ML4JETS

# Classification-Generative Supervised Fine-Tuning

for Hierarchical Tagging of ML4HEP Literature

**Ziyue Chen | IHEP UCAS**

**Collaborators**

Claudius Krause, PhD | MBI Vienna (ÖAW)

Ke Li | Chinese Academy of Sciences (CN)

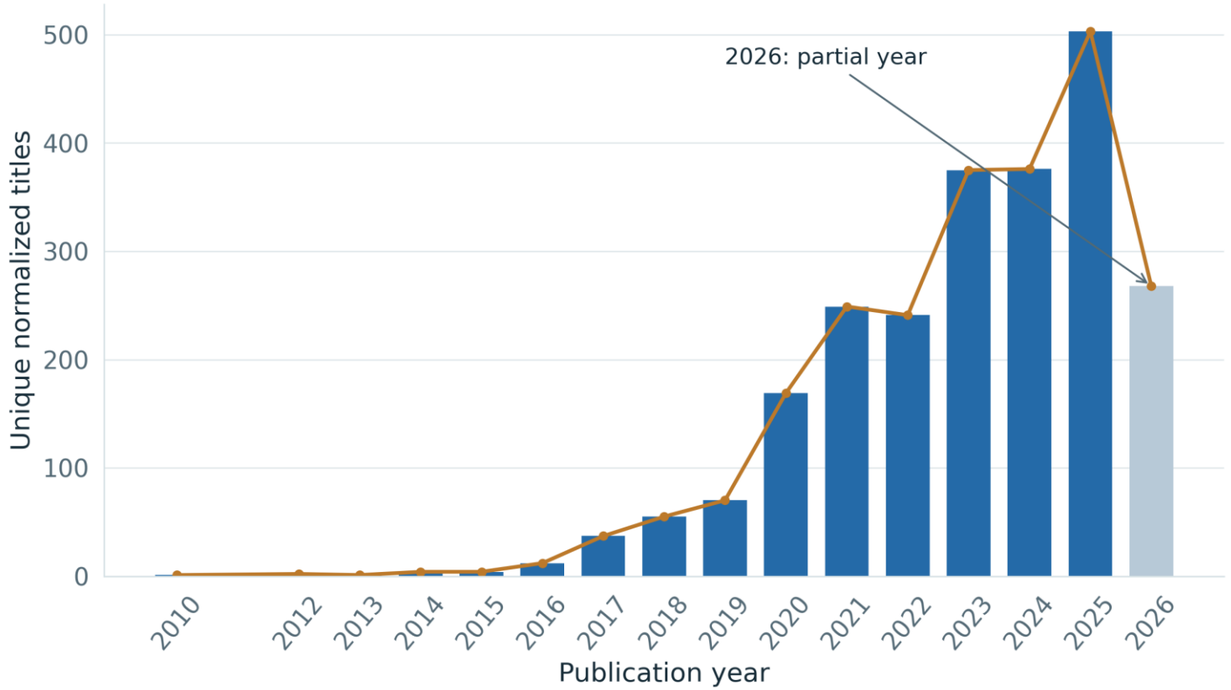
Ramon Winterhalder | Università degli Studi di Milano

Yulei Zhang | University of Washington (US)

# A growing literature makes manual tagging a recurring bottleneck

## ML4HEP literature is growing in volume, topical breadth, and overlap.

- One paper may span methods, datasets, applications, and parent topics.
- Manual tagging repeats these decisions paper by paper.
- Human curation becomes time consuming as the corpus grows.




Source: iml-wg.github.io/HEPML-LivingReview/

## REPORT OUTLINE

- Part 1 — CG-SFT: fine-tune an LLM to assign labels to established HEP/ML papers.
- Part 2 — knowledge base + HEPML Agent: organize the curated literature for domain research.

# Research purpose:  Classify literature using LLM and further organize it into a knowledge base

- **Input: paper title + abstract**
- **Output: 91 labels in a fixed order**

  - Multi-label: several positive labels per paper.
  - Hierarchical: parent–child relations add structure.
  - Example: one paper can cover graph representations and jet tagging.
  - One controlled <yes>/<no> decision per label.
  - Fixed order → checkable vector.
  - No formatting ambiguity.

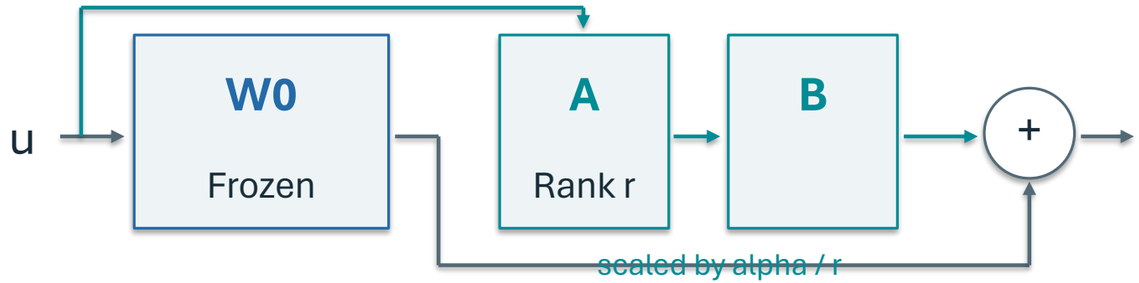
91 labels · 77 parent–child relations · 1,819 records · 186 held-out test papers


```
{
  "instruction": "You are a paper classification assistant. Based on the given paper title and abstract, classify the paper
  "input": "Automated visual inspection of CMS HGCAL silicon sensor surface using an ensemble of a deep convolutional autoen
  "output": "Experimental results : (Performance studies)"
},
{
  "instruction": "You are a paper classification assistant. Based on the given paper title and abstract, classify the paper
  "input": "Ps and Qs: Quantization-aware pruning for efficient low latency neural network inference. Efficient machine lear
  "output": "Classification : (Fast inference / deployment : Hardware/firmware)"
},
{
  "instruction": "You are a paper classification assistant. Based on the given paper title and abstract, classify the paper
  "input": "Evaluating the Impact of Detector Design on Jet Flavor Tagging for Future Colliders. Jet flavor tagging is of ut
  "output": "Classification : (Representations : Graphs)"
},
{
  "instruction": "You are a paper classification assistant. Based on the given paper title and abstract, classify the paper
  "input": "Machine learning method for enforcing variable independence in background estimation with LHC data: ABCDisCoTEC.
  "output": "Decorrelation methods."
},
{
  "instruction": "You are a paper classification assistant. Based on the given paper title and abstract, classify the paper
  "input": "Refining Fast Calorimeter Simulations with a Schr\\\"odinger Bridge. Machine learning-based simulations, especia
  "output": "Generative models / density estimation : (Diffusion Models)"
},
```

# One generative network, specialized at the decision interface

Shared Qwen states feed the native LM head; LoRA and optional decision-row adapters support task adaptation.


## Attention projection: frozen path + low-rank update




$$\mathbf{v} = W_0\mathbf{u} + \frac{\alpha}{r}BA\mathbf{u}$$

## Optional decision-row adapters

$$\ell_d = \ell_d^{(0)} + \Delta\mathbf{w}_d^{\top}\mathbf{h} + \Delta b_d$$

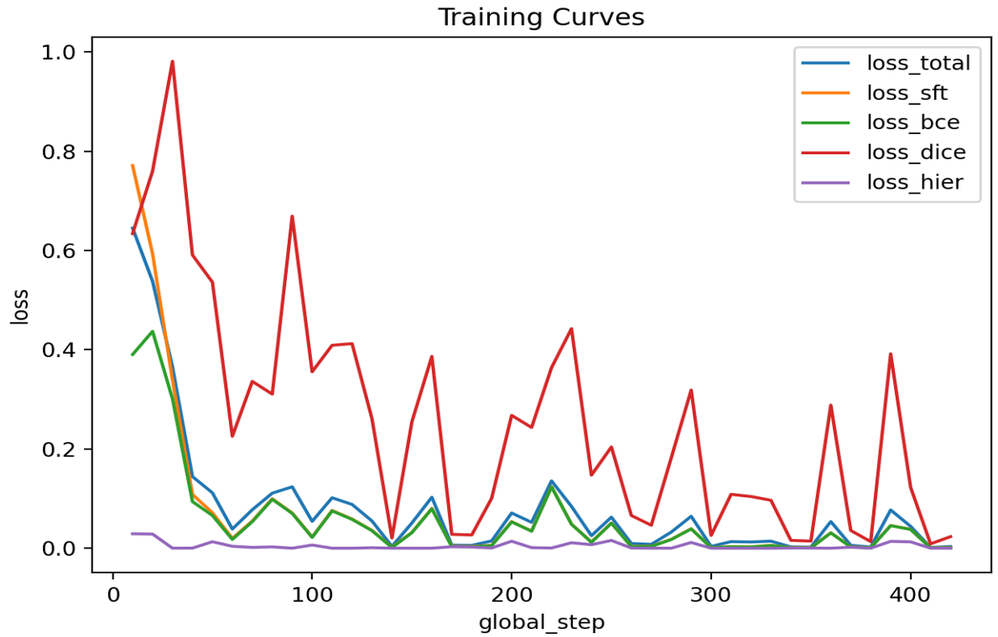
d = <yes> or <no>

Input embeddings also receive two learned row deltas.

# Joint supervision links token generation, binary decisions, and hierarchy

**At position t, the causal logits are read at t – 1.**

- Selective SFT: supervise only the 91 decision positions.
- Binary cross-entropy: separate <yes> from <no>.
- Dice loss: address sparse positive labels.
- Hierarchy term: softly penalize p_child > p_parent.

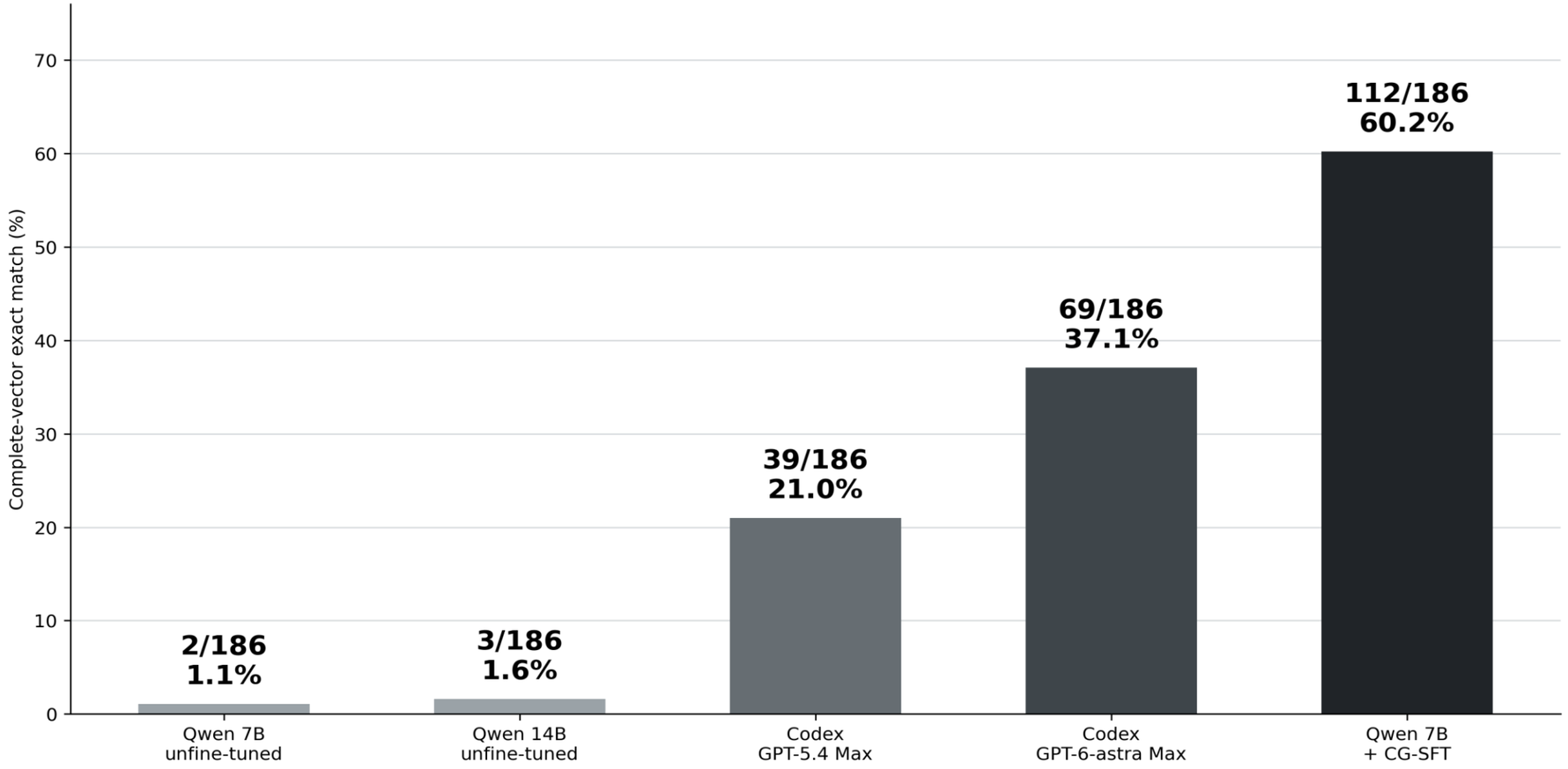



Training curves from the reported run

A strict and reproducible comparison

- **Same 186 held-out papers and title-and-abstract inputs.**
- **Same 91-label taxonomy and fixed order.**
- **Complete-vector exact match: all 91 decisions must agree.**

All 91 decisions must agree for one paper to count as correct.



What the five-way comparison tells us

## TASK COMPLEXITY

This is a difficult text-classification task:

- Each paper can receive many labels
- The labels form a complex parent–child structure
- The model must map paper meaning to a 91-label taxonomy.

## WHAT THE RESULTS SHOW

Before fine-tuning, the small models almost never classify correctly: 2/186 and 3/186 exact matches. CG-SFT raises the 7B model to 112/186. Even GPT-6-astra Max reaches only 69/186, below the fine-tuned small model.

# CG-SFT is meaningful because it matches the task, not just the model size

**Results show that CG-SFT is highly effective for fast, domain-specific paper classification.**

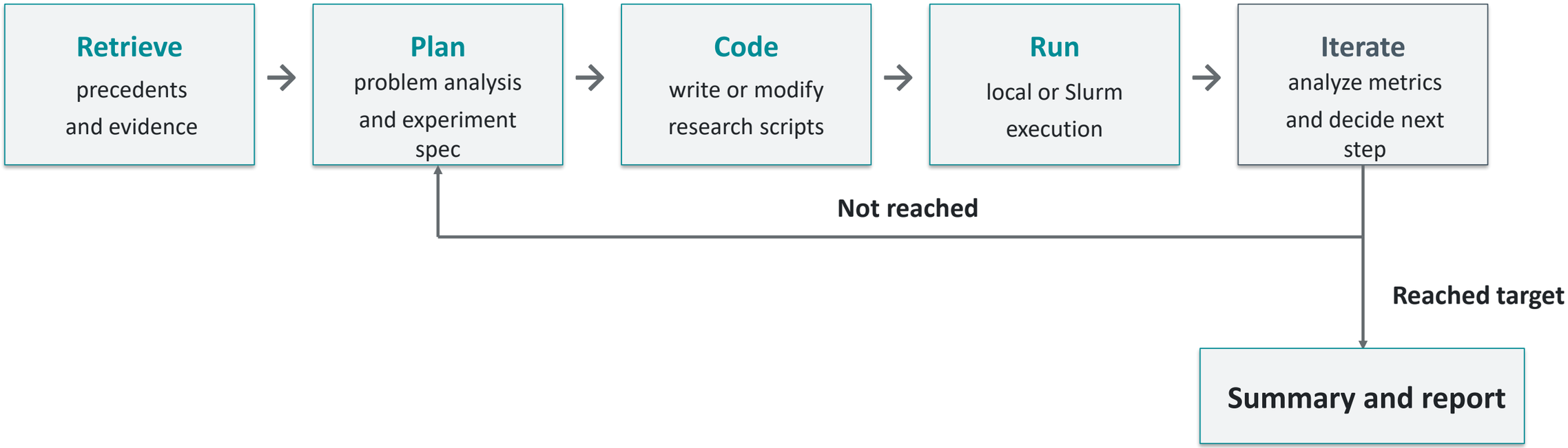
- Large gain on the difficult 91-label task.
- Fixed decision interface makes outputs machine-readable.

- Main limitation: performance depends on human labeling quality.
- Noisy or incomplete labels can be learned and repeated at scale.

**NEXT  Curated papers → knowledge base → HEPML Agent**

A dedicated HEPML Agent for end-to-end research support

## A domain Agent can automate the main research loop:




Knowledge base from the CG-SFT literature; precedent retrieval + persistent experiment records.

# Knowledge base: separate provenance, explicit retrieval, quality boundaries

### Shared literature

Methods: representations, datasets, reported results.

### Project literature

Newly extracted papers remain local; local versions take precedence.

### Research memory

Agent decisions and experiments, each linked to artifacts.

- Retrieval axes · literal terms · relevance · transfer conditions.
- Bounded excerpts with paper ID, source, section, path, and line references.

**Retrieval is evidence, not scientific validation.**

AGENT / EVIDENCE EXAMPLE

Real knowledge-base record: arXiv:2504.01496

## The Agent stores a paper as three linked cards, with source and quality information.

### Problem

- Two undetected neutrinos → underconstrained LHC event.
- Goal: spin-density matrix, entanglement, and Bell nonlocality.

### Result

- ~17% average momentum resolution.
- 6–8% ττ mass resolution; Bell nonlocality above 5σ.

### Solution

- PET transformer body + diffusion generation head.
- DDIM reconstructs the three-momenta of two effective neutrinos.

**The extracted record concerns entanglement and Bell nonlocality in $\tau^+\tau^-$ production at the LHC, enabled by machine-learning neutrino reconstruction.**


```
extensions:
  proposed_benchmark: proposes τ+τ- as the new benchmark system for quantum i
    studies at the LHC, complementing tt̄
schema_version: v3
problem_id: arxiv_2504.01496
title: Entanglement and Bell Nonlocality in τ+ τ- at the LHC using Machine Le
  for Neutrino Reconstruction
physics_task:
  extensions:
    quantum_observables:
    - entanglement
    - Bell nonlocality
    system: τ+τ- two-qubit state
    process: pp → τ+ τ- X
  category: quantum_tomography
  subcategory: spin_density_matrix_reconstruction
  ml_formulation:
  - conditional_generation
  - regression
experiment_contexts:
- extensions:
    process: pp → τ+τ- X
    physics_goal: measure full spin density matrix, entanglement and Bell non
      surpassing 5σ
  experiment: τ+τ- production at the LHC
  facility: CERN Large Hadron Collider (LHC)
  detector: null
  role: quantum tomography and measurement of entanglement and Bell nonlocali
    the τ+τ- two-qubit state using machine-learning neutrino momentum reconst
  conditions: null
  evidence_ids:
  - source_section_001.part_001
  - process_001
- extensions:
    trigger_subchannels:
```

# A real application: JUNO vertex and energy reconstruction

## An Agent-led reconstruction study using MC positron events.

### Retrieve

Read prior decisions and baselines.

### Plan + code

Inspect PMT hit time, charge, and position.
Write reconstruction scripts.

### Run + analyze

Submit Slurm jobs.
Compare vertex and energy metrics.

### Iterate

Choose the next direction.
Store decisions and artifacts.

**Recorded outcome: about 167 mm position resolution with a classic residual head.**
**Decision, artifacts, and source paths remain linked.**

Research loop: retrieve → plan → code → run → analyze → iterate.

Agent conclusion: a closed loop for HEPML research

## The Agent keeps the whole research chain connected:

**Retrieve → analyze → code → run → metrics → iterate → memory**

### External evidence

The arXiv record preserves problem, method, result, source, and quality state.

### Research memory

The JUNO record preserves what was tried, what worked, what failed, and the next decision.

**Research state becomes easier to recover, inspect, and reuse.**
**Scientific review remains necessary.**

Overall conclusion and outlook

## CG-SFT

- CG-SFT is highly effective.
- It greatly improves LLM classification in a specific domain.

## HEPML Agent

- HEP + ML knowledge base built from curated literature.
- Self-iterating HEPML Agent developed.

**Next: stronger label-quality checks, broader evaluation, better evidence retrieval, and transfer to more HEPML tasks.**

**Thank you · Questions and discussion**