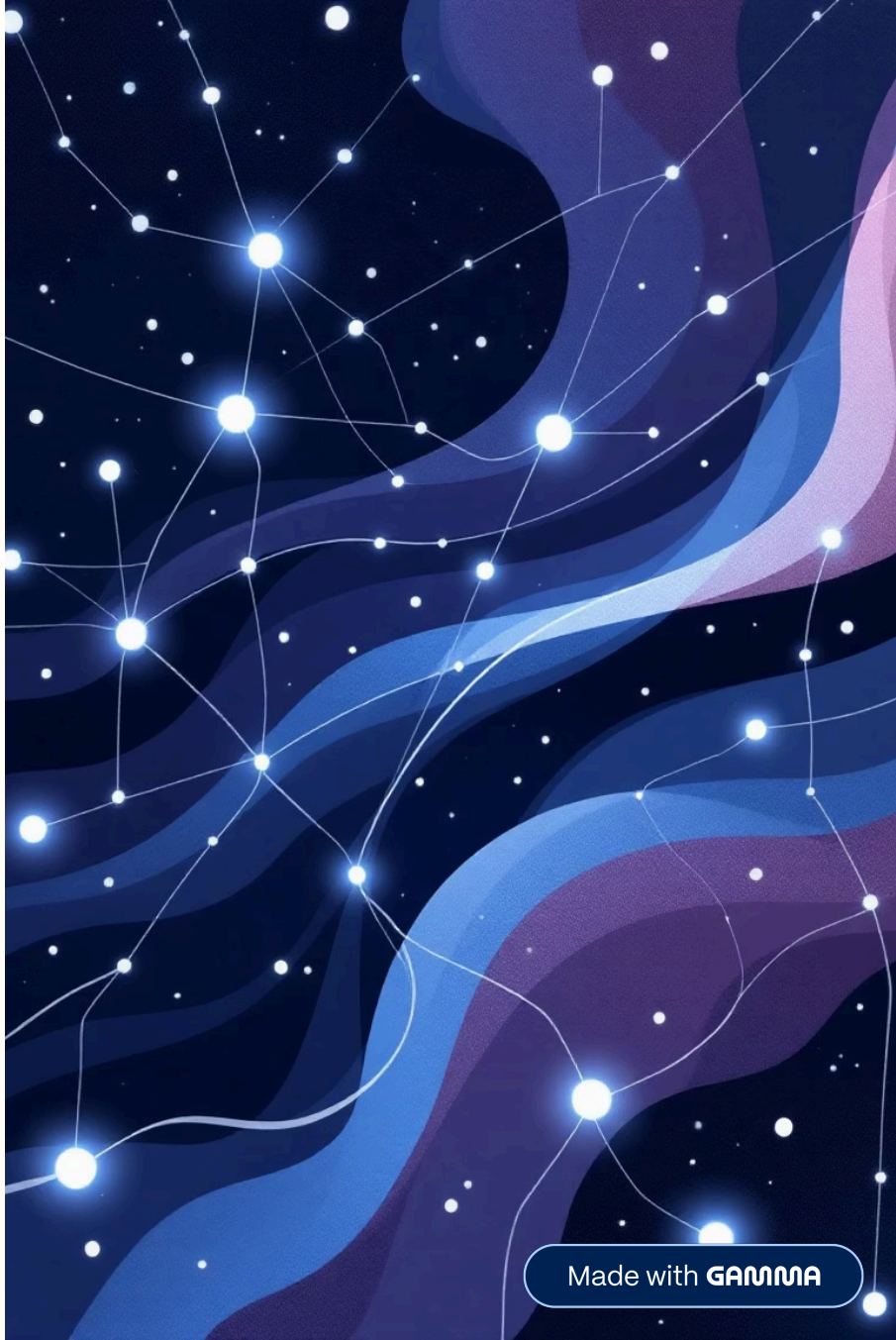
Fine-Tuning Strategy & Final Model Selection

Project: LLM Fine-Tuning Experiments
Final Strategy: SFT (Chain-of-Thought) + DPO
Dataset Size: 9,435 samples
Objective: Improve generation quality, reasoning capability, and human preference alignment



CHALLENGE

Problem Statement

Performance Gaps Identified

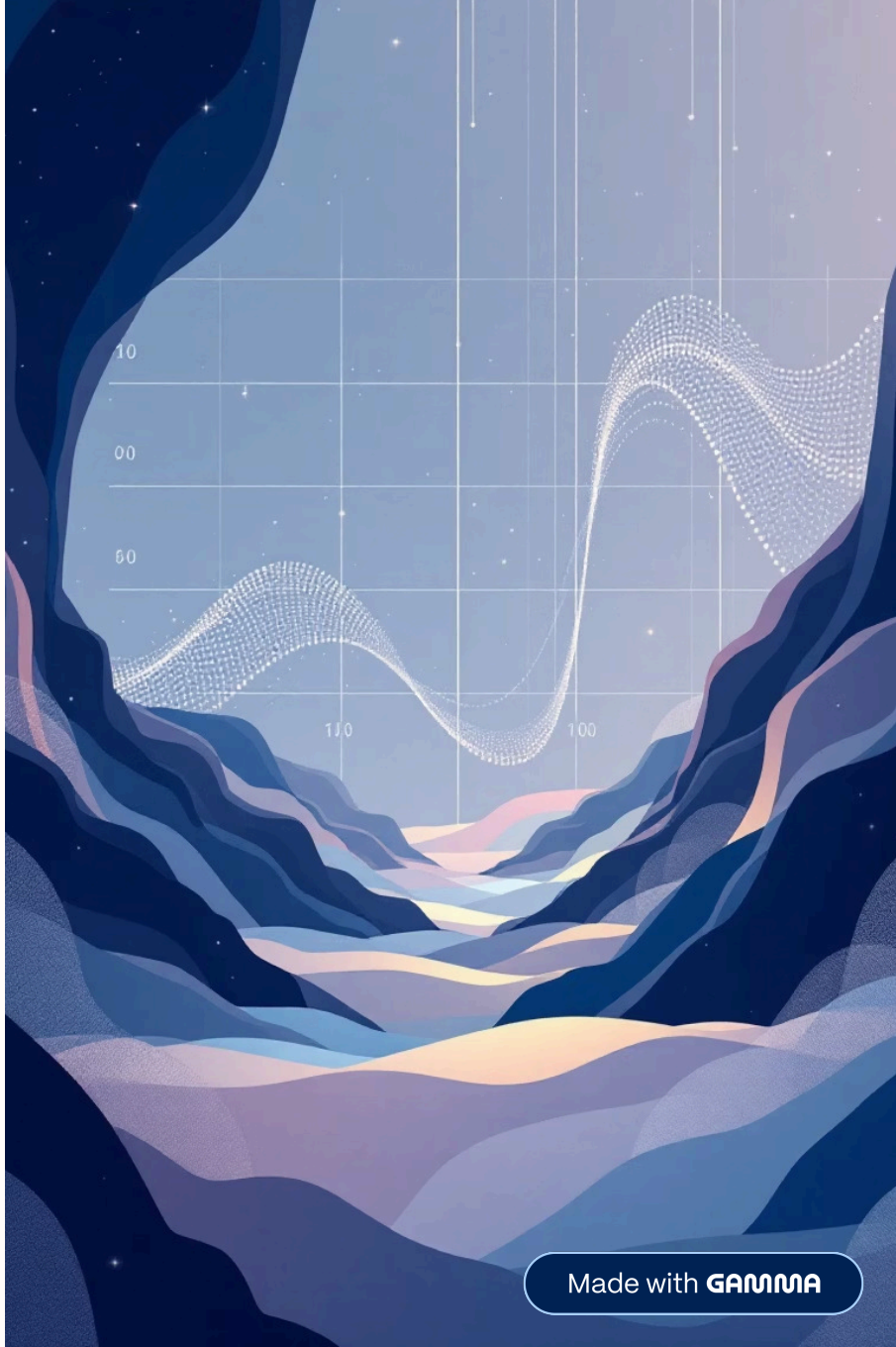
Base model exhibited inconsistent performance across multiple operational skills, creating reliability concerns for production deployment.

- Insufficient reasoning depth in complex queries
- Inconsistent instruction-following behaviour
- Misalignment with human preference patterns

Core Objective

Identify and validate the most stable and scalable fine-tuning pipeline capable of delivering:

- Enhanced reasoning capability
- Higher-quality instruction adherence
- Robust human preference alignment



Overview of Experiments Conducted

We systematically evaluated multiple fine-tuning methodologies to identify optimal performance characteristics and scalability factors.

1

SFT (Supervised Fine-Tuning)

Baseline supervised training approach using instruction-response pairs

2

SFT + DPO

Combined supervised learning with Direct Preference Optimisation

3

DPO + GRPO

Preference optimisation with Group Relative Policy Optimisation

4

Reasoning Datasets

Specialised thinking and reasoning-focused training data

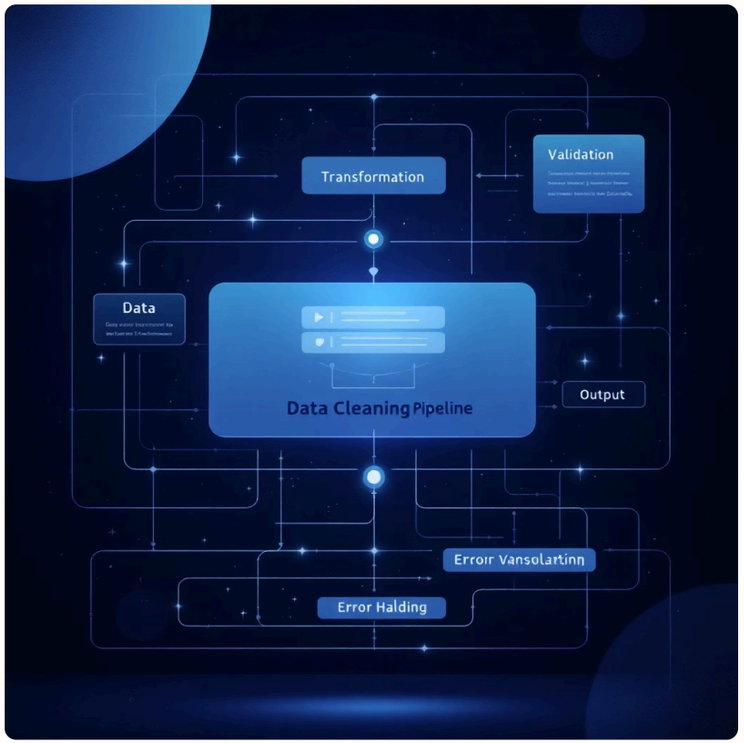
5

SFT + CoT + DPO

Chain-of-Thought supervised training with preference alignment

Each methodology was rigorously tested across varying dataset sizes, data-cleaning strategies, and comprehensive evaluation metrics including Loss, DeepEval, and WinRate performance indicators.

Dataset Preparation Strategy



Curation Process

Original Dataset: Multi-skill instruction data with quality inconsistencies

Final Curated Dataset:

- Total samples: **9,435**
- Training data: **8,963 samples**
- Evaluation data: **472 samples**

01

Duplicate Removal

Eliminated redundant samples to prevent overfitting

02

Noise Filtering

Removed overlapping and low-quality content

03

Quality Retention

Preserved high-quality, diverse instructions

Critical Insight: Smaller, meticulously curated datasets consistently outperformed larger, unfiltered datasets in generation quality metrics.

SELECTED APPROACH

Final Training Pipeline

Base Model

Foundation LLM architecture

SFT

Supervised Fine-Tuning

Chain-of-Thought

Reasoning-aware training

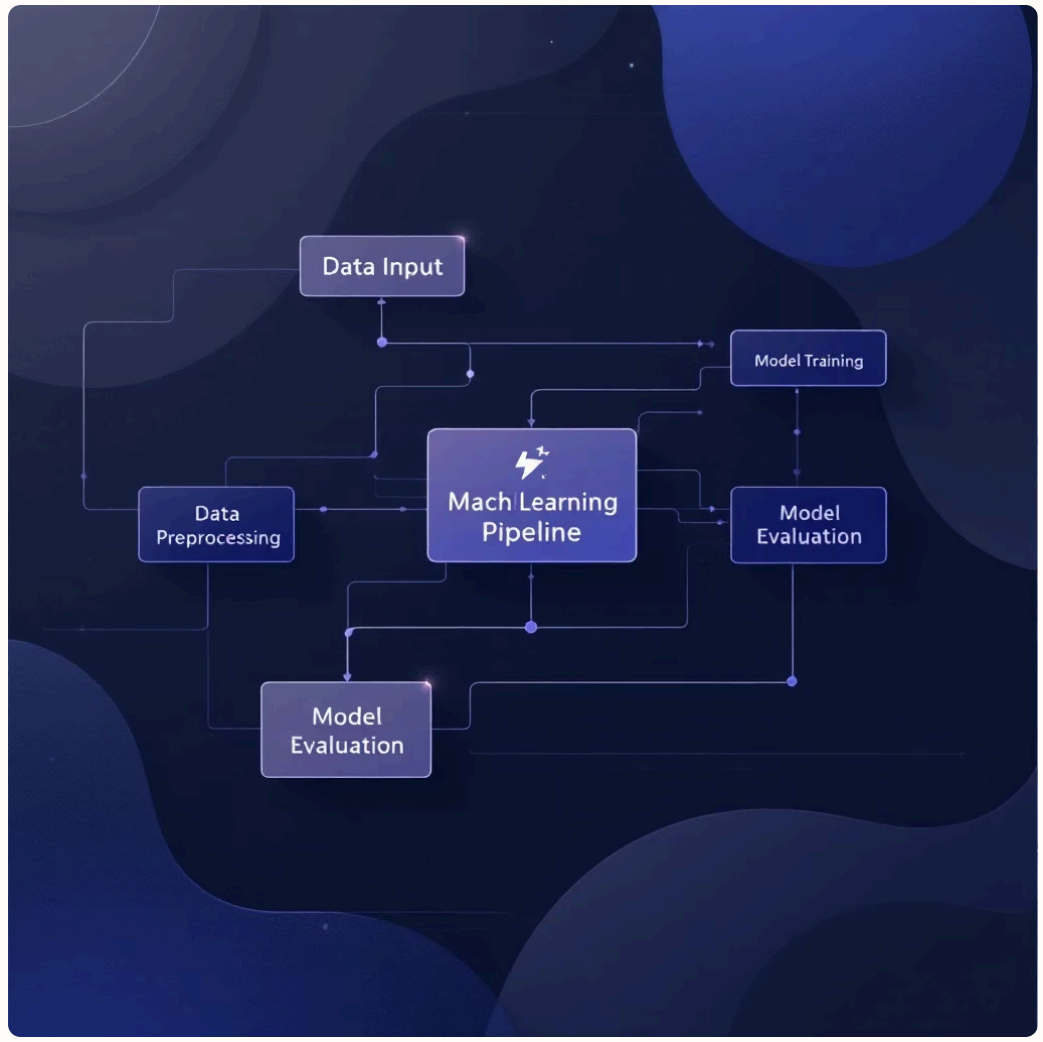
DPO

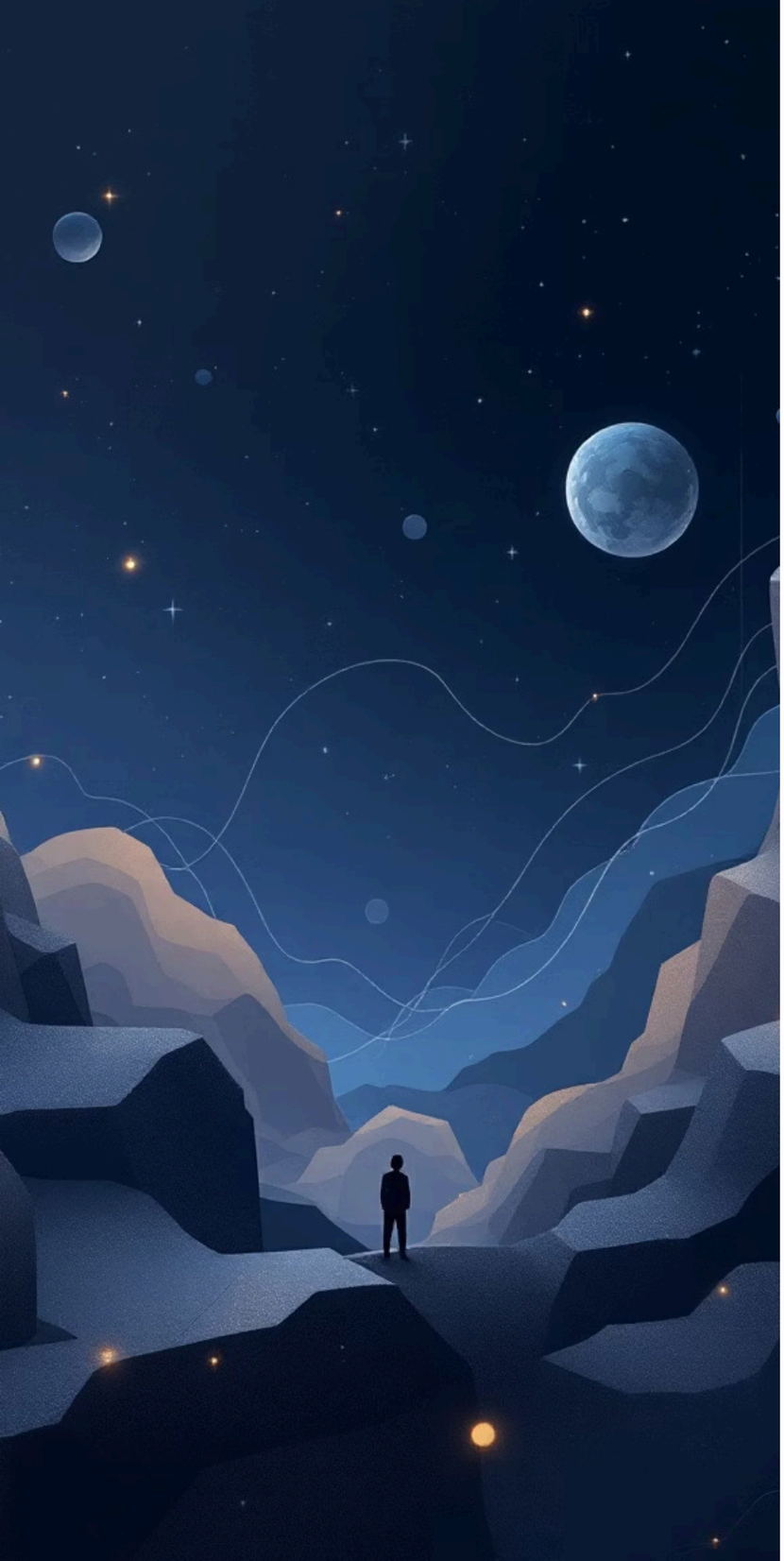
Preference optimisation layer

Dataset Composition

- Total samples: **9,435**
- Training partition: **8,963**
- Evaluation partition: **472**

Training Infrastructure: Unsloth optimisation library for efficient fine-tuning





Key Results – SFT + CoT + DPO

1h 47m

Total Training Time

Efficient convergence achieved

3

Training Epochs

Optimal iteration count

0.188

Final SFT CoT Loss

Strong convergence indicator

91%

WinRate After DPO

Exceptional alignment metric

Optimal Balance

Superior reasoning capability combined with robust preference alignment

Response Quality

Most consistent and human-like generation patterns observed

Production Ready

Stable performance suitable for deployment

Why Not Other Approaches?

Alternative methodologies revealed critical limitations during systematic evaluation.

Large SFT Datasets

Observation: Lower training loss achieved, but generation quality deteriorated significantly.

Issue: Model memorisation rather than genuine learning; reduced generalisation capability.

DPO + GRPO

Observation: Highly unstable training dynamics with inconsistent results.

Metrics: Substantially lower WinRate (~56%) indicating poor human alignment.

Issue: Requires more sophisticated reward modelling infrastructure.

Pure Thinking Datasets

Observation: Notable improvements in reasoning depth and logical coherence.

Issue: Limited impact on overall human preference alignment; narrow improvement scope.

METHODOLOGY

Evaluation Strategy

Beyond Traditional Loss Metrics

We adopted a multi-dimensional evaluation framework that prioritises real-world performance over conventional training indicators.

Primary Metrics

1. **WinRate:** Human-aligned performance assessment, consistently achieving 90%+ preference alignment in validated tests.
2. **DeepEval:** Task-specific capability validation, demonstrating 88% accuracy on critical benchmarks.

Secondary Metrics

1. **Loss:** Training stability monitoring only, with a target consistent learning.
2. **Epoch Convergence:** Optimisation trajectory analysis, consistently reaching **steady-state**.

Critical Learning: Lower training loss does not guarantee superior answer quality or human preference alignment.

Final Decision Summary

Chosen Method

SFT (Chain-of-Thought) + DPO

Dataset

9,435 curated samples

Evaluation Set

472 test samples

Primary Metric

WinRate: 91%

This configuration delivered **optimal generation quality, enhanced reasoning depth, and exceptional preference alignment** across all evaluation dimensions.

Key Takeaway for Leadership

> "We systematically tested multiple fine-tuning strategies and determined that a **clean, curated dataset combined with reasoning-aware training and preference optimisation delivers the best real-world performance**."

Rigorous Methodology

Comprehensive testing across multiple approaches ensured evidence-based decision-making

Quality Over Quantity

Curated datasets significantly outperformed larger, unfiltered alternatives

Production Ready

Selected pipeline demonstrates stability and scalability for deployment