Novelty Assessment Report

Paper: Identity-Free Deferral For Unseen Experts
PDF URL: https://openreview.net/pdf?id=4YG9ufFg58
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Learning to Defer (L2D) improves AI reliability in decision-critical environments, such as healthcare, by training a model to either make its own prediction or delerejector the decision to a human expert. A key challenge is adapting to unseen experts: those who were not involved during the system's training process. Current methods for this task, however, can falter when unseen experts are out-of-distribution (OOD) relative to the training population. We identify a core architectural flaw as the cause: they learn identity-conditioned policies by processing class-indexed signals in fixed coordinates, creating shortcuts that violate the problem's inherent permutation symmetry. We introduce Identity-Free Deferral (IFD), an architecture that enforces this symmetry by construction. From a few-shot context, IFD builds a query-independent Bayesian competence profile for each expert. It then supplies the deferral rejector with a low-dimensional, role-indexed state containing only structural information, such as the model's confidence in its top-ranked class and the expert's estimated skill for that same role, which obscures absolute class identities. We train IFD using an uncertainty-aware, context-only objective that removes the need for expensive query-time expert labels. We formally prove the permutation invariance of our approach, contrasting it with the generic non-invariance of standard population encoders. Experiments on medical imaging benchmarks and ImageNet-16H with real human annotators show that IFD consistently improves generalization to unseen experts, with significant gains in OOD settings, all while using fewer annotations than competing methods.

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: **Learning to Defer to Unseen Human Experts Under Distribution Shift**

A total of **19 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **Deferral Architecture and Representation Design**
- **Uncertainty Quantification for Deferral Decisions**
- **Sequential and Adaptive Deferral Frameworks**
- **Human-in-the-Loop Adaptation Under Distribution Shift**
- **Robustness and Safety Frameworks**
- **Domain-Specific Applications and Simulation**

Complete Taxonomy Tree

- Learning to Defer to Unseen Human Experts Under Distribution Shift Survey Taxonomy
- Deferral Architecture and Representation Design
  - Symmetry-Preserving Deferral Architectures ★ (1 papers)
  - [0] Identity-Free Deferral For Unseen Experts (Anon et al., 2026) View paper
  - Complementarity-Based Deferral Systems (1 papers)
  - [1] Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians (Krishnamurthy, 2023) View paper
- Uncertainty Quantification for Deferral Decisions
  - Uncertainty-Aware Triage and Routing (1 papers)
  - [3] At-cxr: Uncertainty-aware agentic triage for chest x-rays (Li Xueyang, 2025) View paper
  - Uncertainty Evaluation Frameworks (3 papers)
  - [2] A framework for assessing joint human-AI systems based on uncertainty estimation (Emir Konuk, 2024) View paper
  - [9] Is Uncertainty Quantification a Viable (AM Wundram, 2025) View paper
  - [13] Is Uncertainty Quantification a Viable Alternative to Learned Deferral? (Baumgartner, 2025) View paper
  - Conformal and Probabilistic Uncertainty Methods (2 papers)
  - [4] Conformalized Interactive Imitation Learning: Handling Expert Shift and Intermittent Feedback (Zhao Michelle, 2024) View paper
  - [6] Uncertainty Estimation in Deep Learning Models for Reliable Autism Detection: Enhancing Clinical Trust Through Probabilistic Confidence Measures (H Khan, 2022) View paper
  - Distance-Based Reliability Metrics (1 papers)
  - [15] When to Accept Automated Predictions and When to Defer to Human Judgment? (Garcez Artur, 2024) View paper
- Sequential and Adaptive Deferral Frameworks (1 papers)
  - [5] Learning-to-defer for sequential medical decision-making under uncertainty (Shalmali Joshi, 2021) View paper
- Human-in-the-Loop Adaptation Under Distribution Shift
  - Preference-Based Policy Adaptation (1 papers)
  - [14] Deployable Vision-driven UAV River Navigation via Human-in-the-loop Preference Alignment (Wang Zihan, 2025) View paper

    - Drift Detection and Human Oversight (1 papers)
    - [10] A human-centric drift controller framework for adaptive and explainable quality control in manufacturing (Paolo Catti, 2025) View paper
    - Incremental Feature and Model Learning (2 papers)
    - [17] PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data (Mishra, 2023) View paper
    - [19] Human-on-the-loop continual learning: data, knowledge and agents for model adaptation (Rocca, n.d.) View paper
- Robustness and Safety Frameworks
    - Out-of-Distribution Detection and Mitigation (1 papers)
    - [7] A Framework to Enhance Security and Safety of Deep Learning Models Against Out-of-Distribution Examples (Azmoodeh, 2024) View paper
    - Human-AI Collaboration for Robustness (2 papers)
    - [8] Amplifying Human Experience: Enabling Gain-of-Function via AI (Tom Golway, 2025) View paper
    - [16] Robustifying NLP with Humans in the Loop (Kaushik, 2023) View paper
    - Task-Agnostic Continual Learning (1 papers)
    - [18] TAME: Task Agnostic Continual Learning using Multiple Experts (Haoran Zhu, 2022) View paper
- Domain-Specific Applications and Simulation (2 papers)
    - [11] KarmaTS: A Universal Simulation Platform for Multivariate Time Series with Functional Causal Dynamics (Haixin Li, 2025) View paper
    - [12] DeepTag: inferring all-cause diagnoses from clinical notes in under-resourced medical domain (Nie, 2018) View paper

Narrative

Core task: Learning to defer to unseen human experts under distribution shift. This field addresses the challenge of building AI systems that can recognize when to hand off decisions to human experts, particularly when the system encounters data that differs from its training distribution and when expert identities or characteristics are not known in advance. The taxonomy organizes research into several complementary directions: Deferral Architecture and Representation Design focuses on how to structure models that can learn deferral policies without relying on expert-specific features; Uncertainty Quantification for Deferral Decisions develops principled ways to measure confidence and trigger handoffs; Sequential and Adaptive Deferral Frameworks handle multi-step interactions where deferral decisions unfold over time; Human-in-the-Loop Adaptation Under Distribution Shift studies how systems can learn from expert feedback when data distributions evolve; Robustness and Safety Frameworks ensure reliable operation under adversarial or high-stakes conditions; and Domain-Specific Applications and Simulation ground these ideas in real-world settings such as medical diagnosis and autonomous systems.

A particularly active line of work explores how to quantify uncertainty in ways that inform deferral decisions, with approaches ranging from conformal prediction methods like Conformalized Interactive Imitation[4] to joint assessments of model and human uncertainty as in Uncertainty Joint Assessment[2]. Another contrasting theme examines whether deferral architectures should explicitly model complementarity between AI and human strengths, as in Complementarity-Driven Deferral[1], or remain agnostic to expert identity. Identity-Free Deferral[0] sits squarely within the Symmetry-Preserving Deferral Architectures branch, emphasizing that effective deferral can be learned without access to expert-specific identifiers, a design choice that contrasts with sequential frameworks like Sequential Medical Deferral[5] where expert characteristics may be partially observable over time. This work addresses a key open question: how to generalize deferral policies to entirely new experts under shifted data distributions, a setting where traditional expert-aware methods struggle.

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 architectural constraints (permutation invariance, symmetry) that enable generalization to unseen expert identities without relying on fixed encodings. The sibling subtopic emphasizes learning complementary strengths between AI and clinical workflows for deferral decisions. Both address learning-to-defer scenarios, but through fundamentally different mechanisms: structural invariance versus complementarity modeling.

Similarities: - Both address learning-to-defer problems where AI systems must decide when to defer to human experts - Both aim to improve deferral decisions beyond simple uncertainty thresholds - Both consider the relationship between AI capabilities and human expertise in the deferral mechanism

Differences: - Symmetry-Preserving uses formal architectural constraints (permutation invariance) to handle unseen experts; Complementarity-Based learns task-specific strengths between AI and workflows - Symmetry-Preserving focuses on generalization across expert identities through structural guarantees; Complementarity-Based focuses on exploiting AI-human complementarity within clinical contexts - Symmetry-Preserving excludes fixed identity encodings; Complementarity-Based excludes general uncertainty-based methods, suggesting it uses domain-specific complementarity signals rather than generic confidence scores

Suggested Search Directions: - Investigate whether symmetry-preserving architectures could incorporate complementarity signals while maintaining permutation invariance - Explore how clinical workflow complementarity might be modeled in identity-agnostic frameworks - Examine whether complementarity-based systems implicitly learn symmetries when expert identities are unavailable

Sibling Subtopics

- **Complementarity-Based Deferral Systems** (leaves: 1, papers: 1)
- Scope: Systems that learn to exploit complementary strengths between AI predictions and clinical workflows for deferral decisions.
- Exclude: Excludes general uncertainty-based or sequential deferral methods; those belong in respective sibling categories.

Contributions Analysis

Overall novelty summary. The paper introduces Identity-Free Deferral (IFD), an architecture that enforces permutation symmetry to generalize deferral decisions to unseen human experts under distribution shift. According to the taxonomy, this work occupies the 'Symmetry-Preserving Deferral Architectures' leaf, which currently contains only this paper as its sole member. This positioning suggests the paper pioneers a relatively sparse research direction within the broader deferral architecture design space, contrasting with identity-conditioned approaches that dominate existing methods.

The taxonomy reveals that IFD's parent category, 'Deferral Architecture and Representation Design', contains one sibling leaf: 'Complementarity-Based Deferral Systems', which focuses on exploiting complementary AI-human strengths rather than enforcing symmetry constraints. Neighboring branches include 'Uncertainty Quantification for Deferral Decisions' (with four sub-categories spanning conformal methods, distance-based metrics, and triage systems) and 'Sequential and Adaptive Deferral Frameworks'. The scope

notes clarify that IFD's symmetry-preserving design explicitly excludes fixed identity encodings, distinguishing it from methods that condition on expert-specific features.

Among seven candidates examined for the uncertainty-aware training objective contribution, one paper appears to provide overlapping prior work, while six others were non-refutable or unclear. The IFD architecture itself and the formal permutation invariance proof were not examined against any candidates in this limited search. This suggests that while the training methodology may have some precedent in the examined literature, the core architectural innovation and its theoretical guarantees remain less directly challenged within the scope of this analysis, which covered top-K semantic matches rather than an exhaustive field survey.

Based on the limited search scope of seven candidates, the work appears to introduce a novel architectural perspective by formalizing symmetry constraints for expert-agnostic deferral. However, the analysis does not cover the full landscape of identity-conditioned methods or alternative symmetry-preserving designs that may exist outside the top-K semantic neighborhood. The single-paper occupancy of its taxonomy leaf reflects either genuine novelty in this specific direction or the nascent state of research explicitly framing deferral through permutation invariance.

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

Contribution 1: Identity-Free Deferral (IFD) architecture

Description: IFD is a novel architecture for learning to defer that enforces permutation symmetry by construction. It builds query-independent Bayesian competence profiles for experts and supplies the rejector with a low-dimensional, role-indexed state containing only structural information, which obscures absolute class identities and prevents identity-conditioned shortcuts.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Uncertainty-aware, context-only training objective

Description: The authors introduce a training objective that uses only context-derived expert profiles and incorporates risk-sensitive weighting via lower confidence bounds. This eliminates the need for expensive query-time expert annotations while naturally downweighting uncertain profiles to prevent over-deferral.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. No Need for Learning to Defer? A Training Free Deferral Framework to Multiple Experts through Conformal Prediction

URL: View paper

Brief Assessment

Training Free Deferral[23] uses conformal prediction for uncertainty quantification and does not involve training a deferral model. The original paper's contribution focuses on a training objective that learns from context-derived expert profiles, which is fundamentally different from the candidate's training-free approach.

---

2. When Does Confidence-Based Cascade Deferral Suffice?

URL: View paper

Brief Assessment

Confidence Cascade Deferral[24] focuses on cascade deferral between multiple models in sequence, not expert deferral with context-derived profiles. The candidate's uncertainty mechanisms relate to model confidence thresholds, not Bayesian expert competence profiles derived from context sets.

---

3. Towards Uncertainty Aware Task Delegation and Human-AI Collaborative Decision-Making

URL: View paper

Brief Assessment

Uncertainty Aware Delegation[20] focuses on distance-based uncertainty scores for human-AI task delegation in rehabilitation assessment, not on training objectives for expert deferral systems using context sets without query-time labels.

---

4. Expert-Agnostic Learning to Defer

URL: View paper

Prior Art Analysis

Expert-Agnostic Deferral[22] demonstrates prior work that uses context-derived expert profiles with uncertainty-aware weighting, eliminating query-time expert annotations. Both papers construct Bayesian per-class expert profiles from context sets and incorporate risk-sensitive weighting mechanisms. Expert-Agnostic Deferral[22] uses a lower confidence bound (LCB) weighting scheme (equation 3) that downweights uncertain profiles, matching the original paper's claim of using 'risk-sensitive weighting via lower confidence bounds' to 'naturally downweight uncertain profiles to prevent over-deferral.'

Evidence

Evidence 1 - **Rationale**: Both papers eliminate the need for query-time expert labels by using only context-derived information for training the deferral mechanism. - **Original**: we train ifd using an uncertainty-aware, context-only objective that removes the need for expensive query-time expert labels. - **Candidate**: to train ea-l2d without requiring expert annotations m on the query set dq, we adapt the standard surrogate loss lce (eq. (2)). the core modification involves replacing the ground-truth deferral condition i[m = yq] with an evidence-based approximation derived from the expert representation re.

Evidence 2 - **Rationale**: Expert-Agnostic Deferral[22] uses an LCB weighting mechanism that incorporates uncertainty (via standard deviation) to weight deferral decisions, matching the original paper's uncertainty-aware weighting approach. - **Original**: we propose a data-efficient and uncertainty-aware training objectivewhich adapts to new experts using only a small context set, completely eliminating the need for costly expert labels at query-time. - **Candidate**: furthermore, to incorporate the uncertainty inherent in estimating expert performance from limited context data, we weight this condition using a risk-sensitive lower confidence bound (lcb): we yq := max(0, μe k=yq - ασe k=yq), (3) where μe k=yq and σe k=yq are the posterior mean and standard deviat...

Evidence 3 - **Rationale**: Both papers describe mechanisms that downweight uncertain expert profiles to prevent inappropriate deferral, using similar mathematical formulations based on confidence bounds. - **Original**: training uses an uncertainty-aware, context-only surrogate objective, which naturally downweights uncertain profiles and eliminates the need for query-time expert labels. - **Candidate**: this lcb weight tempers the deferral incentive when accuracy estimates are uncertain (i.e., high σe k=yq), particularly useful when context data is sparse.

Evidence 4 - **Rationale**: Both papers construct Bayesian per-class competence profiles from context sets using similar statistical approaches, demonstrating prior work on this methodology. - **Original**: from a few-shot context, ifd builds a query-independent bayesian competence profile for each expert. it then supplies the deferral rejector with a low-dimensional, role-indexed state containing only structural information - **Candidate**: ea-l2d constructs an explicit and interpretable behavioural representation re for arbitrary expert e

from context data de c by modelling the expert's accuracy for each class k ∈ ywith a beta distribution. this approach leverages the beta-binomial conjugate model

---

5. Is Uncertainty Quantification a Viable

URL: View paper

Brief Assessment

Uncertainty Quantification Viable[9] focuses on uncertainty quantification for AI safety and deferral decisions, but does not present a comparable context-only training objective with risk-sensitive weighting via lower confidence bounds for expert deferral systems.

---

6. Learning-to-defer for sequential medical decision-making under uncertainty

URL: View paper

Brief Assessment

Sequential Medical Deferral[5] focuses on sequential decision-making in MDPs using model-based RL to defer based on long-term outcomes, not on context-only training objectives for expert deferral without query-time labels as in the original paper.

---

7. Two-stage learning to defer with multiple experts

URL: View paper

Brief Assessment

Two-Stage Multiple Experts[21] focuses on a two-stage learning framework where a predictor is pre-trained and a deferral function is learned separately, without requiring query-time expert labels. The original paper's uncertainty-aware objective uses lower confidence bounds from Bayesian per-class profiles in a single integrated framework, which is architecturally and methodologically distinct from the two-stage approach.

Contribution 3: Formal proof of permutation invariance

Description: The authors provide formal proofs demonstrating that IFD is invariant to coherent class relabelings, contrasting this with the generic non-invariance of standard population encoders. This theoretical result establishes that IFD blocks identity-conditioned shortcuts by design.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

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

References

- [0] Identity-Free Deferral For Unseen Experts View paper
- [1] Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians View paper
- [2] A framework for assessing joint human-AI systems based on uncertainty estimation View paper
- [3] At-cxr: Uncertainty-aware agentic triage for chest x-rays View paper
- [4] Conformalized Interactive Imitation Learning: Handling Expert Shift and Intermittent Feedback View paper
- [5] Learning-to-defer for sequential medical decision-making under uncertainty View paper
- [6] Uncertainty Estimation in Deep Learning Models for Reliable Autism Detection: Enhancing Clinical Trust Through Probabilistic Confidence Measures View paper
- [7] A Framework to Enhance Security and Safety of Deep Learning Models Against Out-of-Distribution Examples View paper
- [8] Amplifying Human Experience: Enabling Gain-of-Function via AI View paper
- [9] Is Uncertainty Quantification a Viable View paper
- [10] A human-centric drift controller framework for adaptive and explainable quality control in manufacturing View paper
- [11] KarmaTS: A Universal Simulation Platform for Multivariate Time Series with Functional Causal Dynamics View paper
- [12] DeepTag: inferring all-cause diagnoses from clinical notes in under-resourced medical domain View paper
- [13] Is Uncertainty Quantification a Viable Alternative to Learned Deferral? View paper
- [14] Deployable Vision-driven UAV River Navigation via Human-in-the-loop Preference Alignment View paper
- [15] When to Accept Automated Predictions and When to Defer to Human Judgment? View paper
- [16] Robustifying NLP with Humans in the Loop View paper
- [17] PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data View paper
- [18] TAME: Task Agnostic Continual Learning using Multiple Experts View paper
- [19] Human-on-the-loop continual learning: data, knowledge and agents for model adaptation View paper
- [20] Towards Uncertainty Aware Task Delegation and Human-AI Collaborative Decision-Making View paper
- [21] Two-stage learning to defer with multiple experts View paper
- [22] Expert-Agnostic Learning to Defer View paper
- [23] No Need for Learning to Defer? A Training Free Deferral Framework to Multiple Experts through Conformal Prediction View paper
- [24] When Does Confidence-Based Cascade Deferral Suffice? View paper