Assistant Fine-Tuning Performance Analysis

This document summarizes the results of fine-tuning experiments for generating formal postconditions for smart contracts using different GPT models. The analysis is based on 100 total runs.

Overall Performance Analysis

This section presents the overall success rates of each model across all tasks. Success is defined as generating postconditions that pass verification.

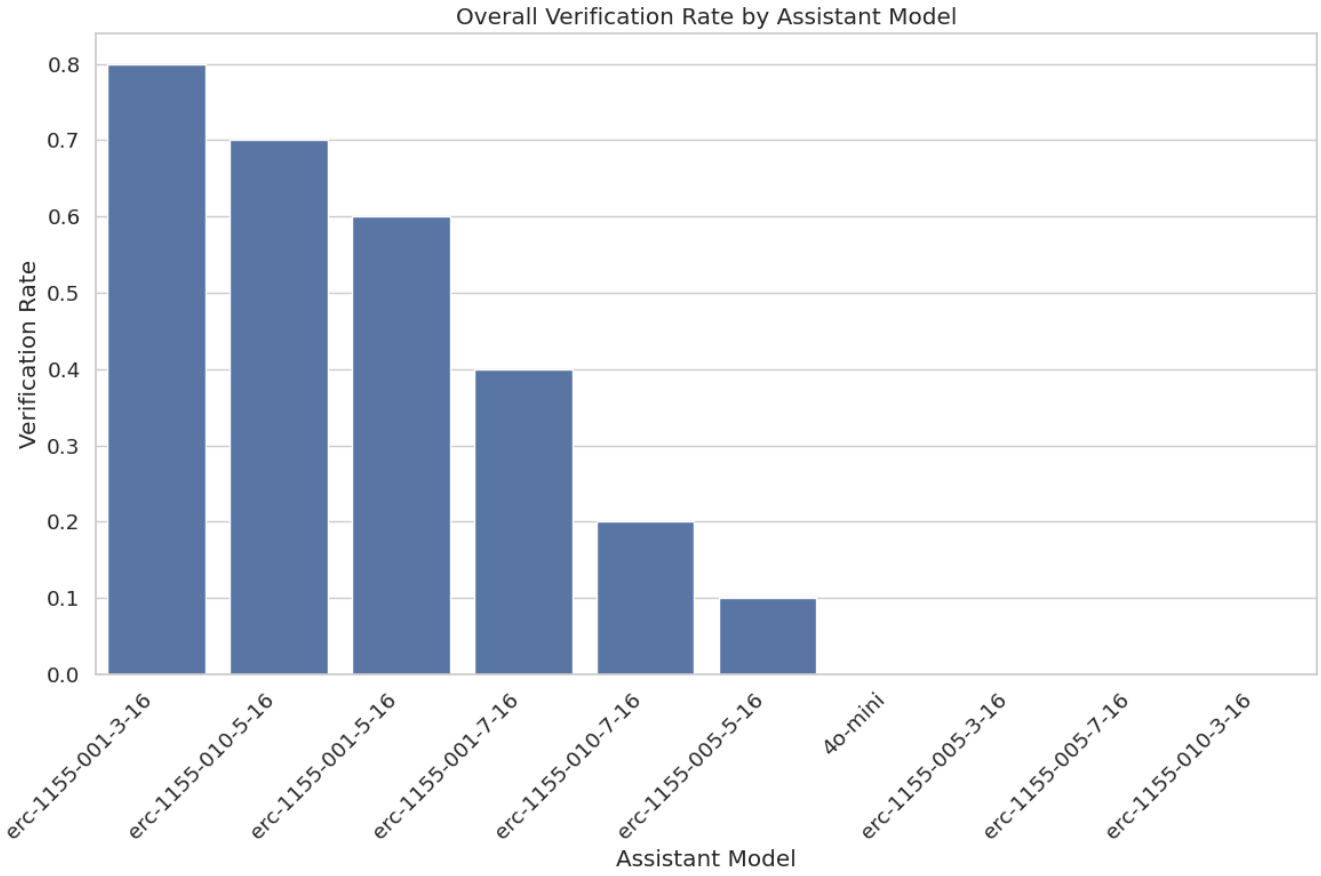
Total Runs Analyzed: 100

Overall Success Rates:

model	verification_rate	verified_count	total_runs
erc-1155-001-3-16	80.00	8	10
erc-1155-010-5-16	70.00	7	10
erc-1155-001-5-16	60.00	6	10
erc-1155-001-7-16	40.00	4	10
erc-1155-010-7-16	20.00	2	10
erc-1155-005-5-16	10.00	1	10
4o-mini	0.00	0	10
erc-1155-005-3-16	0.00	0	10
erc-1155-005-7-16	0.00	0	10
erc-1155-010-3-16	0.00	0	10

Key Observations:

- The 'erc-1155-001-3-16' model achieved the highest overall success rate at 80.00%.
- The average verification rate across all models was 28.00%.
- The 'erc-1155-010-3-16' model had the lowest success rate at 0.00%.



Model Specificity Analysis

This section examines how well each model performs when requested to generate postconditions for a particular contract standard.

Success Rate (%) for each Model on each Requested Type:

model	erc1155
erc-1155-010-7-16	20.00
erc-1155-010-5-16	70.00
erc-1155-010-3-16	0.00
erc-1155-005-7-16	0.00
erc-1155-005-5-16	10.00
erc-1155-005-3-16	0.00
erc-1155-001-7-16	40.00
erc-1155-001-5-16	60.00
erc-1155-001-3-16	80.00
4o-mini	0.00

Successful Runs / Total Runs for each Model on each Requested Type:

model	erc1155
erc-1155-010-7-16	2 / 10

model	erc1155
erc-1155-010-5-16	7 / 10
erc-1155-010-3-16	0 / 10
erc-1155-005-7-16	0 / 10
erc-1155-005-5-16	1 / 10
erc-1155-005-3-16	0 / 10
erc-1155-001-7-16	4 / 10
erc-1155-001-5-16	6 / 10
erc-1155-001-3-16	8 / 10
4o-mini	0 / 10

Efficiency Analysis

This section evaluates the efficiency of the models in terms of the number of iterations and time taken to reach a successful verification or exhaust attempts.

Average Iterations and Time per Model:

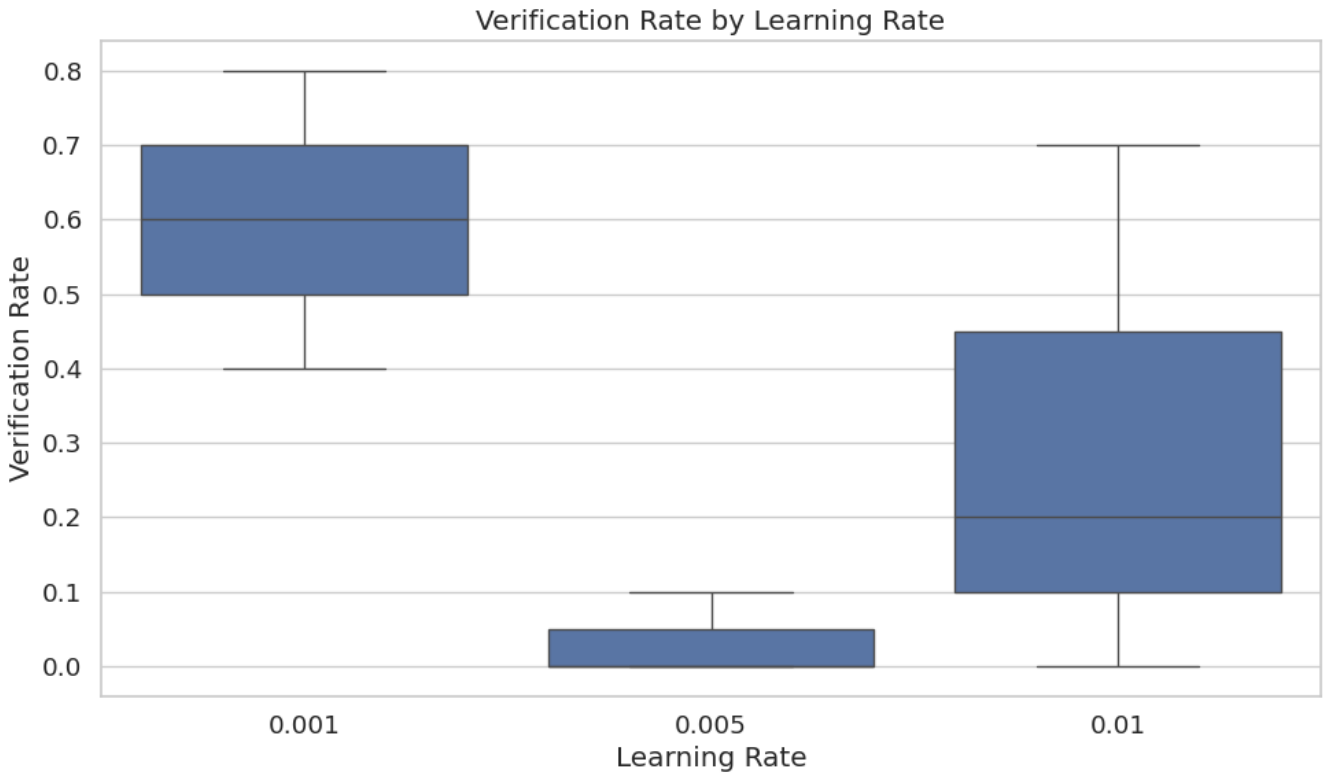
model	avg_fail_iterations	avg_success_iterations	avg_fail_time	avg_success_time	fail_rate
4o-mini	31.9	0.0	842.1562853097915	0.0	100.00
erc-1155-005-3-16	32.6	0.0	673.1885904788971	0.0	100.00
erc-1155-005-7-16	27.6	0.0	738.9068398237229	0.0	100.00
erc-1155-010-3-16	21.3	0.0	504.8798729658127	0.0	100.00
erc-1155-005-5-16	28.0	23.0	760.4511196613312	444.7884430885315	90.00
erc-1155-010-7-16	21.125	14.5	370.1387418806553	276.47833919525146	80.00
erc-1155-001-7-16	21.833333333333332	16.25	417.5228614807129	299.71189588308334	60.00

model	avg_fail_iterations	avg_success_iterations	avg_fail_time	avg_success_time	fail_rate
erc-1155-001-5-16	17.25	7.166666666666667	440.722040951252	179.71604613463083	40.00
erc-1155-010-5-16	17.666666666666668	9.857142857142858	549.7526532014211	331.2002204486302	30.00
erc-1155-001-3-16	16.0	9.875	426.118123292923	258.2129156887531	20.00

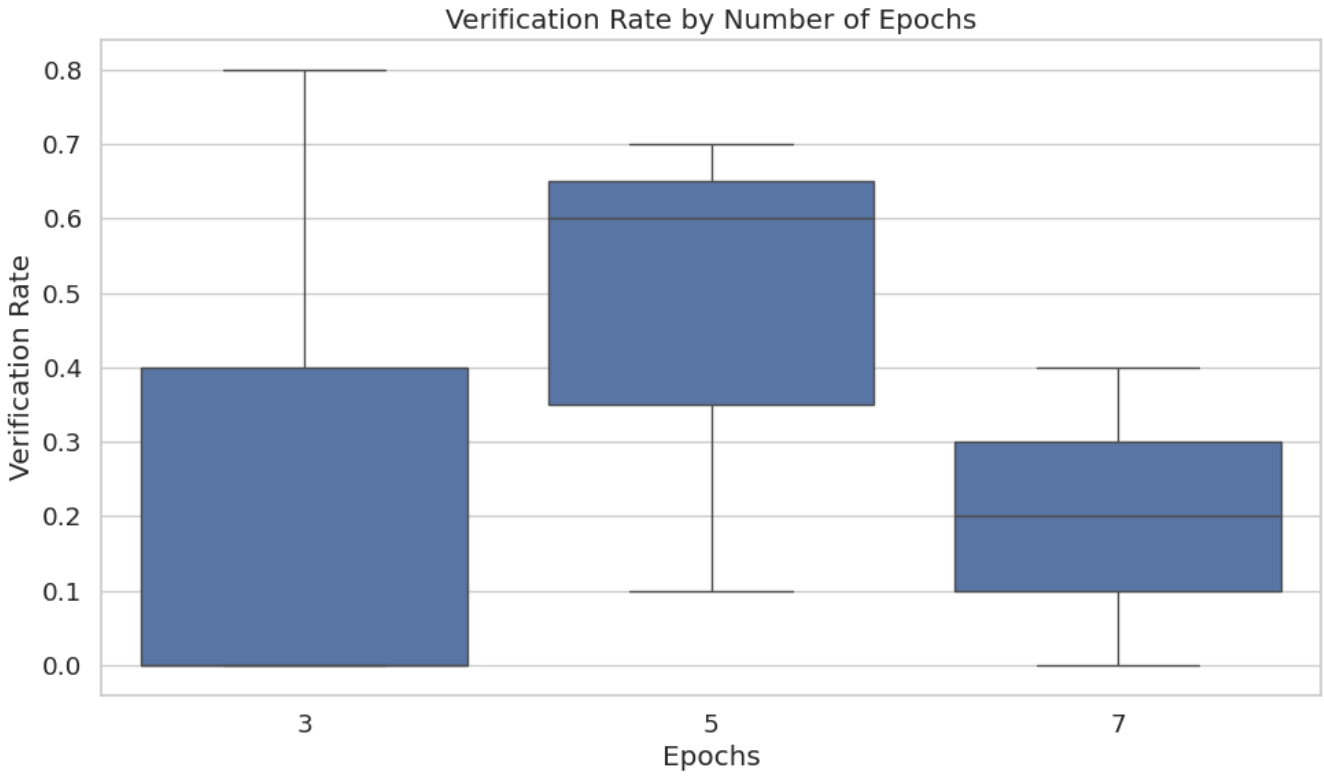
Hyperparameter Analysis

This section analyzes the impact of different hyperparameters (learning rate, epochs, batch size) on model performance.

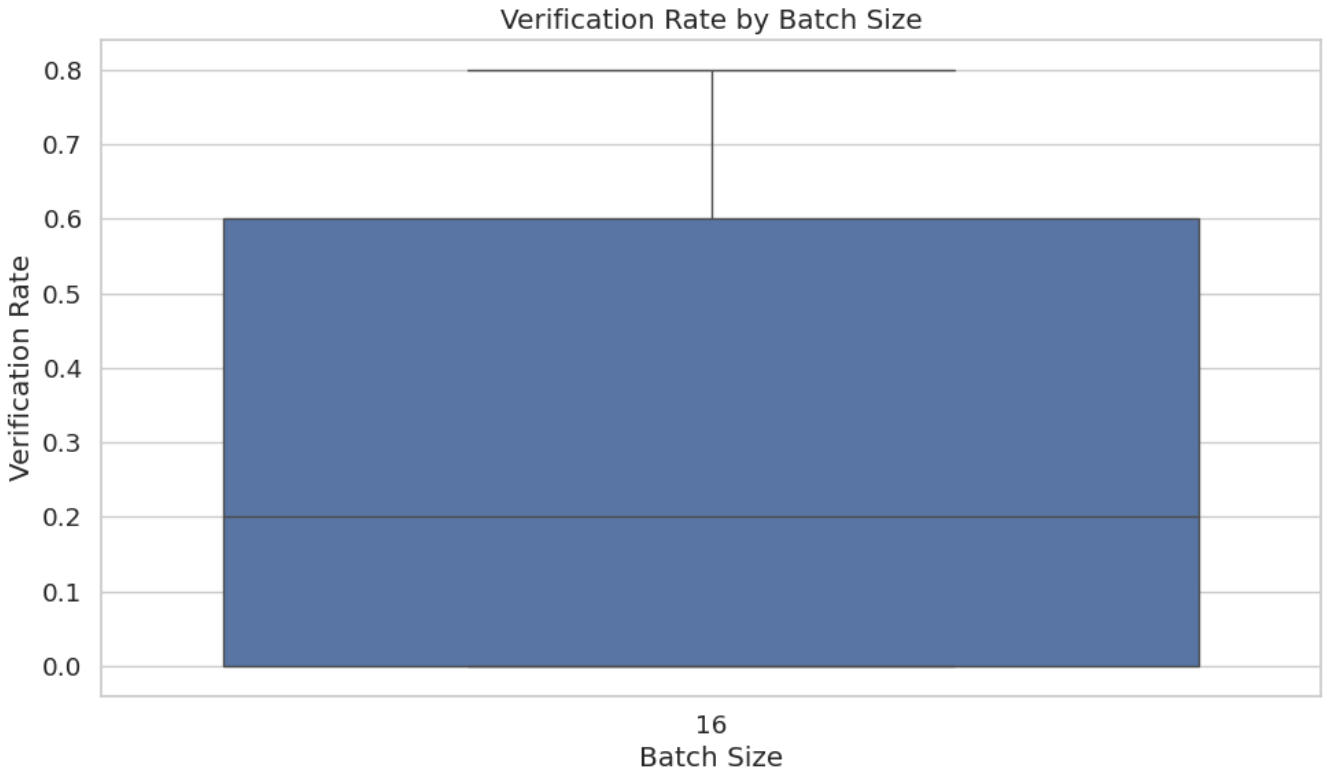
By Learning Rate



By Epochs

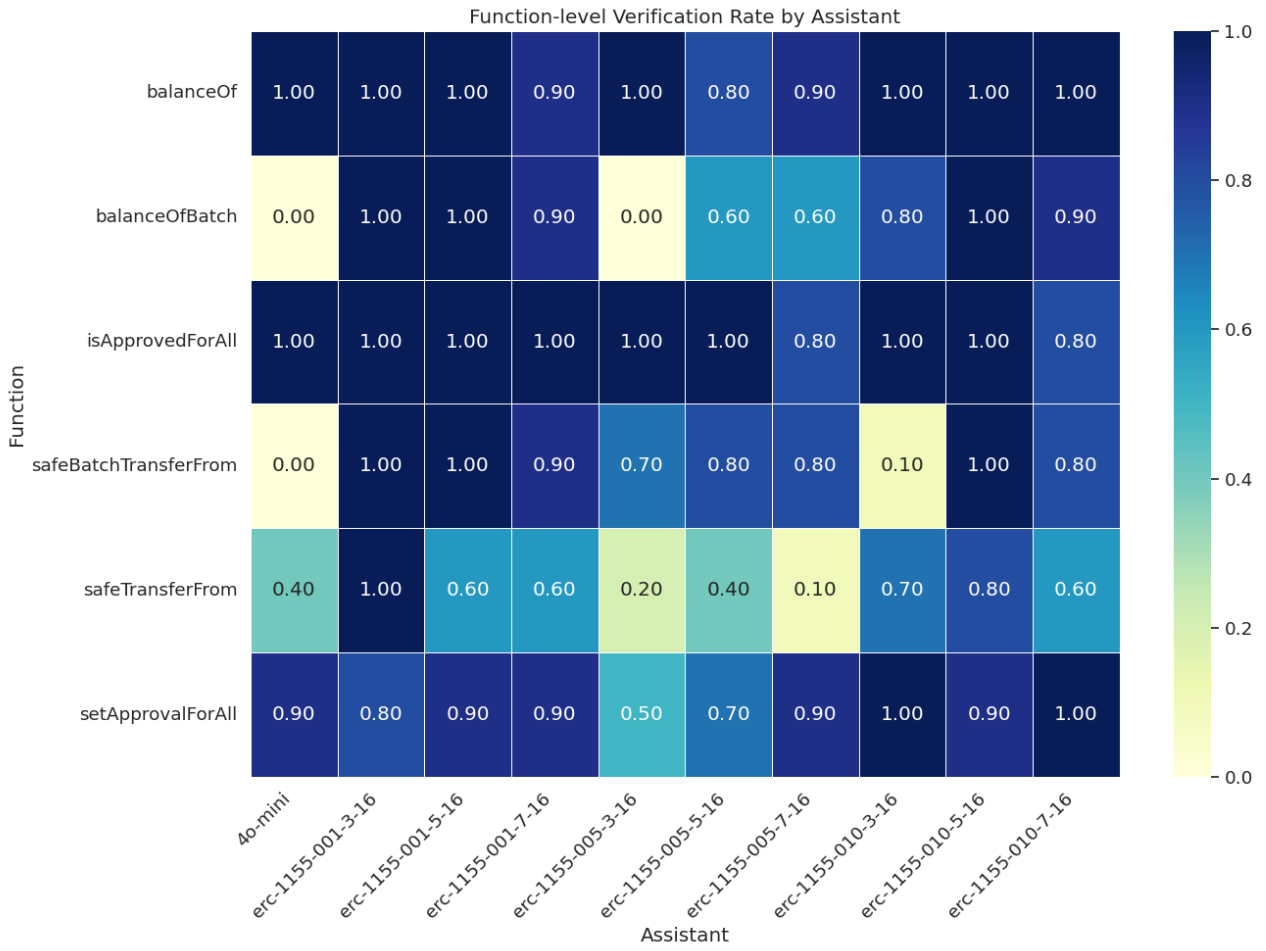


By Batch Size



Function-level Verification Analysis

This section examines which specific functions are most successfully verified by each model.



Overall Conclusion

Based on the analysis, the following conclusions can be drawn:

1. The models `erc-1155-001-3-16`, `erc-1155-010-5-16` and `erc-1155-001-5-16` demonstrated the highest overall verification rates.
2. Fine-tuning generally improved performance compared to the baseline `4o-mini` model (verification rate: 0.00%).
3. The optimal hyperparameters appear to be a learning rate of 0.001, 5 epochs, and a batch size of 16.
4. Successful verification attempts are significantly faster than failed attempts, suggesting that early success indicators can help determine when a model is likely to produce valid postconditions.

Report generated on 2025-05-19 18:05:34