LoRA Fine-Tuning for Immigration Case Outcome Prediction: A Benchmark on USCIS Administrative Appeals

Josue Diaz Flores
jdiaz6@scu.edu

Abstract

This paper documents what happened when I fine-tuned three popular open-weight language models (Gemma 4 E4B, Qwen 2.5 7B, and Llama 3.1 8B) on the same curated dataset of 1,467 USCIS Administrative Appeals Office (AAO) decisions using identical LoRA recipes. The short version: they responded very differently. EB-1A (extraordinary ability) visa adjudication is widely perceived as inconsistent across officers and service centers, which makes outcome prediction a genuinely non-trivial task. I curated the dataset through a five-stage pipeline of scraping, parsing, quality scoring, deduplication, and structured extraction, then fine-tuned each model and evaluated on a 260-case held-out test set. On accuracy, Gemma improved from 9.2% to 64.2%, Qwen regressed from 52.3% to 36.5%, and Llama went from 15.4% to 21.5%. Because the test set is heavily imbalanced (83.8% dismiss), I also report macro-F1. On macro-F1, Gemma LoRA is the strongest trained model at 0.341 and is the only trained configuration to beat an always-dismiss baseline (0.304), with sustain F1 more than twice that of a stratified-random baseline. However, no trained model cleanly beats stratified random (0.356) on aggregate macro-F1, indicating that at this data scale LoRA produces architecture-specific minority-class specialization rather than uniform improvement.

1 Introduction

I wanted to know what would happen if I took three of the most popular open-weight language models in the 7–8B range and fine-tuned all of them on the same legal dataset using the same LoRA recipe. Would they converge? Diverge? Would the smallest model surprise me? This paper is what I found.

I picked EB-1A and O-1A immigration decisions as the domain for three reasons. First, the USCIS Administrative Appeals Office publishes its decisions, so a real dataset exists to scrape. Second, EB-1A is one of the hardest and most sought-after immigration categories in the United States: it offers a path to permanent residency for individuals who can demonstrate sustained national or international acclaim, and the stakes for petitioners are high. Third, adjudication outcomes are widely perceived as inconsistent across officers and service centers even with published USCIS guidance, which makes outcome prediction genuinely non-trivial rather than a pattern-matching exercise.

This paper contributes:

1. A curated dataset of 1,467 AAO decisions on EB-1A/O-1A petitions, processed through a five-stage pipeline and released publicly on HuggingFace.

2. A LoRA fine-tuning benchmark comparing three open-weight models at similar parameter scale: Gemma 4 E4B, Qwen 2.5 7B, and Llama 3.1 8B.

3. An honest per-class analysis of the results, including comparison to simple baselines, that shows LoRA produces architecture-specific minority-class specialization rather than uniform improvement at this data scale.

A note on scope. This work was done independently as a hobby project. I am an undergraduate at Santa Clara University, and this is my first technical writeup of a research experiment. I ran single evaluation runs

without seed averaging, which limits the strength of any individual numerical claim; I flag this explicitly in the limitations section and try to keep the conclusions proportional to the evidence.

2 Dataset

2.1 Data Source

My data comes from the USCIS Administrative Appeals Office (AAO), which publishes non-precedent decisions on its website. I focus exclusively on decisions involving EB-1A (extraordinary ability) and O-1A (individuals with extraordinary ability or achievement) petitions. These decisions are public domain and represent the final administrative adjudication of appealed cases.

Note: USCIS stopped publishing new AAO decisions in March 2025. My dataset represents the complete available corpus at the time of collection.

2.2 Curation Pipeline

I process the raw decisions through a five-stage pipeline:

Stage 1: Scraping. I download all EB-1A/O-1A AAO decision PDFs from the USCIS website, yielding 4,643 documents. I handle pagination and rate limiting, generating a manifest CSV for tracking.

Stage 2: Parsing. I extract text from each PDF using PyMuPDF with Tesseract OCR fallback for scanned documents, achieving a 100% extraction rate (0 failures across 4,643 documents).

Stage 3: Quality Scoring. I use Claude Sonnet 4 to score each case against a rubric covering five dimensions: criteria coverage (30%), analytical depth (30%), evidence discussion (20%), legal reasoning (10%), and outcome clarity (10%). Cases scoring below 7.0 are excluded. Procedural dismissals, jurisdictional issues, and withdrawn appeals are auto-rejected. This stage yields 3,181 scored cases with a mean quality score of 7.66 (range: 7.0–9.1).

Stage 4: Deduplication. I remove near-duplicate decisions using TF-IDF cosine similarity with a threshold of 0.95, retaining the higher-scored copy. This removes 18 duplicate pairs, producing 1,467 final cases.

Stage 5: Structured Extraction. I use Claude Sonnet 4 to decompose each decision into structured components: petitioner background, evidence per criterion, AAO analysis per criterion, outcome reasoning, legal citations, and any fraud or procedural issues.

2.3 Training Data

From the 1,467 curated cases, I generate multi-task training examples in ChatML/ShareGPT format across four task types:

Task Type	Train	Total
Single criterion analysis	6,961	8,716
Outcome prediction	1,176	1,466
Gap identification	1,150	1,435
Criteria analysis	1,179	1,466
Total	**10,466**	**13,083**

Table 1: Training data distribution by task type. Data is split 80/10/10 into train/validation/test.

Class Imbalance. The outcome distribution is heavily skewed: 84.6% of cases are dismissals, with sustain and remand comprising the remaining 15.4% (Table 2).

Outcome	Train	Val	Test
Dismiss	1,111	132	147
Sustain	38	6	1
Remand	27	2	2
Total	**1,176**	**140**	**150**

Table 2: Outcome distribution across splits. The severe class imbalance (84.6% dismiss) means a majority-class baseline achieves ~84% accuracy.

Criteria Distribution. The dataset covers all 10 regulatory criteria, with "original contributions" being the most frequently discussed and "commercial success" the least (Table 3).

Criterion	Train Count
Original contributions	942
Awards	909
Published material	899
Leading role	895
Membership	802
Judging	778
Scholarly articles	589
High salary	499
Exhibition	452
Commercial success	196

Table 3: Frequency of EB-1A criteria discussed in training examples.

3 Method

3.1 Model Selection

I evaluate three open-weight language models spanning different architectures and parameter scales:

- Gemma 4 E4B: Google's mixture-of-experts model
- Qwen 2.5 7B (Qwen Team, 2024): Alibaba's 7B-parameter dense model
- Llama 3.1 8B (Touvron et al., 2023): Meta's 8B-parameter dense model

All models use their instruct-tuned variants as the base for LoRA adaptation.

3.2 LoRA Configuration

I apply LoRA (Hu et al., 2022) to the attention projection matrices (Q, K, V, O) of each model.

Hyperparameter	Value
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05

Table 4: LoRA fine-tuning hyperparameters.

3.3 Evaluation Protocol

I evaluate on a held-out test set of 260 cases. Because the outcome distribution is heavily imbalanced (83.8% dismiss, 11.9% remand, 4.2% sustain), I report both accuracy and macro-F1. Accuracy is dominated by the majority class and can be trivially achieved by an always-dismiss classifier, while macro-F1 weights each class equally and is a more faithful measure of minority-class learning. I also report per-class F1 for dismiss, sustain, and remand. I compare all trained models against two baselines: an always-dismiss classifier and a stratified-random classifier that samples labels from the training distribution with a fixed seed. A small number of predictions across all runs fail regex extraction and are bucketed as “unknown.” These count as incorrect against all trained models uniformly.

4 Results

4.1 Overall Performance

Table 5 summarizes the benchmark results across all three models in both base and LoRA-adapted configurations.

Model	Config	Acc.	Macro-F1	Dismiss F1	Sustain F1	Remand F1
Gemma 4 E4B	Base	9.2%	0.068	0.119	0.085	0.000
	LoRA	64.2%	0.341	0.783	0.239	0.000
Qwen 2.5 7B	Base	52.3%	0.331	0.691	0.157	0.146
	LoRA	36.5%	0.256	0.527	0.179	0.063
Llama 3.1 8B	Base	15.4%	0.161	0.233	0.094	0.156
	LoRA	21.5%	0.192	0.324	0.122	0.130
Always-dismiss		*83.8%*	*0.304*	*0.912*	*0.000*	*0.000*
Stratified-random		*72.3%*	*0.356*	*0.843*	*0.105*	*0.119*

Table 5: Outcome prediction results on 260 held-out test cases. Accuracy is dominated by the 83.8% dismiss majority class, so I also report macro-F1 (unweighted average of per-class F1) which better reflects minority-class learning. Gemma LoRA is the strongest trained model on macro-F1 and the only one to beat the always-dismiss baseline, though no trained configuration cleanly beats stratified-random on aggregate macro-F1.

4.2 Outcome Prediction

Table 5 tells two different stories depending on which metric is emphasized. On raw accuracy, Gemma LoRA shows the largest improvement of any configuration, jumping +55.0 points from 9.2% to 64.2%. However, even this best trained result sits nearly 20 points below the 83.8% achieved by a trivial always-dismiss classifier. Because the test set is 83.8% dismiss, accuracy is largely a measure of how closely a model matches the majority prior, not how well it learns the minority classes that matter most for adjudication review.

Macro-F1 gives a fairer picture. Gemma LoRA achieves 0.341, the highest macro-F1 of any trained model and the only trained configuration to beat the always-dismiss baseline of 0.304. Its sustain F1 of 0.239 is more than 2× the stratified-random baseline (0.105), driven by correctly identifying 8 of 11 true sustain cases (72.7% recall). Qwen base is a close second at 0.331, and is the only trained configuration with non-trivial remand F1 (0.146). Gemma base, by contrast, collapses to 0.068 macro-F1: despite predicting “sustain” for 224 of 260 cases, it gets sustain F1 of only 0.085 because its precision is near zero.

Gemma 4 E4B. LoRA corrects a severely miscalibrated prior. The base model predicts sustain for 86% of cases, far from the 4.2% true rate, and gets almost nothing right as a result. LoRA shifts sustain

predictions down to 56 cases, closer to the true distribution, and produces the strongest minority-class F1 of any trained model. As a side effect, Gemma LoRA is also 2.68× faster than Gemma base at inference (1,948s vs. 5,225s for 260 cases) despite carrying adapter weights, likely because it produces shorter, more confident outputs (2,854 vs. 3,663 average tokens).

Qwen 2.5 7B. LoRA actively damages a previously reasonable model. Qwen base is the second-strongest configuration on macro-F1, but LoRA drops it to 0.256 by reducing dismiss F1 (0.691 to 0.527) and collapsing remand F1 (0.146 to 0.063), while only marginally improving sustain F1 (0.157 to 0.179). The net effect is that Qwen LoRA performs worse on every outcome class except sustain, and even there the improvement is small.

Llama 3.1 8B. Modest gains across the board (macro-F1 0.161 to 0.192), but both configurations remain below the always-dismiss baseline. Llama base achieves the highest remand F1 (0.156) of any configuration, though this advantage is small in absolute terms.

Comparison to stratified random. A stratified-random classifier that samples labels from the training distribution achieves macro-F1 of 0.356, narrowly above Gemma LoRA (0.341) and ahead of every other trained configuration. This does not invalidate the trained results. Gemma LoRA still dominates on sustain F1 (0.239 vs. 0.105), and its performance reflects genuine though imperfect learning rather than distribution matching. But it does indicate that at this data scale, no trained model has learned the full minority-class structure cleanly. I return to this point in Section 5.3.

5 Discussion

5.1 Why Does LoRA Help Gemma But Hurt Qwen?

The divergent responses to identical LoRA training data across architectures is the central finding of this work. The per-class F1 breakdown in Table 5 allows a more specific diagnosis than accuracy alone permits.

Base Model Calibration. Gemma base and Qwen base begin from very different priors. Gemma base predicts “sustain” for 86% of cases, which is catastrophically miscalibrated given the 4.2% true sustain rate, producing near-zero F1 across all classes. Qwen base, by contrast, is already reasonably calibrated and achieves macro-F1 of 0.331. For Gemma, LoRA has room to correct a broken prior; for Qwen, there is little prior to correct, and the fine-tuning signal instead disrupts an already-functioning calibration.

Direction of the Qwen Regression. Qwen LoRA does not simply overfit to the dominant “dismiss” class as might be expected from the 83.8% imbalance. Instead, the per-class F1 breakdown shows LoRA raising Qwen’s sustain F1 slightly (0.157 $\rightarrow$ 0.179) while collapsing its remand F1 (0.146 $\rightarrow$ 0.063) and reducing its dismiss F1 (0.691 $\rightarrow$ 0.527). The net effect is sustain over-prediction at the cost of both other classes, suggesting LoRA is not pulling Qwen toward the majority class but toward a spurious pattern in the sustain training examples that generalizes poorly.

Architecture Sensitivity. Gemma 4 E4B uses a mixture-of-experts architecture, which may respond differently to low-rank updates than the dense transformer architectures of Qwen and Llama. I cannot distinguish this hypothesis from the base-calibration hypothesis without training ablations, but the fact that Llama (dense) shows only modest gains while Gemma (MoE) shows the largest absolute improvement is suggestive.

Training Data Scale. With ~10K training examples, the dataset is small relative to typical LoRA fine-tuning budgets. The observation that no trained model cleanly beats stratified-random baselines on aggregate macro-F1 suggests the dataset may be below the effective threshold at which dense models can learn minority-class structure reliably.

5.2 Class Imbalance and Metric Choice

The 83.8% dismiss rate in the test set has substantial implications for metric choice. An always-dismiss classifier achieves 83.8% accuracy, which is higher than any trained model in this benchmark, but gets macro-F1 of only 0.304 because it produces zero F1 on sustain and remand. Accuracy therefore rewards models that default to the majority class without doing any meaningful legal reasoning, which is the opposite of what a useful adjudication-assistance tool should do.

Macro-F1 is a more appropriate headline metric for this task. Under macro-F1, Gemma LoRA is the strongest trained model at 0.341, the only trained configuration to beat the always-dismiss baseline, and has sustain F1 more than twice the stratified-random baseline. Qwen base is a close second. The +55 accuracy-point improvement for Gemma LoRA over its base is real, but it is better understood as evidence that LoRA corrected a broken prior than as evidence that the resulting model is usable in deployment.

Three caveats are worth making explicit. First, Gemma LoRA's sustain precision is only 14.3%: it flags 56 cases as sustain for every 8 correct identifications, so any deployment would require human review of a large false-positive pool. Second, remand is effectively invisible to every model; only Qwen base and Llama base produce remand F1 above 0.14, and none of the LoRA configurations improve on this. Third, a stratified-random classifier (macro-F1 0.356) narrowly beats every trained model on aggregate, which is discussed below.

5.3 Limitations

Relationship to simple baselines. A stratified-random classifier achieves macro-F1 of 0.356, narrowly higher than Gemma LoRA (0.341) and every other trained configuration. Stratified random is a strong baseline on small imbalanced datasets because it exactly matches the class prior without learning anything, and its per-class F1 scores are smoothed across all classes rather than concentrated. Trained models can still win on individual classes. Gemma LoRA's sustain F1 of 0.239 is 2.3× the stratified-random baseline of 0.105, reflecting genuine though imperfect minority-class learning. But no trained configuration Pareto-dominates stratified random across all three classes. This suggests that at this data scale, LoRA produces architecture-specific class specialization rather than uniform improvement over the prior.

- **Dataset scope.** AAO decisions represent appealed cases, which are not representative of all EB-1A adjudications. Cases decided at the field office level (the vast majority) are not included.
- **Temporal coverage.** USCIS stopped publishing AAO decisions in March 2025. Policy changes after this date are not reflected.
- **Class imbalance and remand learnability.** The severe skew toward dismissals limits my ability to evaluate sustain and remand performance robustly. Remand in particular is effectively unlearnable at this data scale: no trained configuration exceeds remand F1 of 0.16, and the best remand F1 belongs to an untrained base model (Llama 3.1 8B at 0.156).
- **Sustain precision.** Even the best-performing configuration (Gemma LoRA) has sustain precision of only 14.3%, meaning deployment would generate a large false-positive pool for any sustain-flagging workflow.
- **LLM-generated labels.** Case quality scores and structured components were extracted using Claude Sonnet 4. Errors or biases in the extraction model propagate into the training and evaluation data.
- **Single evaluation run.** I report single-run results without confidence intervals. Variance across random seeds may be significant, particularly for sustain and remand metrics computed over 11 and 31 test cases respectively.

6 Conclusion

I set out to answer a simple question: what happens when you fine-tune three popular open-weight models on the same specialized legal dataset using the same LoRA recipe? The answer turned out to be more interesting than I expected. On accuracy, Gemma 4 E4B improves by +55 points, Llama 3.1 8B shows modest gains, and Qwen 2.5 7B regresses by 15.8 points. On macro-F1, Gemma LoRA is the strongest trained configuration (0.341) and the only one to beat an always-dismiss baseline, with sustain F1 more than twice that of a stratified-random baseline.

However, no trained configuration cleanly beats stratified random on aggregate macro-F1. This does not mean the trained models learned nothing: Gemma LoRA in particular makes genuine minority-class predictions that random sampling cannot explain. But it does mean that at this data scale and with this severely imbalanced class distribution, LoRA produces architecture-specific specialization rather than uniform improvement. Qwen LoRA's regression is also specific and diagnosable: rather than collapsing to the majority class as might be expected, it degrades dismiss and remand performance while barely improving sustain. The main takeaway is that accuracy alone, and any single-metric headline in general, can badly mislead when the underlying class distribution is imbalanced and the dataset is small.

I am releasing the curated dataset publicly on HuggingFace, and I hope it is useful to anyone who wants to benchmark their own models on real immigration adjudication decisions.

Future Work.

- Multi-seed evaluation with confidence intervals, especially for minority classes.
- Ablation studies on LoRA rank, training data volume, and task mixture.
- RAG integration with the USCIS Policy Manual to ground predictions in published guidance.
- Human evaluation with immigration attorneys as an external validity check.

References

Kazarian v. USCIS, 596 F.3d 1115 (9th Cir. 2010). Established the two-step framework for EB-1A adjudication.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shuxian Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hugo Touvron, Louis Martin, Kevin Stone, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

U.S. Citizenship and Immigration Services. 2024. USCIS policy manual, volume 6, part F: Extraordinary ability or achievement. https://www.uscis.gov/policy-manual. Accessed: 2025-02-01.