wang, 2023) View paper

- [6] Integrating Generative AI into Circular Supply Chain Safety Management: A Forward-Looking Perspective (Zhuowen Chen, 2025) View paper
- [21] Performance Evaluation of Cloud Native Applications: A Systematic Mapping Study (Francisco Airton Silva, 2025) View paper
- [31] 26 Evaluating AI models trained with varying amounts of expert feedback for chronic graft-versus-host disease skin assessment in photos of patients with diverse skin tones (Andrew McNeil, 2025) View paper

Narrative

Core task: evaluating and training AI systems using complementary labels from domain experts. The field organizes around five main branches that reflect different facets of expert-driven supervision. Complementary-Label Learning Frameworks and Theory develops formal methods for learning from partial or indirect annotations, addressing scalability and bias when experts cannot label every instance exhaustively. Medical Imaging and Diagnostic AI Evaluation focuses on radiology, pathology, and clinical imaging tasks where expert radiologists or pathologists provide ground truth or quality assessments, as seen in works like ARTHUR DIANA Ultrasound[4] and Follicular Lymphoma Grading[9]. Clinical Decision Support and Language Model Evaluation examines how large language models can be assessed or guided by clinicians, for instance through LLM Medical Graders[3] or consensus protocols like LLM Clinical Consensus[34]. General AI Evaluation and Annotation Quality tackles cross-domain challenges such as handling inconsistent annotations (Inconsistent Human Annotations[26]), defining what constitutes expertise (Defining Expert[27]), and ensuring robust evaluation pipelines (AI Ground Truth[5]). Domain-Specific Applications and Specialized Tasks spans diverse settings from wildlife monitoring (Wildlife Camera Traps[20]) to surgical skill assessment (Surgical Expertise Assessment[10]), illustrating how expert feedback adapts to varied problem contexts.

Several active lines of work explore trade-offs between annotation cost, label quality, and model performance. One recurring theme is how to aggregate or reconcile disagreements among multiple experts (Multiple Expert Annotators[22], Heterogeneous Expert Consistency[7]) while preserving the nuanced information that partial or complementary labels provide. Another thread investigates hybrid supervision strategies that combine weak labels, semi-automated tools, and targeted expert input (Hybrid Supervision Radiographs[40], Semi-Automated Quality Assurance[42]). Partitioned Human Supervision[0] sits within the Scalable Oversight and Unbiased Estimation cluster, emphasizing methods that partition the supervision task to reduce expert burden while maintaining unbiased learning guarantees. This approach contrasts with works like Recycling Weak Labels[39], which reuse noisy annotations more opportunistically, and aligns closely with efforts such as Measurement Error Correction[32] that formally account for imperfect or incomplete expert signals. The central question across these directions is how to design training and evaluation protocols that respect expert time constraints without sacrificing the reliability or interpretability that domain knowledge brings.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on unbiased estimation and scalable oversight methods using complementary labels from partitioned domain experts, emphasizing theoretical guarantees and practical supervision at scale. The sibling subtopics address orthogonal aspects: one concentrates on dataset creation and benchmarking infrastructure for complementary-label research, while the other explores integration with additional weak supervision signals beyond pure complementary labels. Together, these subtopics partition the complementary-label learning space by methodology (pure vs. hybrid), resource type (algorithms vs. datasets), and supervision complexity.

Similarities: - All subtopics operate within the complementary-label learning paradigm where annotations specify what a sample is NOT rather than what it IS - Share the common goal of enabling learning when traditional full supervision is expensive or impractical - Address challenges in leveraging domain expert knowledge for AI system training - Exclude domain-specific applications in favor of methodological or infrastructural contributions

Differences: - Original leaf emphasizes unbiased accuracy estimation and theoretical properties of complementary-label methods, while siblings focus on practical resources (datasets) or methodological extensions (multi-source integration) - Original leaf targets scalable oversight with partitioned experts as a specific use case, whereas Weak and Multi-Source Integration combines complementary labels with other supervision paradigms like noisy or distant labels - Complementary-Label Datasets subtopic provides evaluation infrastructure rather than learning algorithms - Original leaf excludes multi-label and weak supervision integration (explicitly deferred to siblings), while Weak and Multi-Source Integration makes this integration its core focus

Suggested Search Directions: - Theoretical analysis of bias-variance tradeoffs when combining complementary labels with other weak supervision sources - Benchmark studies comparing pure complementary-label methods against hybrid weak supervision approaches - Scalability analysis of complementary-label methods on large-scale human-annotated datasets - Expert partitioning strategies and their impact on unbiased estimation guarantees

Sibling Subtopics

- **Complementary-Label Datasets and Benchmarks** (leaves: 1, papers: 1)
- Scope: Creation and analysis of human-annotated datasets specifically designed for complementary-label learning research.
- Exclude: Excludes datasets for other weak supervision paradigms or domain-specific annotation tools; see respective application categories.
- **Weak and Multi-Source Supervision Integration** (leaves: 1, papers: 2)
- Scope: Frameworks combining complementary labels with other weak supervision signals such as noisy labels or distantly supervised data.
- Exclude: Excludes pure complementary-label methods or domain-specific applications; see Scalable Oversight or Medical Imaging categories.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Scalable oversight framework via partitioned human supervision using complementary labels

Description: The authors introduce a framework that exploits partitioned human expertise to collect complementary labels (indicating incorrect options) at scale for superhuman tasks. This enables evaluation and training of AI systems without requiring full ground truth or comprehensive expert verification.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Beyond Manual Annotation: A Human-AI Collaborative Framework for Medical Image Segmentation Using Only â Better or Worseâ Expert Feedback

URL: View paper

Brief Assessment

Better or Worse Feedback[52] focuses on medical image segmentation using binary preference feedback ('better or worse') to guide an interactive clicking agent, not on partitioned human supervision with complementary labels for evaluating superhuman AI tasks across multiple domains.

2. A Human-Centric Assessment Framework for AI

URL: View paper

Brief Assessment

Human-Centric AI Assessment[53] focuses on a Turing-test-inspired framework where domain experts evaluate AI solutions without knowing the solver's identity. It does not address partitioned human supervision, complementary labels, or scalable oversight for superhuman tasks as defined in the original paper.

3. Federated benchmarking of medical artificial intelligence with MedPerf

URL: View paper

Brief Assessment

MedPerf[51] focuses on federated benchmarking infrastructure for medical AI models across institutions, not on partitioned human supervision or complementary labeling methods for scalable oversight.

Contribution 2: Unbiased estimator of top-1 accuracy from complementary labels with variance analysis and mixture estimators

Description: The authors derive an unbiased linear correction estimator for accuracy using only complementary labels, analyze its variance properties, and propose two mixture estimators (inverse-variance weighted and maximum-likelihood) that combine ordinary and complementary labels with finite-sample deviation guarantees.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels

URL: View paper

Prior Art Analysis

Unbiased Risk Estimators[54] demonstrates that unbiased estimators for complementary labels were previously established in the literature. The candidate paper explicitly derives an unbiased risk estimator for complementary labels (Proposition 1 and the uniform assumption formulation in Equation 5), which directly corresponds to the linear correction estimator claimed as novel in the original paper. Furthermore, the candidate analyzes variance properties of complementary label estimators and discusses the relationship between variance and sample size requirements, covering the same theoretical ground as the original contribution. The candidate also addresses mixture approaches by discussing correction methods, though not in the exact same framework as the original paper's inverse-variance weighted and maximum-likelihood estimators.

Evidence

Evidence 1 - **Rationale**: This pair shows that the candidate paper already derived an unbiased risk estimator for complementary labels, which is the core of the original paper's first contribution. The candidate's Proposition 1 establishes the mathematical foundation for unbiased estimation from complementary labels. - **Original**: we derive anunbiasedestimator of top-1 accuracy from complementary labels and quantify how many complementary labels are needed to match the variance of ordinary labels. - **Candidate**: proposition 1. the ordinary risk can be transformed as r(g; ℓ) = e(x,y)~d[e⊤ y (t-1)ℓ(g(x))]. (2) that is, we obtain an unbiased risk estimator (ure): r(g; ℓ) = e(x,y)~d[ℓ(y,g(x))] = r(g; ℓ) (3) where ℓis the following rewritten loss: ℓ(y,g(x)) = e⊤ y (t-1)ℓ(g(x)). (4)

Evidence 2 - **Rationale**: Both papers derive the same linear correction formula for the uniform complementary label case. The candidate's Equation 5 is mathematically equivalent to the original's Corollary 1, showing that the specific form of the unbiased estimator was already known. - **Original**: corollary 1.under the assumption in eq. (1), the estimator bacomp = (k-1)ˆq-(k-2)(2) is unbiased fora, wherek≥3is the number of choices. - **Candidate**: by plugging in the uniform assumption t = 1 k-1 (1k -ik), we have the following formulation of ℓ, ℓ(y,g(x)) = -(k-1)ℓ(y,g(x)) + k∑ j=1 ℓ(j,g(x)). (5)

Evidence 3 - **Rationale**: While the candidate focuses on gradient analysis rather than direct variance formulas for accuracy estimation, it establishes the theoretical framework for analyzing properties of unbiased estimators from complementary labels, including their statistical behavior. This shows prior work on analyzing the quality and properties of complementary label estimators. - **Original**: sincewis 1here "weakly-correct" means that the model prediction isnotequal to the complementary label. for example, whenk= 4, avoiding the complementary label still leaves a 2 3 chance of being wrong. under review as a conference paper at iclr 2026 bernoulli with meanq= a+k-2 k-1 , we obtain var(ba... - **Candidate**: proposition 3. the gradient of an unbiased risk estimator is unbiased to the ordinary risk gradient. that is, for an instance (x,y) we have, ey|y [∇θℓ(y,g(x))] = ∇θℓ(y,g(x)) (14) thus, the gradient of the complementary loss ℓis unbiased with respect to the gradient of the ordinary loss

Evidence 4 - **Rationale**: The candidate discusses methods for combining and correcting complementary label estimators, showing that the concept of mixing or correcting complementary label-based estimates was already explored in prior work, though with different specific techniques than the original paper's IVW and ML estimators. - **Original**: we further introduce two estimators to combine scarce ordinary labels with abundant complementary labels. we provide finite-sample deviation guarantees for both complementary-only and the mixed estimators. - **Candidate**: risk correction methods: ishida et al. (2019) proposed two correction methods to mitigate the problem. first, the non-negative loss correction (nn), which enforces nonnegativity to the decomposed risk of each class. second, namely the gradient ascent correction (ga) which enforces a reverse gradient...

Contribution 3: Demonstration of evaluation and agentic training using complementary labels

Description: The authors empirically validate that their estimators enable both evaluation of large language models without ground truth and training of agentic AI systems by using complementary labels as fitness signals in agent search pipelines, demonstrating improved downstream performance.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data

URL: View paper

Brief Assessment

Candidate Pseudolabel Learning[59] focuses on fine-tuning vision-language models using candidate pseudolabels (sets of potential labels) for image classification tasks, not on evaluating or training language models with complementary labels (labels indicating what is NOT correct) as fitness signals in agent search pipelines.

2. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents: V. Oliveira et al.

URL: View paper

Brief Assessment

Weak Supervision Legal NER[65] focuses on named entity recognition in legal documents using weak supervision methods for labeling, not on evaluating or training language models with complementary labels as fitness signals in agent search pipelines.

3. Cold-start active learning through self-supervised language modeling

URL: View paper

Brief Assessment

Cold-Start Active Learning[61] focuses on active learning for text classification using self-supervised language modeling loss, not on evaluation or training methods using complementary labels as fitness signals for agentic AI systems.

4. Language models in the loop: Incorporating prompting into weak supervision

URL: View paper

Brief Assessment

Language Models Loop[63] focuses on using language models as labeling functions within a weak supervision framework (Snorkel) to create classifiers with limited labeled data. This differs from the original paper's use of complementary labels for evaluating and training agentic AI systems without ground truth in superhuman tasks.

5. Pre-Trained Vision-Language Models as Noisy Partial Annotators

URL: View paper

Brief Assessment

Noisy Partial Annotators[62] focuses on using pre-trained vision-language models (CLIP) to generate noisy partial labels for image classification, not on evaluating or training language models with complementary labels as fitness signals in agent search pipelines. The candidate addresses a different problem domain (computer vision with CLIP annotations) rather than agentic AI systems or LLM evaluation.

6. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision

URL: View paper

Brief Assessment

Weak-to-Strong Generalization[56] focuses on using weak model supervision to elicit capabilities from stronger models in standard supervised learning settings, not on complementary label-based evaluation or agentic training frameworks.

7. Difference-Complementary Learning and Label Reassignment for Multimodal Semi-Supervised Semantic Segmentation of Remote Sensing Images

URL: View paper

Brief Assessment

Difference-Complementary Learning[58] focuses on multimodal remote sensing image segmentation using optical and SAR data fusion with semi-supervised learning. It does not address evaluation or training of language models using complementary labels without ground truth, which is the core contribution of the original paper.

8. Pre-trained Vision-Language Models Assisted Noisy Partial Label Learning

URL: View paper

Brief Assessment

Noisy Partial Label[60] focuses on learning from noisy partial labels (candidate label sets) annotated by pre-trained vision-language models for image classification tasks, not on evaluating or training language models using complementary labels as fitness signals in agent search pipelines.

9. Learning with biased complementary labels

URL: View paper

Brief Assessment

Biased Complementary Labels[64] focuses on learning classifiers from biased complementary labels in traditional supervised learning settings, not on evaluating or training agentic AI systems. The paper addresses multi-class classification with biased transition probabilities but does not discuss LLM evaluation, agent search pipelines, or agentic workflows.

10. Simvlm: Simple visual language model pretraining with weak supervision

URL: View paper

Brief Assessment

SimVLM[57] focuses on vision-language pretraining with weak supervision from noisy image-text pairs, not on complementary label-based evaluation or agentic training frameworks.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Scalable Oversight via Partitioned Human Supervision View paper
- [1] Scientific discovery in the age of artificial intelligence View paper
- [2] Machine learning tools match physician accuracy in multilingual text annotation View paper
- [3] Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders View paper
- [4] Automated ultrasound system ARTHUR V.2.0 with AI analysis DIANA V.2.0 matches expert rheumatologist in hand joint assessment of rheumatoid arthritis patients View paper
- [5] Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what View paper
- [6] Integrating Generative AI into Circular Supply Chain Safety Management: A Forward-Looking Perspective View paper
- [7] Weakly supervised visible-infrared person re-identification via heterogeneous expert collaborative consistency learning View paper

- [8] Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare View paper
- [9] A study of criteria for grading follicular lymphoma using a cell type classifier from pathology images based on complementary-label learning View paper
- [10] Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation View paper
- [11] Expert of experts verification and alignment (eval) framework for large language models safety in gastroenterology View paper
- [12] FindTheFlaws: Annotated Errors for Detecting Flawed Reasoning and Scalable Oversight Research View paper
- [13] A human-ai collaborative approach for clinical decision making on rehabilitation assessment View paper
- [14] A case for moving beyond â□□gold dataâ□□ in AI safety evaluation View paper
- [15] 11plus-bench: Demystifying multimodal llm spatial reasoning with cognitive-inspired analysis View paper
- [16] Expert-guided explainable few-shot learning for medical image diagnosis View paper
- [17] A Generative Foundation Model for Chest Radiography View paper
- [18] Improving medical multi-modal contrastive learning with expert annotations View paper
- [19] Agent-in-the-loop to distill expert knowledge into artificial intelligence models: a survey View paper
- [20] Human supervision is key to achieving accurate AI-assisted wildlife identifications in camera trap images View paper
- [21] Performance Evaluation of Cloud Native Applications: A Systematic Mapping Study View paper
- [22] Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis View paper
- [23] Building Foundations for Inclusiveness through Expert-Annotated Data. View paper
- [24] Man and the machine: Effects of AI-assisted human labeling on interactive annotation of real-time video streams View paper
- [25] Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists View paper
- [26] The impact of inconsistent human annotations on AI driven clinical decision making View paper
- [27] What Makes An Expert? Reviewing How ML Researchers Define" Expert" View paper
- [28] Enabling Domain Expert Evaluation of Emerging AI Technologies in Healthcare Settings View paper
- [29] CheXseg: combining expert annotations with DNN-generated saliency maps for X-ray segmentation View paper
- [30] Advancing AI Factuality via Comprehensive Evaluation View paper
- [31] 26 Evaluating AI models trained with varying amounts of expert feedback for chronic graft-versus-host disease skin assessment in photos of patients with diverse skin tones View paper
- [32] Correcting the Measurement Errors of AI-Assisted Labeling in Image Analysis Using Design-Based Supervised Learning View paper
- [33] Semi-Supervised Learning in Prostate MRI Tumor Segmentation Approaches Fully-Supervised Performance on External Validation View paper
- [34] 2025 Expert consensus on retrospective evaluation of large language model applications in clinical scenarios View paper
- [35] Reasoning Visual Language Model for Chest X-Ray Analysis View paper
- [36] Modeling annotator expertise: Learning when everybody knows a bit of something View paper
- [37] MATRIX: Multi-Agent simulaTion fRamework for safe Interactions and conteXtual clinical conversational evaluation View paper
- [38] Unlocking new frontiers in leukemia diagnostics through large language Modelâ□□Driven report generation View paper
- [39] Recycling weak labels for multiclass classification View paper
- [40] Impact of hybrid supervision approaches on the performance of artificial intelligence for the classification of chest radiographs View paper
- [41] Synergistic AI-resident approach achieves superior diagnostic accuracy in tertiary ophthalmic care for glaucoma and retinal disease View paper
- [42] A semi-automated quality assurance tool for cardiovascular magnetic resonance imaging: application to outlier detection, artificial intelligence evaluation and trainee â□¦ View paper
- [43] Exploring Diagnostic Prompting Approach for Multimodal LLM-based Visual Complexity Assessment: A Case Study of Amazon Search Result Pages View paper
- [44] CLImage: Human-Annotated Datasets for Complementary-Label Learning View paper
- [45] Robustness of human vs. AI measurements under progressive image degradation View paper
- [46] Machine-learning for quantitative histopathology of piglet intestinal tissues: challenges with limited training data View paper
- [47] A Multi-faceted Analysis of Cognitive Abilities: Evaluating Prompt Methods with Large Language Models on the CONSORT Checklist View paper
- [48] Augmenting Document-level Relation Extraction with Efficient Multi-Supervision View paper
- [49] Human Experts vs. Large Language Models: Evaluating Annotation Scheme and Guidelines Development for Clinical View paper
- [50] First, do NOHARM: towards clinically safe large language models View paper
- [51] Federated benchmarking of medical artificial intelligence with MedPerf View paper
- [52] Beyond Manual Annotation: A Human-AI Collaborative Framework for Medical Image Segmentation Using Only â□□Better or Worseâ□□ Expert Feedback View paper
- [53] A Human-Centric Assessment Framework for AI View paper
- [54] Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels View paper
- [55] Learning with Multiple Complementary Labels View paper
- [56] Weak-to-strong generalization: Eliciting strong capabilities with weak supervision View paper
- [57] Simvlm: Simple visual language model pretraining with weak supervision View paper
- [58] Difference-Complementary Learning and Label Reassignment for Multimodal Semi-Supervised Semantic Segmentation of Remote Sensing Images View paper
- [59] Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data View paper
- [60] Pre-trained Vision-Language Models Assisted Noisy Partial Label Learning View paper
- [61] Cold-start active learning through self-supervised language modeling View paper
- [62] Pre-Trained Vision-Language Models as Noisy Partial Annotators View paper
- [63] Language models in the loop: Incorporating prompting into weak supervision View paper
- [64] Learning with biased complementary labels View paper
- [65] Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents: V. Oliveira et al. View paper