Bardic Fine-Tuning: Poetic Conversationalists

Akeela Darryl Fattha – akeelaf.2022@scis.smu.edu.sg

February 27, 2026

Abstract

This report details the fine-tuning of an open-weight Large Language Model (Mistral Nemo 12B) to respond to general conversational prompts exclusively in a poetic, bardic format. The project explores the transition from Llama 3.1 8B to Mistral Nemo, the curation of a custom poetic dataset, and the comparative analysis of LoRA and DoRA fine-tuning techniques across various hyperparameters.

1 Task Definition and Dataset

1.1 Objective

The objective of this project is to fine-tune an LLM to act as a ”Bard,” responding to user queries not with standard prose, but with structured, coherent, and contextually relevant poetry. The model must maintain conversational capabilities while strictly adhering to the poetic format trained.

1.2 Dataset Curation

The dataset was constructed in multiple phases to ensure high-quality poetic generation (see `notebooks/01_Data_Generator.ipynb` and `notebooks/01_Data_Refiner.ipynb` for implementation details):

- **Initial Seed:** Sourced from the Poem Comprehensive Dataset (PCD) (Yousef et al., 2018). Mistral Small Creative (via OpenRouter) was used to generate potential user prompts and English translations/meanings of the original verses.
- **Refinement and Expansion:** The initial dataset proved limiting. It was expanded to include various length poems that directly answer queries. Furthermore, specific subsets of standard conversational datasets were translated into poetic formats to teach the model how to handle normal chat interactions poetically. Specifically, the `wikihow` subset of the LIMA dataset was used to teach the model how to provide instructional responses in poetry, while the `Open QA`, `Closed QA`, and `Summarize` categories from the NoRobots dataset were selected to mitigate catastrophic forgetting of factual knowledge by teaching the model to embed factual answers within its poetic structures.

2 Experimental Setup

2.1 Model Selection

Initial experiments were conducted using Llama 3.1 8B (Base and Instruct). However, these models struggled to consistently adopt the poetic format and often failed to produce high-quality results regardless of the configuration. The base model was subsequently switched to the Unsloth-quantized Mistral Nemo 12B checkpoint (`unsloth/Mistral-Nemo-Base-2407`).

The increased parameter count and different architectural nuances of Mistral Nemo yielded significantly better adherence to the poetic constraints.

2.2 Fine-Tuning Methodology

The models were fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) methods, specifically Low-Rank Adaptation (LoRA) and Weight-Decomposed Low-Rank Adaptation (DoRA), utilizing the Unsloth library for optimized training.

The "control" configuration established for baseline comparisons utilized LoRA with a Rank of 32, Alpha of 64, a learning rate of 2e-4, and a cosine scheduler, trained on the full conversational data format. Variations in data formatting were also tested, including the Alpaca format and training exclusively on response loss (masking the user prompts). However, these alternative formatting approaches yielded inferior results compared to the standard conversational format.

Over 10 hyperparameter configurations were tested. Key variations explored relative to the Control baseline (Rank 32, Alpha 64, LR 2e-4, Cosine Scheduler, Weight Decay 0.001, conversational format) included:

- **Adapter Category:** LoRA and DoRA (Adapter).
- **Rank/Alpha:** 16/32, 32/64, and 64/128.
- **Learning Rate:** 2e-4 and 5e-4.
- **Weight Decay:** 0.001 and 0.01.
- **Dropout:** 0% and 5%.
- **Dataset Size:** Subset of 1k vs full 7k.
- **Formatting/Loss:** Conversational vs Alpaca format; full-conversation loss vs response-only loss.

3 Training Dynamics and Signals

After the dataset was upgraded, validation loss was introduced as the primary metric for checkpoint selection. Qualitative observations indicated severe overfitting when training on the full 7k sample dataset for extended periods, leading to infinite repetition loops during generation. Consequently, early stopping was employed, saving the model at the lowest validation loss (normally around 1k samples processed) rather than allowing it to plateau or degrade.

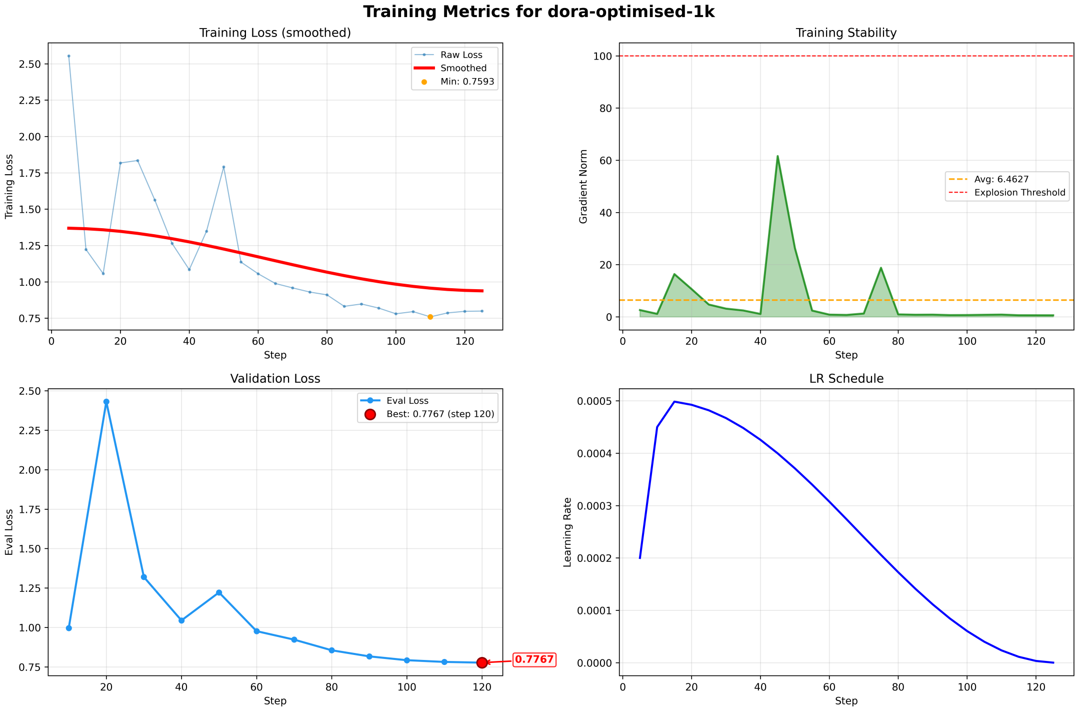

Figure 1: Training vs. Validation Loss Curves for the best performing model.

4 Evaluation

4.1 Methodology

Evaluation was conducted on a held-out set of 50 general conversational prompts. Because the outputs are highly subjective (poetry), a rubric-based quantitative metric was defined with concrete examples:

1. **Quality Score (1-5):** Ranging from 0 (Total Gibberish/Infinite Repetition) to 5 (Excellent, creative poetry with non-direct analysis).

4.1.1 Quality Rubric with Examples

Score	Example Output
0 – Gibberish	*"repetition repetition repetition repetition..."* (infinite loop or broken tokens)
1 – Basic Conversation	*"The answer to your question is yes. Yes, it is true."* (standard prose, no poetic structure)
2 – Prose-like	*"A journey begins with a single step, and steps lead us forward through time and space."* (metaphorical but lacks consistent structure or rhythm)
3 – Basic Poem-like	*"A step is taken, then another one, / The path ahead has just begun."* (clear poetic form with simple rhyme)
4 – Obvious, Structured	*"In morning's light the journey starts to bloom, / A single stride transcends the shadowed room, / Each step becomes a verse of forward grace, / A measured march through time and sacred space."* (A rhyme scheme employed, consistent meter, direct meaning)
5 – Excellent, Non-direct Analysis	*"The genesis lies not within the sole, / But in the courage housed within the soul; / To move is but to shed what binds us fast, / And every dawn erases every past."* (sophisticated language, non-obvious interpretation, thematic depth)

Table 1: Quality scoring rubric with representative examples from evaluation outputs.

4.1.2 Additional Metrics

2. **Format Adherence:** Binary metric indicating if the output was a recognizable poem (presence of line breaks, rhyme schemes, or poetic structure).

3. **Failure Rate:** Percentage of outputs resulting in infinite repetition or broken tokens.

4.2 Results

4.2.1 Quantitative Results

A total of 12 configurations were tested across LoRA, DoRA, and various hyperparameter combinations. Performance was evaluated on 50 diverse conversational prompts. Table 2 presents the complete results, with visual examples available in Appendix A.

Configuration	**Quality**	**Format (%)**	**Failure (%)**
Base (Zero-shot)	0.70	8.0	56.0
LoRA 16/32 (1k)	2.56	80.0	20.0
LoRA 32/64 (1k)	2.62	80.0	6.0
LoRA 32/64 (Full)	1.72	44.0	28.0
LoRA Fast Learner (5e-4)	1.86	50.0	26.0
LoRA Dropout 5% (1k)	2.08	58.0	14.0
LoRA Dropout 5% (Full)	2.56	74.0	6.0
LoRA Decay 0.01 (1k)	2.38	68.0	6.0
DoRA 32/64 (1k)	2.40	68.0	10.0
DoRA 32/64 Optimised (Full)	1.98	52.0	20.0
DoRA 64/128 (1k)	1.98	56.0	22.0
DoRA 64/128 (Full)	0.00	0.0	98.0
DoRA 32/64 Optimised (1k) (Dropout = 0, Decay = 0.001)	**2.80**	**82.0**	**2.0**

Table 2: Comprehensive results of choice hyperparameter configurations tested on 50 evaluation prompts. Bold indicates best performance in each metric.

4.2.2 Qualitative Observations

Beyond the numerical metrics, manual inspection of the generated outputs revealed distinct behavioral shifts across the different training configurations:

- **Base Model (Zero-Shot):** Without fine-tuning, the base model struggled significantly to maintain the poetic persona. Even with a strong system prompt, responses often felt like standard role-play rather than genuine poetry, and the model frequently reverted to normal prose or failed to follow the intended format entirely.

- **LoRA (Full Dataset vs. 1k Subset):** Training on the full 7k dataset resulted in severe degradation. The model frequently entered infinite repetition loops or leaked the training system prompt before outputting unrelated text. Conversely, the LoRA model trained on just 1k samples produced much higher quality outputs. The poems were coherent, comfortable to read, and demonstrated a good balance of length and structure, though they sometimes lacked deep analytical thought.

- **DoRA (1k Subset):** The DoRA configuration trained on 1k samples yielded the most impressive qualitative results. The generated poems exhibited varying lengths, strong rhyming schemes, and a non-direct, analytical approach to answering the prompts. While a small fraction (roughly 10%) still exhibited minor flaws, the overall quality and creativity were markedly superior to the LoRA counterparts. A subsequent run using the full set of optimizations (5% dropout, 0.01 decay, 5e-4 learning rate) on the full dataset still underperformed, indicating that the improvements did not transfer cleanly to full-data DoRA.

- **Impact of Dropout:** Interestingly, applying a 5% dropout rate during LoRA training on the full 7k dataset mitigated some of the catastrophic repetition seen in the standard full-dataset runs. The resulting poems were of varying lengths and legitimately high quality, suggesting that dropout effectively combated the severe overfitting observed earlier. **However**, this did not translate to the DoRA configurations, where the dropout might have caused the catastrophic failure when training on the full dataset, indicating that DoRA may be more sensitive to overfitting to self-destruction in this context.

5 Analysis and Discussion

5.1 Improvements and Best Performer

The DoRA configuration trained on 1k samples emerged as the best performer. It generated poems of varying lengths, exhibited great qualitative depth (non-direct analysis), and frequently maintained rhyme schemes.

5.2 Key Findings

- **System Prompt Necessity:** Even after fine-tuning, the models required a system prompt to trigger the poetic generation reliably. Without it, the models defaulted to standard prose or became incoherent. A potential reason for this is that the base model (Mistral Nemo) has strong pre-existing conversational priors. The fine-tuning process likely learned to associate the specific poetic behavior with the presence of the system prompt context, rather than overwriting the model's default conversational behavior entirely. Training on the response loss only always yielded in worse results.

- **Overfitting vs. Dataset Size:** Training on the full 7k dataset often led to catastrophic repetition. Reducing the training duration/samples (e.g., 1k samples) significantly improved generation quality and stopped the infinite looping, suggesting the model quickly memorized the format but struggled to generalize if pushed too far, despite the improved validation loss.

- **Rank/Alpha Impact:** Increasing Rank/Alpha from 16/32 to 32/64 did not drastically change the qualitative output, confirming that repetition issues were tied to overfitting rather than bottlenecked adapter capacity.

- **DoRA's Superiority for Language Tasks:** The decision to test Weight-Decomposed Low-Rank Adaptation (DoRA) was motivated by literature suggesting its learning patterns closely mimic full fine-tuning, making it particularly effective for complex reasoning and instruction tuning (Liu et al., 2024). This claim held true in our experiments; the DoRA configuration consistently produced more creative, structurally sound, and analytically deep poetry compared to standard LoRA, without adding any inference overhead.

5.3 Limitations and Failure Modes

A primary failure mode observed was the model entering an infinite repetition loop, especially when encountering concepts it struggled to articulate poetically. Additionally, some configurations produced poems that were excessively long, failing to output the EOS (End of Sequence) token. Factual responses also tended to degrade after fine-tuning; this is likely salvageable by increasing the proportion of factual data beyond the current 10% share in the dataset.

5.4 Future Improvements

To better guide the model during training and evaluation, future iterations should incorporate additional loss functions or automated metrics. For instance, a semantic similarity loss could be introduced to ensure the generated poem accurately reflects the intended meaning of the prompt. Alternatively, a secondary classifier model could be used to calculate a "poem-like" score, penalizing the model when it deviates into standard prose.

6 Reproducibility and Ethical Considerations

6.1 Reproducibility

All training was conducted using the `unsloth` library. Depending on the memory requirements of the configuration, training was executed either locally on a Windows 11 system equipped with an NVIDIA RTX 4070 Ti, or on a GPU cluster for configurations requiring at least 20GB of VRAM. The environment is managed via `uv`. The base model used is Mistral Nemo, and the dataset is a custom blend of PCD, LIMA, and NoRobots. Hyperparameters for the best run: DoRA, Rank 32, Alpha 64, LR 5e-4, Cosine Scheduler.

Code and Artifacts: All code, datasets, and trained adapters are available in the official GitHub repository: `https://github.com/acitea/fine-poems`. The repository contains all notebooks for dataset generation, training, evaluation, and inference, along with some pre-trained adapter weights and evaluation results.

6.2 Ethical Considerations

Fine-tuning an LLM to respond exclusively in poetry introduces unique risks. Harmful, biased, or dangerous information could be masked within the aesthetic appeal of a poem, potentially bypassing standard safety filters or making malicious content seem benign or profound. Future work should include a safety-alignment phase to ensure the ”Bard” refuses harmful requests gracefully rather than fulfilling them poetically.

7 References

- Yousef, W. A., Ibrahime, O. M., Madbouly, T. M., Mahmoud, M. A., El-Kassas, A. H., Hassan, A. O., & Albohy, A. R. (2018). *Poem Comprehensive Dataset (PCD).* `https://hci-lab.github.io/ArabicPoetry-1-Private/#PCD`

- Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., & Chen, M.-H. (2024). *DoRA: Weight-Decomposed Low-Rank Adaptation.* arXiv preprint arXiv:2402.09353. `https://arxiv.org/abs/2402.09353`

- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. (2023). *LIMA: Less Is More for Alignment.* Hugging Face repository. `https://huggingface.co/datasets/GAIR/lima`

- Rajani, N., Tunstall, L., Beeching, E., Lambert, N., Rush, A. M., & Wolf, T. (2023). *No Robots.* Hugging Face repository. `https://huggingface.co/datasets/HuggingFaceH4/no_robots`

- Unsloth AI. (2024). *Mistral-Nemo-Base-2407* (quantized). Hugging Face repository. `https://huggingface.co/unsloth/Mistral-Nemo-Base-2407`

A Quality Distribution Visualizations

This appendix provides quality distribution charts for all tested configurations, showing the frequency of each quality score (0-5) across the 50 evaluation prompts.

A.1 Base Model

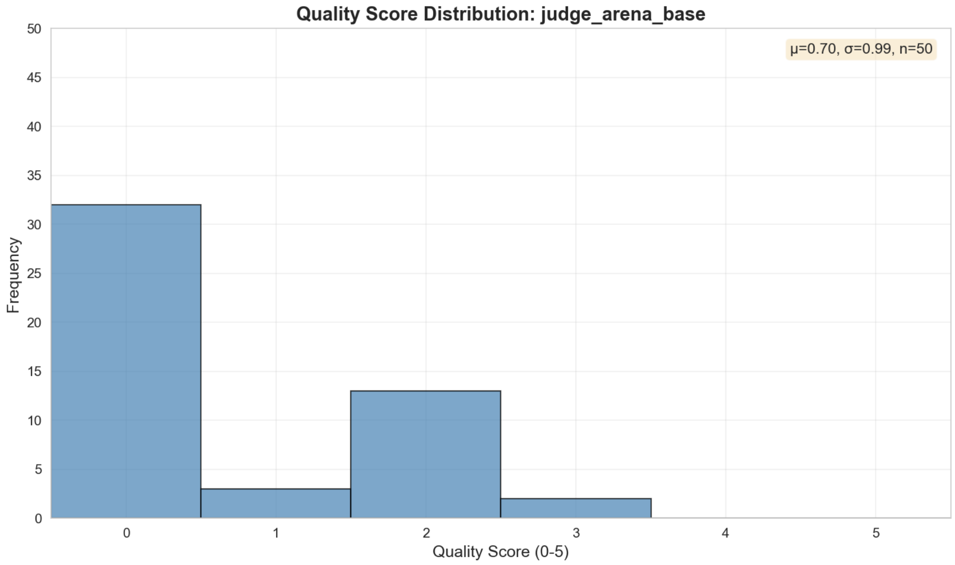

Figure 2: Quality distribution for base Mistral Nemo model (zero-shot with system prompt). Note the high concentration of low scores (0-1), indicating frequent failures.

A.2 LoRA Configurations

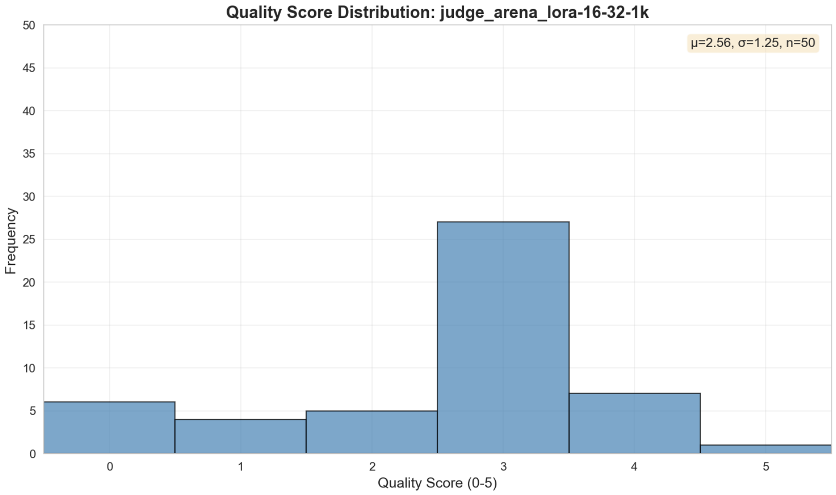

Figure 3: LoRA 16/32, 1k samples

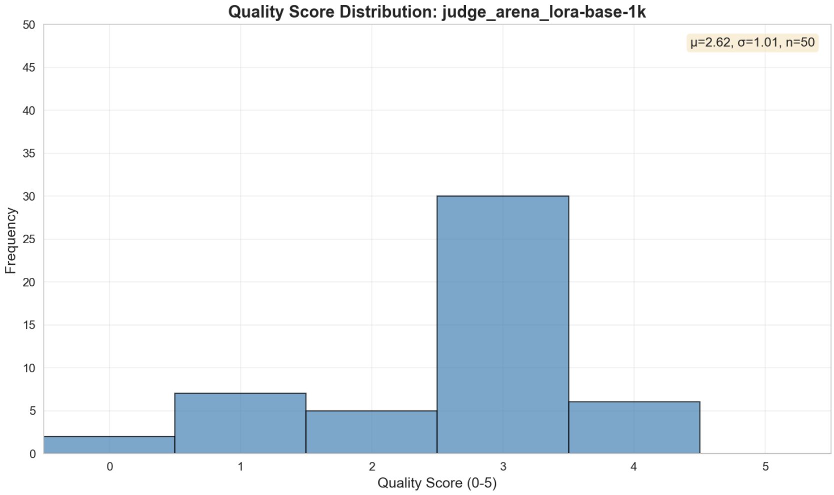

Figure 4: LoRA 32/64, 1k samples (Best LoRA configuration)

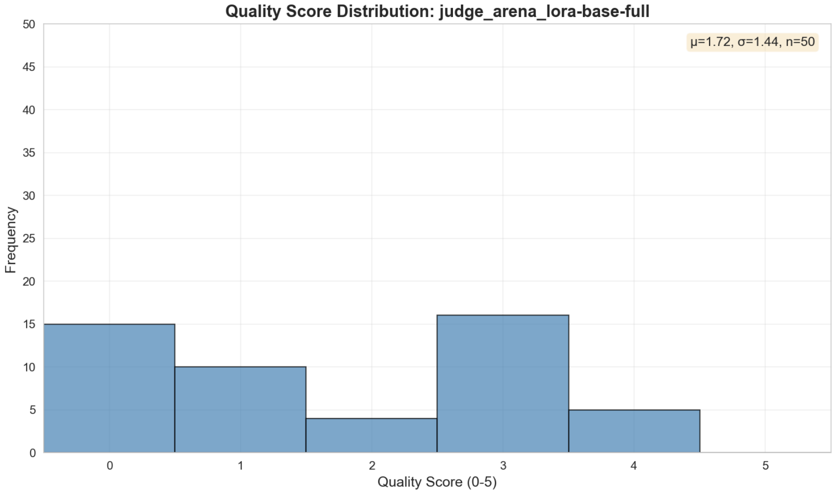

Figure 5: LoRA 32/64, Full dataset (7k samples). Shows increased failure rate compared to early stopping at 1k.

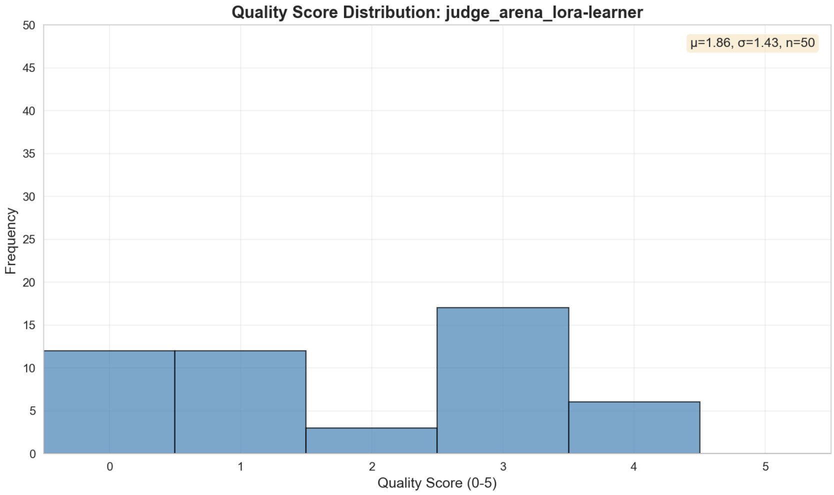

Figure 6: LoRA Fast Learner (LR 5e-4)

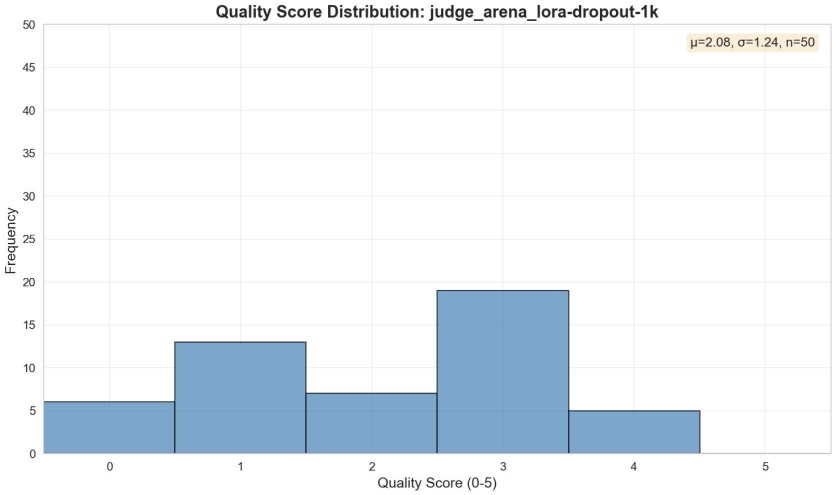

Figure 7: LoRA with 5% Dropout, 1k samples

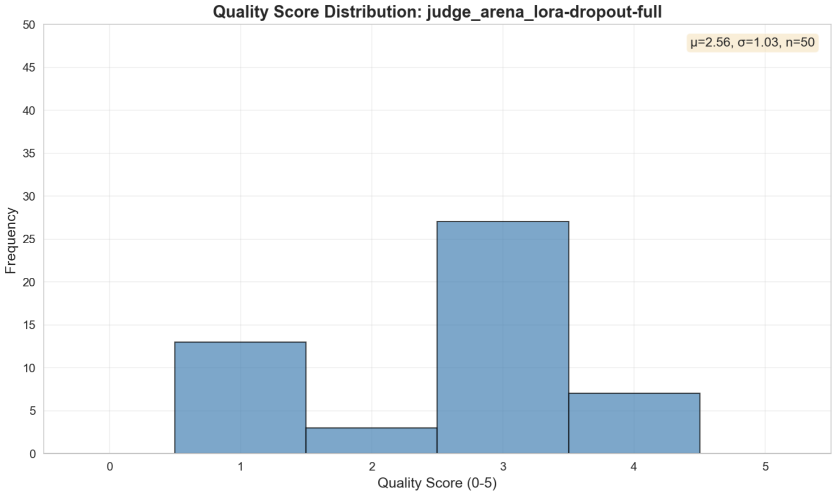

Figure 8: LoRA with 5% Dropout, Full dataset. Dropout mitigates some overfitting.

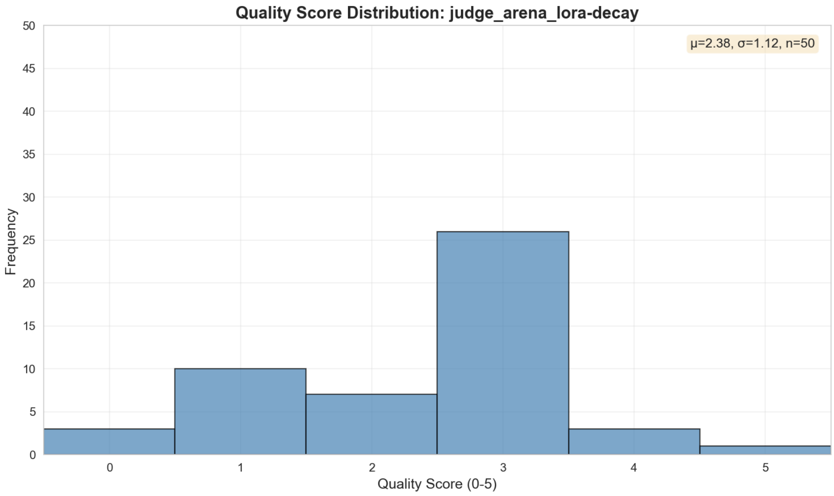

Figure 9: LoRA with Weight Decay 0.01

A.3 DoRA Configurations

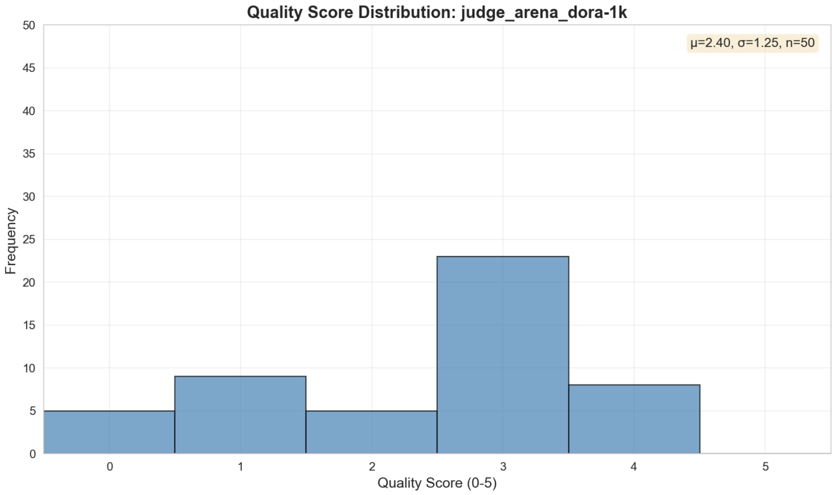

Figure 10: DoRA 32/64, 1k samples

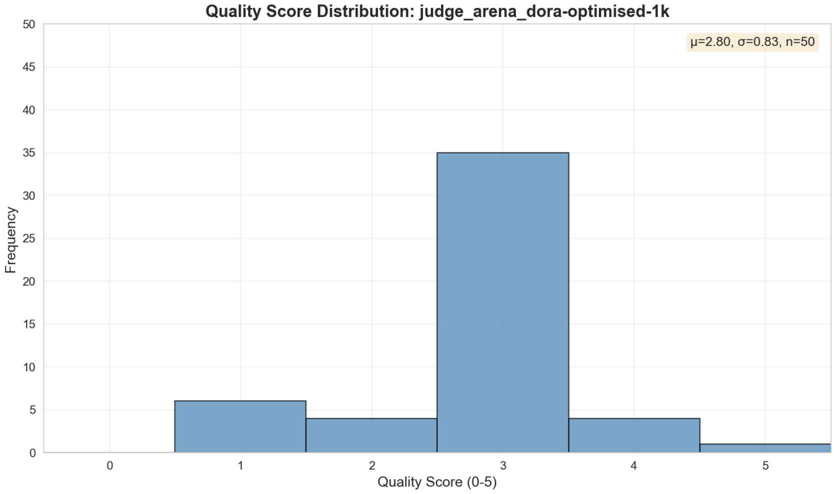

Figure 11: DoRA 32/64 Optimised, 1k samples, No Dropout, Decay 0.001, 5e-4 LR (Best overall configuration). Note the strong concentration in quality scores 2-4.

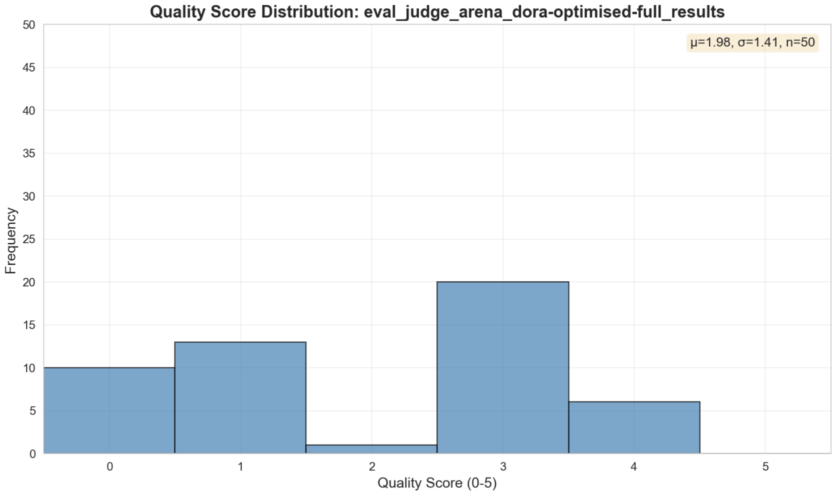

Figure 12: DoRA 32/64 Optimised, Full dataset (7k samples), Dropout 5%, Decay 0.01, LR 5e-4. Shows significant degradation compared to 1k samples, with increased failure rate and lower quality scores.

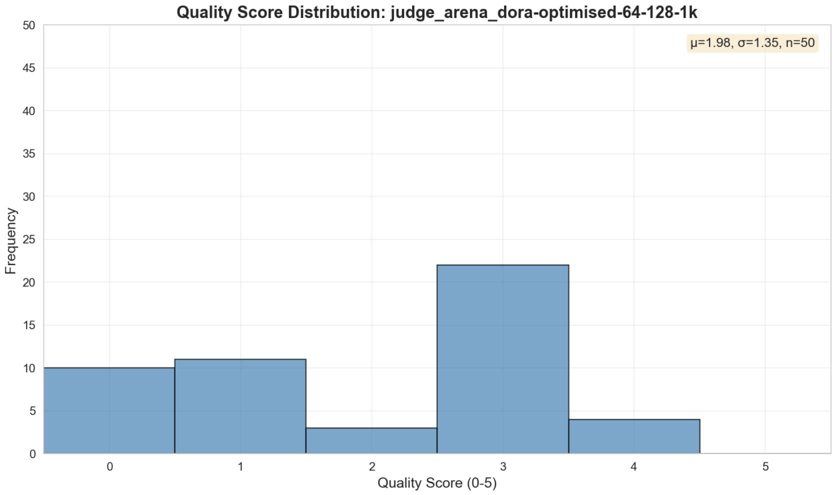

Figure 13: DoRA 64/128, 1k samples

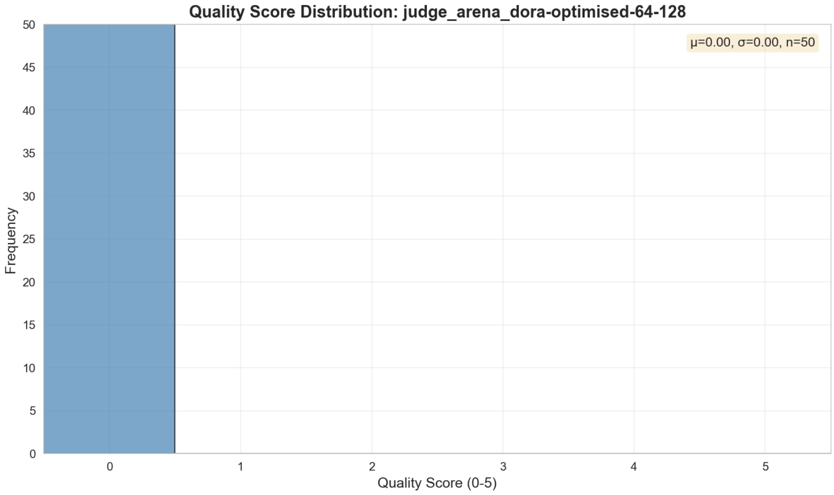

Figure 14: DoRA 64/128, Full dataset. Catastrophic failure with 98% failure rate.