LLM POST-TRAINING LAB Part-2

From Base Model to Post-Training Tutor

Hiteshwari

SVNIT

PHASE 1 RECAP

Overview of Previous Building Blocks

RECAP OF PREVIOUS PHASE WORK

Base Model

Llama 3.2 3B Instruct 4-bit quantized architecture as the foundational starting point.

Stage 1: Pretraining

Continued Pretraining for domain adaptation using specialized corpus text data.

Stage 2: SFT

Instruction-response training utilizing LoRA/QLoRA for behavioral adaptation.

Stage 3: DPO

Preference alignment via chosen vs. rejected response optimization.

NEW PHASE ADDITIONS AND FOCUS

The Transition

Shift from **training the model** to **evaluating the model**. This phase establishes a rigorous testing framework for the post-training pipeline.

Core Focus Areas

- Multi-stage model evaluation
- LLM-as-a-Judge implementation
- Literature review on DPO scaling
- Held-out preference test sets

PROOF-OF-CONCEPT DATASET OVERVIEW

Stage	Dataset	Size	Purpose
Continued Pretraining	Domain Corpus	Small	Domain Adaptation
SFT	Instruction Set	100 samples	Instruction Following
DPO	Preference Set	60 pairs	Preference Alignment
Evaluation	Eval Questions	49 samples	Model Comparison
DPO Diagnostic	Held-out Pairs	31 pairs	Likelihood Test

Note: These represent proof-of-concept scale datasets for pipeline validation.

STRATEGIC DATASET SELECTION

SFT DATA (100)

Supervised Fine-Tuning

Teaches instruction following, answer structure, and domain-specific explanations.

DPO DATA (60)

Direct Preference

Refines behavior by distinguishing between preferred and rejected response quality.

EVAL QUESTIONS (49)

Rigorous Evaluation

Provides a fixed baseline to measure impact across all model versions.

COMPREHENSIVE EVALUATION PIPELINE

01

GENERATE

Run eval questions through Base, SFT, and DPO models.

02

JUDGE

LLM-as-a-Judge scores on 5 key dimensions.

03

COMPARE

Pairwise analysis between model checkpoints.

04

ANALYZE

Final diagnostic review and trend extraction.

LLM-AS-A-JUDGE OVERALL SCORES

+0.49

Base → SFT Gain

+0.04

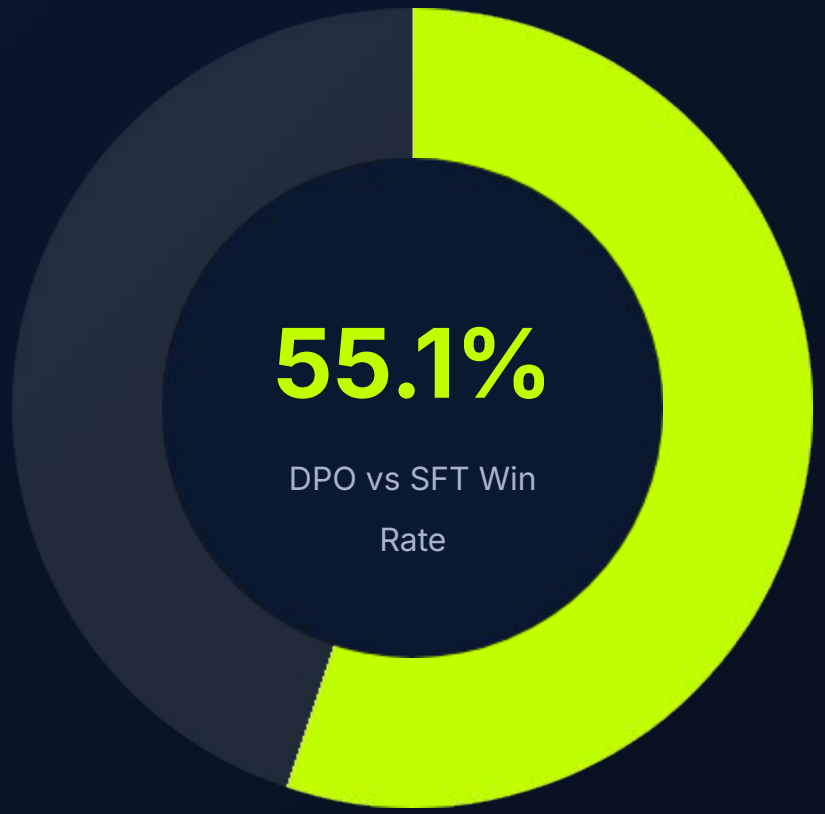
SFT → DPO Gain

DETAILED METRICS BY STAGE

Model	Correctness	Relevance	Clarity	Helpfulness
Base	3.65	4.22	3.59	3.47
SFT	4.08	4.45	4.10	3.98
DPO	4.12	4.49	4.16	4.06

💡 **Key Finding:** SFT produces the most significant leap across all evaluative dimensions.

PAIRWISE MODEL PREFERENCE RESULTS



SFT vs Base	**81.63%**
DPO vs SFT	**55.10%**
DPO vs Base	**81.63%**

Interpretation: DPO is preferred over SFT, but the margin remains modest in this proof-of-concept setting.

THE SFT-DPO IMPROVEMENT GAP

12.2x

GAIN MULTIPLIER

SFT provided over 12 times the numerical gain of DPO (0.49 vs 0.04).

The Central Research Question

Should DPO normally produce a large improvement over SFT, or is this small marginal gain expected for small backbones?

LITERATURE REVIEW: SMALL MODEL GAINS

> *"DPO yields small, task-dependent gains over strong SFT and can match competitive SFT accuracy without a warm start when the preference construction closely parallels the supervised objective."*
>
> **— Feng & Yang, 2026: Interaction in Small Language Models**

Finding

A large DPO improvement is not guaranteed in small-scale regimes.

Implication

Our +0.04 result is consistent with recent findings on marginal returns.

DRIVERS OF SIGNIFICANT DPO GAINS

Literature identifies specific conditions where

DPO achieves higher delta

01 Larger, more diverse preference datasets

02 Broadened evaluation benchmark scales

03 Higher complexity / task difficulty

04 Optimized preference-data construction

05 Modified objectives (e.g., Smaug / DPOP)

INITIAL DPO DATASET INVESTIGATION

Check 01

Uniqueness

60 pairs → 60 unique prompts.

No prompt duplication found.

Check 02

Similarity

TF-IDF analysis of Chosen vs Rejected responses:

0.146
Mean TF-IDF

0.131
Median

DATASET PROMPT OVERLAP ANALYSIS

EXACT OVERLAP STATUS

Measuring the shared prompts between training stages:

DPO Overlap 20%

SFT Overlap 12%

KEY INSIGHT

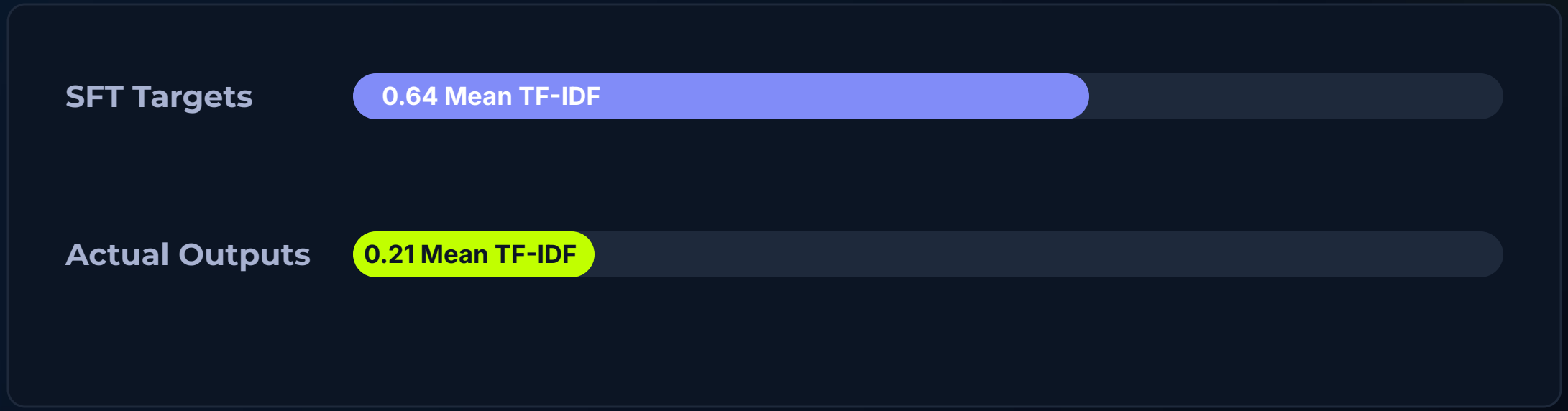
88 / 100

SFT instructions have no exact match in the DPO prompt set.

Note: Exact overlap does not capture conceptual coverage, indicating a potential distribution shift.

COMPARISON: LABELS VS. GENERATIONS

How similar are DPO's "chosen" responses to SFT's training targets versus actual model output?



Observation: DPO chosen responses are not strongly similar to actual SFT model generations.

DPO is not merely reproducing existing SFT behaviors.

KEY DATASET ANALYSIS TAKEAWAYS

Duplicate Analysis

No widespread duplicate DPO prompts or near-duplicate response pairs detected.

Prompt Overlap

Limited exact prompt overlap observed between the SFT and DPO stages (12-20%).

Target Similarity

DPO targets show low similarity to actual SFT outputs, suggesting novel preference signals.

Caveat: These results do not definitively prove that dataset coverage is the primary cause of the small DPO gain.

HELD-OUT PREFERENCE TEST RESULTS

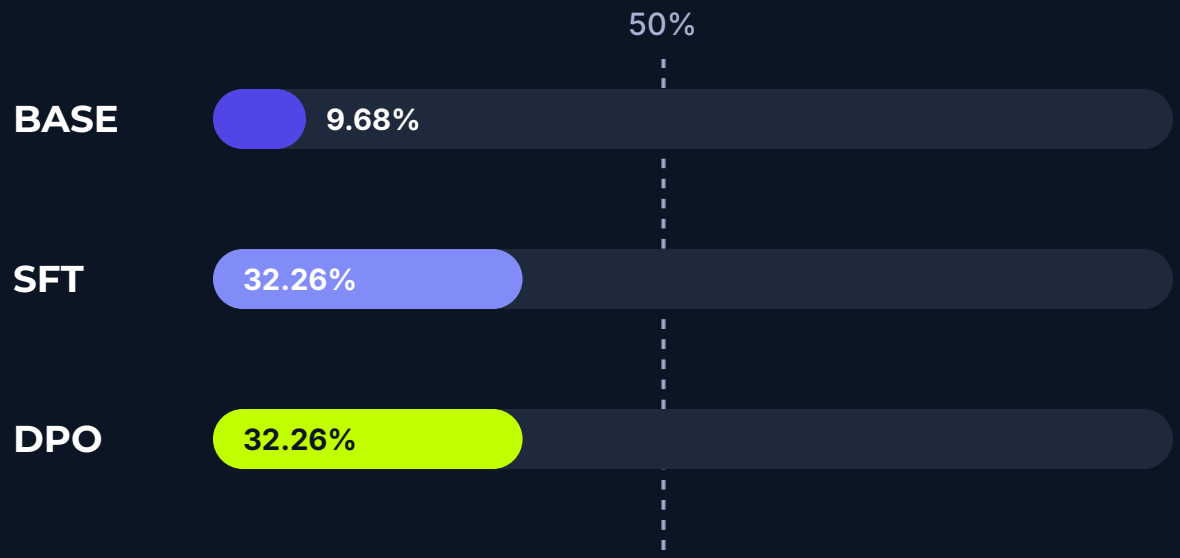
Methodology

Diagnostic test using 31 held-out pairs to evaluate model preference accuracy.

Decision Rule (Correct If):

```
logP(y_cho|x) > logP(y_rej|x)
```

Preference Test Accuracy



ACCURACY DIAGNOSTIC AND ANALYSIS

32.26%

STATIC ACCURACY

Key Observation:

Both fine-tuned models achieve **32.26%** accuracy, substantially outperforming the BASE model at **9.68%**.

However, performance remains below the 50% binary reference.

Required Validation Steps

01
Verify scoring directionality
Confirm correct scoring parameters and verify successful adapter loading.

02
Audit pair construction
Analyze and evaluate the integrity of the held-out validation pairs.

03
Analyze log-probability margins
Inspect raw margins and confidence deltas between evaluating models.

PROJECT ROADMAP AND NEXT STEPS

01

Validate Test

Audit the 31-pair preference results.

02

Compare Coverage

Covered vs. Uncovered eval analysis.

03

Scale Up

Implementation of larger training sets.

FUTURE PREFERENCE DATA STRATEGY

Key architectural questions for the next phase development:

DATA QUALITY & DIVERSITY

- How should quality be judged objectively?
- How diverse should preference pairs be?
- What filtering mechanisms for low-quality data?

DATASET SCALE & SPLITS

- How large should the preference dataset become?
- How to cleanly separate train/eval splits?
- Optimizing construction of rejected responses.

Strategic Goal: **Higher-quality + broader preference coverage.**

SCALING: CURRENT VS. NEXT PHASE

Component	Current (PoC)	Next Phase
SFT Dataset	100 examples	Large-scale SFT set
DPO Dataset	60 preference pairs	Diverse preference corpus
Evaluation	49 static questions	Extensive benchmark suite
Analysis	Initial Diagnostics	Deep behavioral evaluation
Environment	Google Colab	Scalable infrastructure

CORE PROJECT LEARNINGS

MODEL & METHODOLOGY

- Base → SFT improvement is robust and consistent.
- SFT → DPO gains are marginal but theoretically valid.
- Small model settings produce mixed preference behavior.

DATA & EVALUATION

- Current datasets serve as pipeline proofs only.
- Held-out tests require deeper margin analysis.

Strategic Next Priority: **Scaling data is key to unlocking further gains.**

FINAL VISION

RESEARCH INVESTIGATION FLOW

1. Build
2. Evaluate
3. Observe
4. Question
5. Investigate
6. Scale
7. Experiment

Core Research Question

How does data scale and preference optimization affect the incremental value of DPO over SFT?

THANK YOU FOR YOUR ATTENTION