COLM 2025

University of Washington
Allen Institute for AI
Carnegie Mellon University

The Delta Learning Hypothesis

Preference Tuning on Weak Data can Yield Strong Gains

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap
Ranjay Krishna, Pang Wei Koh, Jerry Li

GitHub Repo Available

PRESENTATION OVERVIEW

Research Roadmap

01 **Introduction & Motivation**

Challenging conventional wisdom about strong supervision requirements

02 **The Delta Learning Hypothesis**

Formalizing the concept of learning from relative quality differences

03 **Controlled Experiments**

Validating through systematic experimentation with stylistic and semantic deltas

04 **Large-Scale Post-Training**

Testing at scale: 8B LLM post-training without strong supervision

05 **Theoretical Analysis**

Understanding mechanisms through logistic regression analysis

06 **Key Findings & Implications**

Synthesis, practical impact, and future research directions

CHAPTER 01

Introduction & Motivation

Challenging the conventional wisdom:

Can models learn from weak supervision?

INTRODUCTION

The Problem: Strong Data Builds Strong Models... Or Does It?

Conventional Wisdom

Common principle in ML: Improving performance requires training on data that exceeds a model's current capabilities.

This has driven progress across the LM pipeline: pretraining corpus curation, rejection sampling for finetuning, and preference tuning with human annotators identifying best outputs.

The Inherent Limitation

Implication: Model capability may be upper-bounded by the strength of supervision available.

Many desirable tasks are difficult to support with strong data due to high costs or tasks exceeding human expertise.

The Challenge

1. **High Collection Costs**

   E.g., synthesizing scientific literature at PhD level requires domain experts and significant resources.

2. **Exceeds Human Expertise**

   E.g., formulating a unified theory of physics requires superhuman reasoning capabilities.

Research Question

How might we build models that exceed the capabilities demonstrated in their training data?

BREAKTHROUGH INSIGHT

The Core Insight: Learning from Paired Weakness

Key Discovery

Preference pairs with individually weak data points can improve a stronger model beyond the strength of each individual sample.

This challenges the assumption that we need high-quality, better-than-current-policy responses.

Pilot Result: 8B Llama 3 Experiment

Setup

Preference tune 8B Llama 3 using **paired outputs from weaker models** (3B, 1.5B)

Outcome with DPO

Consistent performance gains

Outcome with SFT

Performance degradation

The Paradox

The Intuition

Pairwise Contrast

Valuable signal exists in the **pairwise contrast** between chosen and rejected responses.

Directional Signal

The relative difference provides a direction for improvement.

Preference Tuning Leverages

Preference tuning can **leverage this contrast** to improve the model.

> This observation motivates the Delta Learning Hypothesis

CHAPTER 02

The Delta Learning Hypothesis

Formalizing the intuition:

relative quality differences drive learning

THE DELTA LEARNING HYPOTHESIS

What is the Delta Learning Hypothesis?

Core Concept

Training on paired responses **(x, yc, yr)** enables learning from the relative quality difference — the **delta** — between chosen and rejected responses.

Even if both responses have low absolute quality compared to the model being trained, as long as y_c is better than y_r along some informative axes, the model can learn from this delta and improve.

Formal Statement

Let **μ(x, y)** be the utility of response y to prompt x. We wish to improve model M.

1. **Low Absolute Utility**
   $\mu(x, y_c) \leq$ M's capability, so SFT on (x, y_c) hurts.
2. **Extrapolated Gain**
   Preference tuning on the pair improves M beyond $\mu(x, y_c)$.

The Intuition

- **The delta defines a meaningful direction** of improvement.
- A **strong model may learn to generalize** along this direction.
- The model can improve beyond the absolute quality of the preferred example.

Key Requirements

- **Informative delta:** y_c must be better than y_r along meaningful dimensions.
- **Preference tuning:** algorithm must leverage pairwise contrast effectively.
- **Generalization:** model must extrapolate from delta to unseen examples.

CHAPTER 03

Controlled Experiments

Validating the hypothesis through systematic experimentation

CONTROLLED EXPERIMENT 1

Stylistic Delta: Number of Bold Sections

Toy Setting: Explicit Quality Definition

We explicitly define $\mu(x, y)$ as: "the number of Markdown-denoted bold section headers in y" (e.g., **example header**).

This provides a **measurable and controllable metric** to test the hypothesis in a clean setting.

Experimental Setup

1 Dataset Construction

Build dataset of prompts x matched with responses y_k containing varying numbers k of bolded sections.

2 Training

Tune Llama-3.2-3B-Instruct with DPO on pairs (x, y_{k_i}, y_{k_j}) where $k_i > k_j$ (more sections as chosen).

3 Controls

(1) Reverse preferences ($k_i < k_j$); (2) Equal sections ($k_i = k_j$); Compare to SFT on chosen y_{k_i}.

4 Evaluation

Measure average number of bolded sections generated on held-out test prompts.

Key Results

SFT Result

SFT only helps when training responses are **higher quality than model's baseline**.

When responses contain fewer sections than baseline (5.9), SFT decreases sections generated.

DPO Result

Even when responses are **individually weak**, pairing them with positive delta massively boosts section generation.

Model learns to make **nearly every word a new section header**!

Negative Controls

Preference tuning with negative or zero delta does not yield gains.

The positive delta is critical for learning!

CONTROLLED EXPERIMENT 2

Semantic Delta from Weaker Models

Beyond Style: Testing Semantic Quality

This experiment tests whether delta learning extends beyond one-dimensional style features to general semantic quality. We create a delta between self-generated outputs and outputs from a weaker model.

Experimental Design

1 Self-Generated Responses

Use model M to greedy decode responses $y_M = M(x)$. By construction, $\mu(x, y_M)$ = M's capability.

2 Weaker Model Responses

Use smaller model m from same family: $y_m = m(x)$. On average $\mu(x, y_M) > \mu(x, y_m)$.

3 Training

Tune M with DPO to prefer self-generated y_M over weaker y_m. Compare to SFT on y_M.

4 Guarantee

M never observes chosen response higher quality than it can produce.

Results

SFT on Self-Generated

Reduces average performance by 1.2 points. Possibly due to overfitting on self outputs at expense of broader ability.

DPO (Self over Weaker)

Creates positive delta driving learning beyond baseline. Yields small but consistent gains on nearly all benchmarks, with 0.4-point average gain.

Negative Control

Flipping preference order (preferring weaker over self) eliminates gains and worsens performance (-0.7 points).

Improvement comes specifically from the positive delta!

CHAPTER 04

Large-Scale Post-Training

Testing delta learning at scale:

8B LLM post-training without strong supervision

BASELINE COMPARISON

The Tülu 3 Baseline: State-of-the-Art Open Recipe

Why Tülu 3?

Tülu 3 represents the current state-of-the-art recipe in open-source post-training. It comprises 8B and 70B models achieving performance that matches or exceeds equivalently-sized proprietary m odels.

Tülu 3 Preference Data Pipeline

1. **Starting Point: 271k Diverse Prompts**
2. **Response Generation: Strong LLMs**
   Llama-3.1-70B-Instruct, Qwen-2.5-72B-Instruct, etc.
3. **Quality Scoring: GPT-4o Judge**
   Frontier LLM scores all responses on quality.
4. **Preference Pair Formation**
   Highest-scoring response as chosen, lower-scoring as rejected.
5. **DPO Tuning**
   Tune Tülu-3-8B-SFT on preference pairs to get Tülu-3-8B-DPO.

Cost Analysis

GPT-4o Annotation

~$10,000 USD

Substantial cost for quality scoring alone, excluding strong model generati on.

Key Assumptions

Strong Generation

Requires **70B+ models** to generate high-quality chosen responses.

Strong Judgment

Requires **GPT-4o** for accurate quality annotation.

Scale Constraint

Assumes access to stronger-than-student supervision .

OUR APPROACH

Our Simple Recipe: No Strong Models Required

Simplification Strategy

We intervene by removing all use of strong models (>3B parameters) while keeping starting checkpoint (Tülu-3-8B-SFT) and prompts fixed to isolate our changes.

Our Recipe Steps

1 Chosen Response Generation

Generate **all chosen responses** with a single small model (e.g., Qwen 2.5 3B Instruct) that is near or below Tülu-3-8B-SFT capability.

FLOPs Reduction: ~6% of original (~10x reduction)

2 Forming Preference Pairs

Eliminate GPT-4o entirely. Use **model size as proxy for quality**. Pair chosen response with next-smallest model.

E.g., Qwen 2.5 3B Instruct (chosen) ↔ Qwen 2.5 1.5B Instruct (rejected)

Three Datasets Created

1 Qwen-2.5-3B over 1.5B

Chosen: Qwen 2.5 3B Instruct
Rejected: Qwen 2.5 1.5B Instruct

2 Qwen-2.5-1.5B over 0.5B

Chosen: Qwen 2.5 1.5B Instruct
Rejected: Qwen 2.5 0.5B Instruct

3 Llama-3.2-3B over 1B

Chosen: Llama 3.2 3B Instruct
Rejected: Llama 3.2 1B Instruct

Key Advantage

No strong supervision needed!

Reduces reliance on expensive models

EXPERIMENTAL RESULTS

Striking Results: Matching Tülu 3 Performance

Main Finding

Our simple recipe matches Tülu 3 in performance!

Tuning with weak preference data achieves **comparable gains** to Tülu 3's approach, which relies on vastly stronger supervision.

Performance Comparison

Qwen-3B over 1.5B — +0.4 avg

Best setup, **beats Tülu 3** by 0.4 points average

Qwen-1.5B over 0.5B — +2.1 avg

Chosen model **11.4 points worse** than base, yet still yields +2.1 gain

Llama-3.2-3B over 1B — +5.5 GSM

Boosts **GSM8K accuracy by 5.5 points despite Llama being weaker than Tülu**

Quality Comparison

Tülu 3 Data

Absolute Quality:	4.44/5
Supervision:	GPT-4o + 70B models

Our Weak Data

Absolute Quality:	3.98/5
Supervision:	3B + 1.5B models

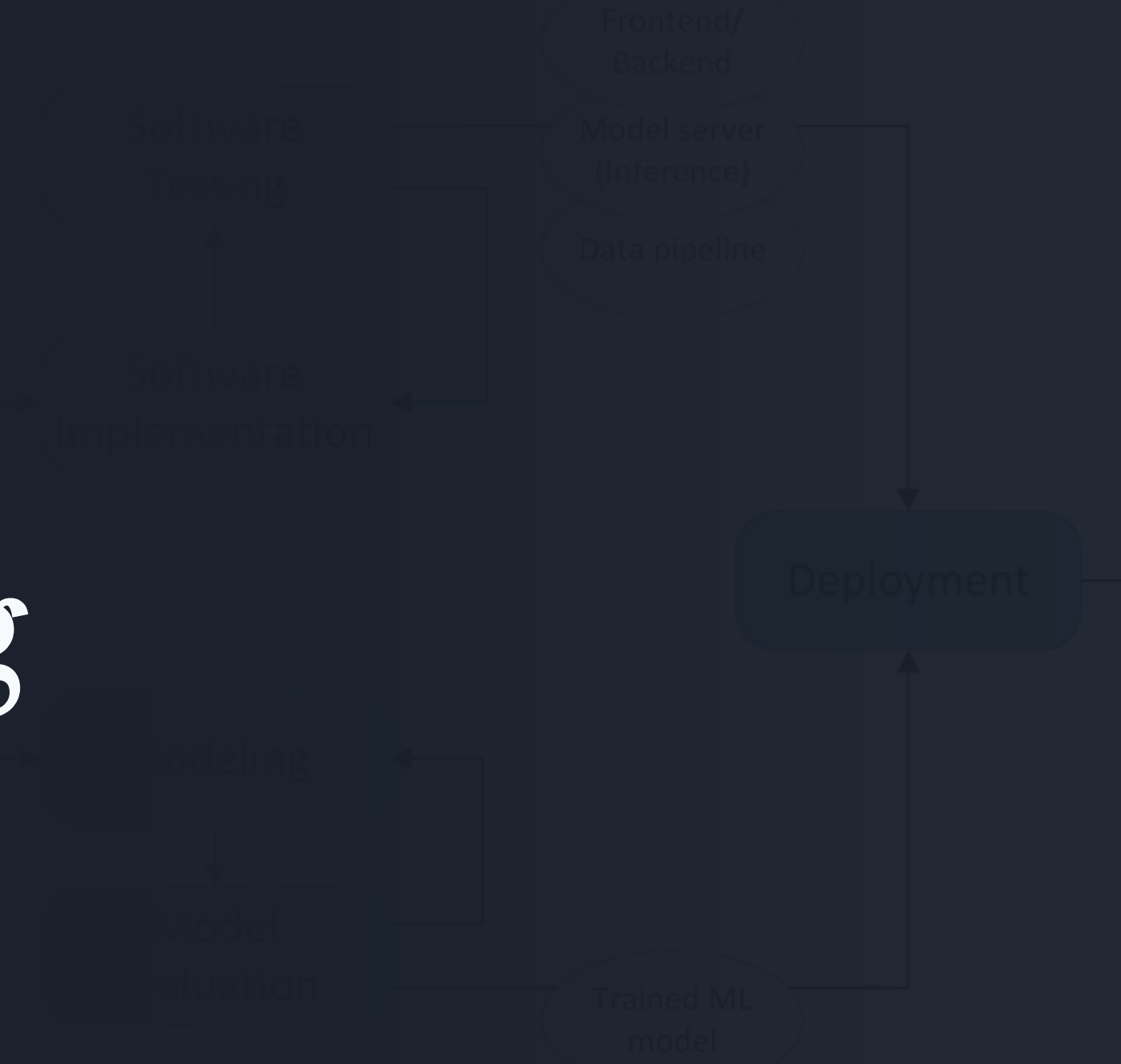
The quality delta between chosen and rejected responses suffices to produce comparable gains!

Recipe generalizes across model families!

CHAPTER 05

Analysis & Theoretical Understanding

Understanding what makes delta learning work

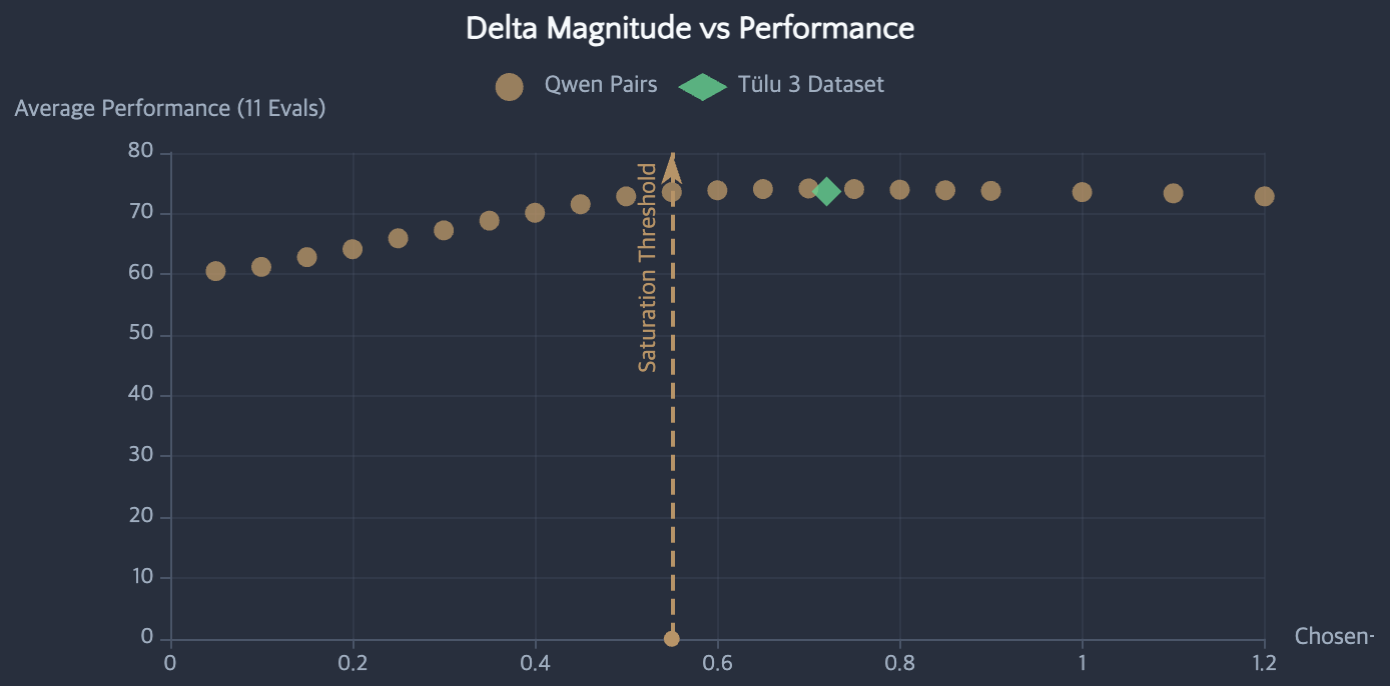


ANALYSIS: FACTOR 1

How Does Quality Delta Magnitude Affect Learning?

Experimental Design

Constructed **21 preference datasets** using all Qwen 2.5 Instruct model pairs (0.5B to 72B). Quantified response quality with GPT-4o annotations on 1-5 scale. Plot shows performance after tuning vs average pairwise delta.



Key Findings

Strong Correlation

Delta magnitude strongly predicts performance, up to $\Delta \approx 0.55$, then plateaus.

Tülu 3 Alignment

Tülu 3 dataset follows same trend, explaining why weak data matches it.

Not All Deltas Work

Pairing 72B Qwen with 32B or 14B hurts performance (both responses much stronger than base).

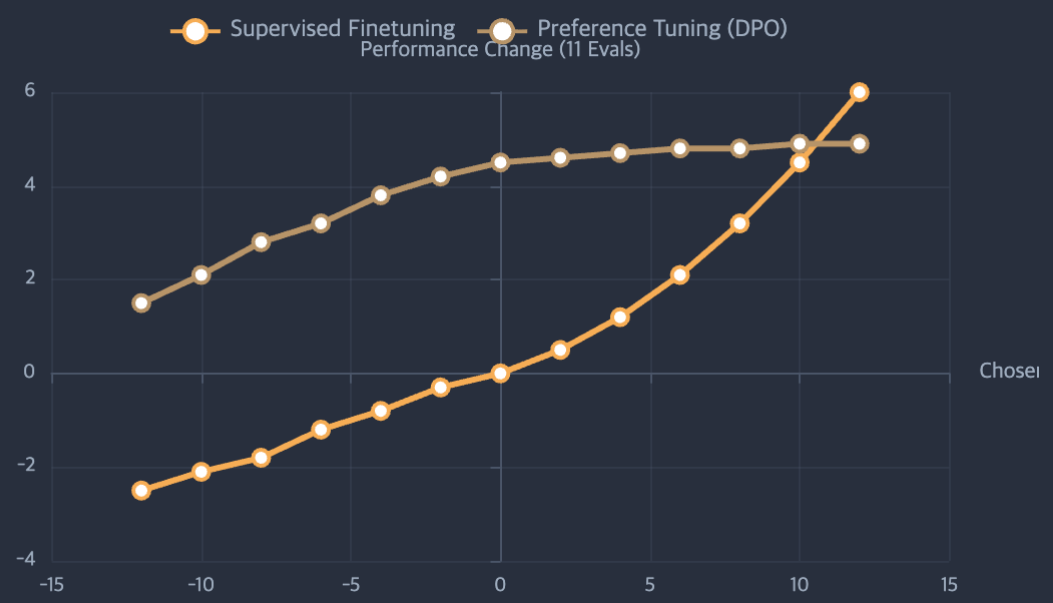
Outlier: Qwen-1.5B

Smaller gains than delta predicts because **chosen model substantially weaker** than base.

Optimal delta in ~[0.55, 0.8] range

ANALYSIS: FACTOR 2

How Does Chosen Response Absolute Quality Affect Learning?



SFT scales with quality; DPO is less dependent

Key Insights

SFT: Scales with Quality

- Gains only when **chosen model stronger than base**
- Gains scale monotonically with chosen quality

DPO: Less Dependent

- Gains even when **chosen responses weaker than base**
- Diminishing returns once chosen quality reaches base capability

Implication

This **saturation effect** explains why weak preference data matches Tülu 3 despite lower absolute quality.

ANALYSIS: ABLATIONS

Ablations: Model Size Heuristic & Base Model Choice

Model Size Heuristic

Using Tülu 3's GPT-4o judge to re-label our best weak dataset (Qwen-3B over 1.5 B).

Results

Agreement Rate: 80.5%

Model size vs GPT-4o preferences

GPT-4 vs Human: ~65%

Estimated from prior work

Performance with either GPT-4o labels or model size heuristic is comparable!

Base Model Generality

Test recipe on **OLMo-2-7B-SFT** with prompts from OLMo 2 Preference Dataset.

Setup & Results

Data Generation

Chosen: Qwen-2.5-3B-Instruct

Rejected: Qwen-2.5-1.5B-Instruct

Performance: +0.2 avg

Matches OLMo 2 preference data

Our simple recipe generalizes across base models!

Additional Finding: Delta learning also succeeds with SimPO (alternative to DPO), yielding 5.2 point gain!

THEORETICAL ANALYSIS

Theoretical Analysis: Logistic Regression Setup

Why Theoretical Analysis?

To deepen intuition for why delta learning works, we analyze binary logistic regression where a student model is trained to prefer pseudo-labels from one teacher over another.

Problem Setup

Preliminaries

- Binary classification with **intercept-free logistic regression**
- Inputs: $x \sim N(0, Id)$ (isotropic Gaussian)
- Labels: $y^* \in \{0,1\}$ from ground-truth θ^*

Model Accuracy

Classification accuracy proportional to cosine similarity with θ^*:

$$\cos(\theta, \theta^*) = \langle \theta, \theta^* \rangle / \|\theta\|_2$$

Student & Teachers

- **Student θ_0**: model being improved
- **Teachers θ_c, θ_r**: with $\alpha_c > \alpha_r$ (θ_c stronger)

Training Procedure

1. Generate Labels

$$y_c = 1\{\langle \theta_c, x \rangle \geq 0\}$$
$$y_r = 1\{\langle \theta_r, x \rangle \geq 0\}$$

2. Preference Pair

Form pair (x, y_c, y_r) with annotation $y_c \succ y_r$

3. SGD Updates

Optimize naïve preference loss:

$$L_{pref} = -(\log p_\theta(y_c|x) - \log p_\theta(y_r|x))$$

Some annotations may be incorrect, but learning succeeds if y_c more correct than y_r on average.

THEORETICAL GUARANTEE

Theoretical Guarantee: Delta Learning Succeeds

Key Insight

Preference tuning pushes student towards **θc** and away from **θr**. Since **θc** better aligned with θ*, the difference vector θc − θr is positively aligned with θ* regardless of absolute alignment.

High-Dimension Effect

Orthogonal Errors

Teachers' errors are **essentially orthogonal** to student's errors in high dimensions.

Rare Amplification

Such amplification rarely happens, creating a training "sweet spot".

Improvement Guarantee

Student can improve from useful signal without overfitting to teachers' errors.

Theorem 6.1 (Informal)

Given student θ_0 and teachers θ_c, θ_r with $\alpha_c > \alpha_r$, if **Condition C1** holds:

$$\kappa = (\alpha_c - \alpha_r)(1 - \alpha_0^2) - \text{noise} > 0$$

Then training for **T steps** with proper hyperparameters yields:

$$\cos(\theta_T, \theta^*) > \cos(\theta_0, \theta^*) + \Theta(\kappa^2)$$

Corollary 6.2

In **high dimensions**, most teacher pairs with performance gap suffice to improve student.

Example

80% student, 70% & 60% teachers: **d > 2000** suffices for 90% of pairs.

Dimension threshold is **mild** for modern ML models.

PROOF SKETCH

Proof Sketch: Why Delta Learning Works

1 Exact Gradients

Population update traces parametric ray: $\ell(\lambda) = \theta_0 + \lambda v \Delta$

2 Empirical SGD

Control distance between exact iterate θ_T and SGD iterate θ_T using **martingale concentration**.

Geometric Analysis

Define Alignment Map

$$f(\lambda) = \cos(\ell(\lambda), \theta^*)$$

Improvement Condition

Learning succeeds if $f'(0) > 0$

Reduces to Condition C1

$$\kappa = (\alpha_c - \alpha_r)(1 - \alpha_0^2) - \text{noise} > 0$$

Quantify Gain

Improvement: $\Gamma \geq \Theta(\kappa^2)$

Concentration Analysis

Bound Deviation

With proper η, B, T, control:

$$\|\theta_T - \theta_T\|_2 \leqslant \Gamma/2$$

Triangle Inequality

Decompose improvement:

$$\text{Gain} \geqslant \Gamma - \Gamma/2 = \Gamma/2$$

Final Guarantee

SGD achieves $\geq \Theta(\kappa^2)/2$ improvement.

FUTURE RESEARCH

Limitations & Future Directions

Limitations

- **Limited scope:** Based on in-depth analysis of single algorithm (DPO) and few base models.
- **Evaluation coverage:** Doesn't capture all behaviors (multilingual, domain-specific).
- **Generalization:** Extending findings to other algorithms, larger scales, and new tasks needed.

Research Opportunities

1 Algorithm Extensions

Test with other preference tuning algorithms.

2 Scale & Model Varieties

Validate at **larger scales** and across **different architectures**.

3 Task Generalization

Test on new tasks beyond current benchmarks.

Open Questions

What Makes Delta Informative?

How to characterize semantic deltas that drive effective learning?

How to Scale Delta-Based Learning?

What are the limits as models and data grow?

Task & Algorithm Dependencies?

To what extent are dynamics dependent on specific tasks?

Improving Safety?

How to curate prompts and deltas that effectively improve safety?

Data Generation Strategies?

Can we generate targeted corruptions or lightweight human edits?

Exciting research directions ahead!

CONCLUSION

A New Paradigm for Learning

Core Contribution

We demonstrated that models can learn surprisingly well from the delta between paired weak data points, challenging conventional wisdom about strong supervision requirements.

Key Insights

Relative Quality Suffices

Relative quality differences provide sufficient learning signal even when absolute quality is weak.

Magnitude & Saturation

Delta magnitude predicts performance with saturation effects beyond threshold.

Simple Heuristics Work

Simple heuristics like model size differences suffice to create informative deltas.

Theoretical Understanding

Logistic Regression Proof

Delta learning works with **high probability in high dimensions**.

Directional Signal

Difference vector provides directionally correct signal even with weak teachers.

High-Dimension Protection

Orthogonality in high dimensions **protects against overfitting** to teacher errors.

Practical Recipe Impact

- Eliminates reliance on **strong model distillation**
- Enables **simpler, cheaper post-training**
- Matches Tülu 3 **without GPT-4o**

Delta learning opens new frontiers for efficient model improvement!

THANK YOU

Questions & Discussion

GitHub Repo Available

scottgeng@cs.washington.edu

> The delta between weak and weaker can be stronger than the sum of its parts
>
> — The Delta Learning Hypothesis

COLM 2025 | University of Washington, Allen Institute for AI, Carnegie Mellon University
Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Ranjay Krishna, Pang Wei Koh, Jerry Li