# 1. Project Background and Objective

This project is based on the **Kaggle “LLM Classification Fine-tuning” competition**, focusing on classifying LLM prompts through fine-tuning.
Across multiple iterations, the system evolved from a simple baseline to a robust fine-tuning pipeline integrating **data cleaning, bidirectional augmentation, checkpoint resuming, and OOF evaluation**.

The main objectives were to:

- Reduce **LogLoss** (better calibration).
- Improve **F1** and **Accuracy**.
- Establish a **generalizable LLM fine-tuning and evaluation framework**.

---

# 2. Experimental Setup

| Component | Description |
|---|---|
| Python Environment | Python 3.10 + PyTorch 2.1 + transformers 4.37 |
| Hardware | NVIDIA A100 40GB |
| Base Model | DeBERTa-v3 / AutoModelForSequenceClassification |
| Dataset | JSONL format with text and label fields |
| Training Params | lr: 2e-5–1e-4; batch: 8–16; epoch: 1–5 |
| Evaluation Metrics | LogLoss, Accuracy, F1 (macro) |
| Framework | HuggingFace Trainer + datasets + evaluate |

---

# 3. Version Structure and Evolution

## 3.1 Overall Architecture per Version

| Version | Structural Overview |
|---|---|
| **v1.0** | DataLoader → Tokenizer → Trainer |
| **v5.0** | DataLoader → Trainer + Metrics (per-epoch eval) |
| **v7.0** | DataLoader → Trainer(resume) → A/B augmentation (prototype) |
| **v8.0** | DataLoader → Full A/B augmentation → Trainer |
| **v14.0** | DataCleaner (UTF-8 + flatten) → OOF Exporter |
| **v17.0** | Modular pipeline (load, tokenize, augment) + runtime monitor |
| **v18.0** | Bidirectional inference averaging + improved logging |

| Version | Structural Overview |
|---|---|
| **v20.0** | Simplified pipeline with cleaning temporarily disabled |
| **v22.0** | Restored cleaning pipeline + fast 1-epoch training |

## 3.2 Changes and Performance Comparison

| Version | Added / Modified Modules | Removed / Disabled | Param Changes | LogLoss/F1 | Explanation |
|---|---|---|---|---|---|
| **v1.0** | Baseline HF Trainer + compute_metrics | — | epoch=3 | 1.09 / 0.28 | Baseline, no enhancement |
| **v5.0** | Per-epoch eval + logging | — | batch=8→16 | 1.0858 / 0.347 | Smoother convergence |
| **v7.0** | resume_from_checkpoint + A/B proto | No UTF-8 cleaning | lr=3e-5 | 1.089 / 0.307 | Stability improved |
| **v8.0** | Full A/B augmentation | — | epoch=5 | 1.0895 / 0.361 | Better generalization |
| **v14.0** | utf8_clean + flatten + oof_export | — | lr=2e-5 | 1.093 / 0.25 | Enhanced data flow stability |
| **v17.0** | Modular pipeline, bidirectional logic | Removed old collator | batch=4 | 1.044 / 0.457 | Major boost |
| **v18.0** | Averaged A/B inference + improved logging | — | epoch=2 | 1.069 / 0.418 | Stable but slightly overfit |
| **v20.0** | Cleaning disabled, OOF off | utf8_clean / flatten commented | lr=1e-4 | 1.158 / 0.441 | Overfitting |
| **v22.0** | Restored utf8_clean + flatten | — | epoch=1 | 1.075 / 0.450 | Recovered stability |

# 4. Training Log Comparison

## v5.0

| Epoch | Train Loss | Val Loss | LogLoss | Acc | F1 |
|---|---|---|---|---|---|
| 1 | 1.0906 | 1.0895 | 1.0895 | 0.361 | 0.277 |
| 2 | 1.0706 | 1.0867 | 1.0867 | 0.394 | 0.318 |
| 3 | 1.0748 | 1.0858 | 1.0858 | 0.408 | 0.347 |

### v17.0

| Epoch | Train Loss | Val Loss | LogLoss | Acc | F1 |
|---|---|---|---|---|---|
| 1 | 0.974 | 1.068 | 1.068 | 0.431 | 0.438 |
| 2 | 0.932 | 1.044 | 1.044 | 0.460 | 0.457 |

### v20.0

| Epoch | Train Loss | Val Loss | LogLoss | Acc | F1 |
|---|---|---|---|---|---|
| 1 | 0.984 | 1.105 | 1.105 | 0.451 | 0.443 |
| 2 | 0.940 | 1.158 | 1.158 | 0.441 | 0.441 |

### v22.0

| Epoch | Train Loss | Val Loss | LogLoss | Acc | F1 |
|---|---|---|---|---|---|
| 1 | 0.967 | 1.075 | 1.075 | 0.452 | 0.450 |

# 5. Performance Trend Analysis

## Figure 1: LogLoss Trend (v1.0–v22.0)

> LogLoss dropped from 1.09 → 1.04 → 1.07 → 1.15 → 1.075.
> Best at v17.0; overfit at v20.0; stabilized again at v22.0.

## Figure 2: F1 Trend (v1.0–v22.0)

> F1 consistently improved with A/B augmentation and data cleaning.

## Figure 3: Accuracy–LogLoss Scatter

> Negative correlation between LogLoss and Accuracy indicates better calibration and generalization over time.

# 6. Feature Impact Validation

| Module | Introduced | Removed | LogLoss Δ | F1 Δ | Impact |
|---|---|---|---|---|---|
| UTF-8 Cleaning | 14.0 | 20.0 | ↑ +0.08 | ↓ 0.01 | Prevented text corruption |

| Module | Introduced | Removed | LogLoss Δ | F1 Δ | Impact |
|---|---|---|---|---|---|
| List Flattening | 14.0 | 20.0 | ↑ +0.05 | ↓ 0.02 | Preserved semantic structure |
| A/B Augmentation | 7.0–17.0 | — | ↓ −0.04 | ↑ +0.05 | Main improvement driver |
| Resume Checkpoint | 7.0 | — | Stable convergence | ↑ | Enhanced reproducibility |
| OOF Export | 14.0 | 20.0 | ↑ +0.02 | ↓ 0.01 | Enabled calibration stage |

## 7. Issues and Optimization Plan

### Key Issues

- Missing **EarlyStoppingCallback** → overfitting in v20.0.
- Lacked **temperature scaling** calibration.
- Class imbalance affected F1 stability.
- No multi-fold OOF validation.

## 🔥 Post-v17.0 Enhancement Roadmap & Execution Plan

To ensure a structured, controllable, and reversible evolution of the model, we propose a staged enhancement plan built on top of **v17.0**, following the principle of **incremental integration with switch-based components**, enabling safe A/B testing and rollback when needed.

### ✅ Overall Optimization Strategy

- **Principle**: Maintain the core v17.0 architecture while introducing improvements incrementally, each guarded by feature flags for safe experimentation.
- **Goal**: Improve model stability, reduce Log Loss, enhance long-context robustness, and build a scalable training pipeline suitable for future ensembling and multi-model fusion.

### 📌 Planned Optimization Roadmap (Phase-based)

| Phase | Focus Area | Key Enhancements | Expected Gains |
|---|---|---|---|
| **Stage 1: Stabilize & Correct** | Immediate stability & low-cost gains | A/B dual-order inference, temperature scaling, Early Stopping, UTF-8 cleaning | −0.02 LogLoss, improved training stability |

| Phase | Focus Area | Key Enhancements | Expected Gains |
|---|---|---|---|
| **Stage 2: Alignment & Generalization** | Train–test alignment & long-context robustness | Smart truncation + sliding window aggregation, Warmup + Scheduler, unified seeds and logging | −0.04~0.06 LogLoss on long samples, reduced variance |
| **Stage 3: Upper-bound Boost** | Model capacity and fusion enhancements | 5-fold OOF training, DeBERTa-large + LoRA, pairwise fusion, structured feature injection | Higher ceiling & improved performance on complex samples |

---

### 🧪 Validation Mechanism (Ensuring "Real Gain")

- **Multi-seed evaluation**: Each update is validated on ≥3 random seeds; report mean±std.
- **Bucket-based evaluation**: Report LogLoss/F1 across **Short / Medium / Long** samples and **A/B/TIE** classes to track targeted improvements.
- **OOF vs Public tracking**: Monitor the delta between OOF and Public/Private LB to prevent leaderboard overfitting.
- **Rollback anchor**: Keep the original v17.0 submission and logs as a reference baseline for regression checks.

---

### 🧭 Milestone-Based Delivery Plan

- **M1 (v17.1)**: Launch A/B dual-order inference + temperature scaling + Early Stopping
- **M2 (v17.2)**: Align train–inference length strategies & standardize text-cleaning module
- **M3 (v17.3)**: Introduce 5-fold OOF training + optional DeBERTa-Large + LoRA branch
- **M4 (v17.4, optional)**: Add Pairwise Reward Model fusion + explicit feature injection layer

---

## 8. Conclusion

Across 22 versions, the model evolved through four major phases:

1. **Foundation (v1.0–v5.0):** established Trainer and metrics.
2. **Stabilization (v7.0–v8.0):** resume checkpoint and initial augmentation.
3. **Enhancement (v14.0–v17.0):** complete data cleaning and modularization.
4. **Regression & Recovery (v18.0–v22.0):** temporary overfitting fixed via restored cleaning.

Final performance (v22.0):
**LogLoss ≈ 1.075, F1 ≈ 0.450, Accuracy ≈ 0.452.**

✅ **Core Finding:**
*Data cleaning and bidirectional A/B augmentation are the decisive contributors to model stability and performance.*

---

# 9. Appendix

## 9.1 Full Metrics per Version

| Version | LogLoss | Accuracy | F1 |
|---|---|---|---|
| 1.0 | 1.09 | 0.36 | 0.28 |
| 5.0 | 1.0858 | 0.408 | 0.347 |
| 7.0 | 1.089 | 0.384 | 0.307 |
| 8.0 | 1.0895 | 0.369 | 0.361 |
| 14.0 | 1.093 | 0.370 | 0.250 |
| 17.0 | 1.044 | 0.460 | 0.457 |
| 18.0 | 1.069 | 0.421 | 0.418 |
| 20.0 | 1.158 | 0.441 | 0.441 |
| 22.0 | 1.075 | 0.452 | 0.450 |

---

## 9.2 Feature Matrix

| Feature | 1.0 | 5.0 | 7.0 | 8.0 | 14.0 | 17.0 | 18.0 | 20.0 | 22.0 |
|---|---|---|---|---|---|---|---|---|---|
| Resume Training | | | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ |
| A/B Augmentation | | | ⚙️ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ |
| UTF-8 Cleaning | | | | | ✅ | ✅ | ✅ | ❌ | ✅ |
| List Flattening | | | | | ✅ | ✅ | ✅ | ❌ | ✅ |
| OOF Export | | | | | ✅ | ✅ | ✅ | ❌ | ✅ |
| Early Stopping | | | | | | | | | 🚫 |
| Temperature Calibration | | | | | | | | | 🚧 |

---