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 80 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.

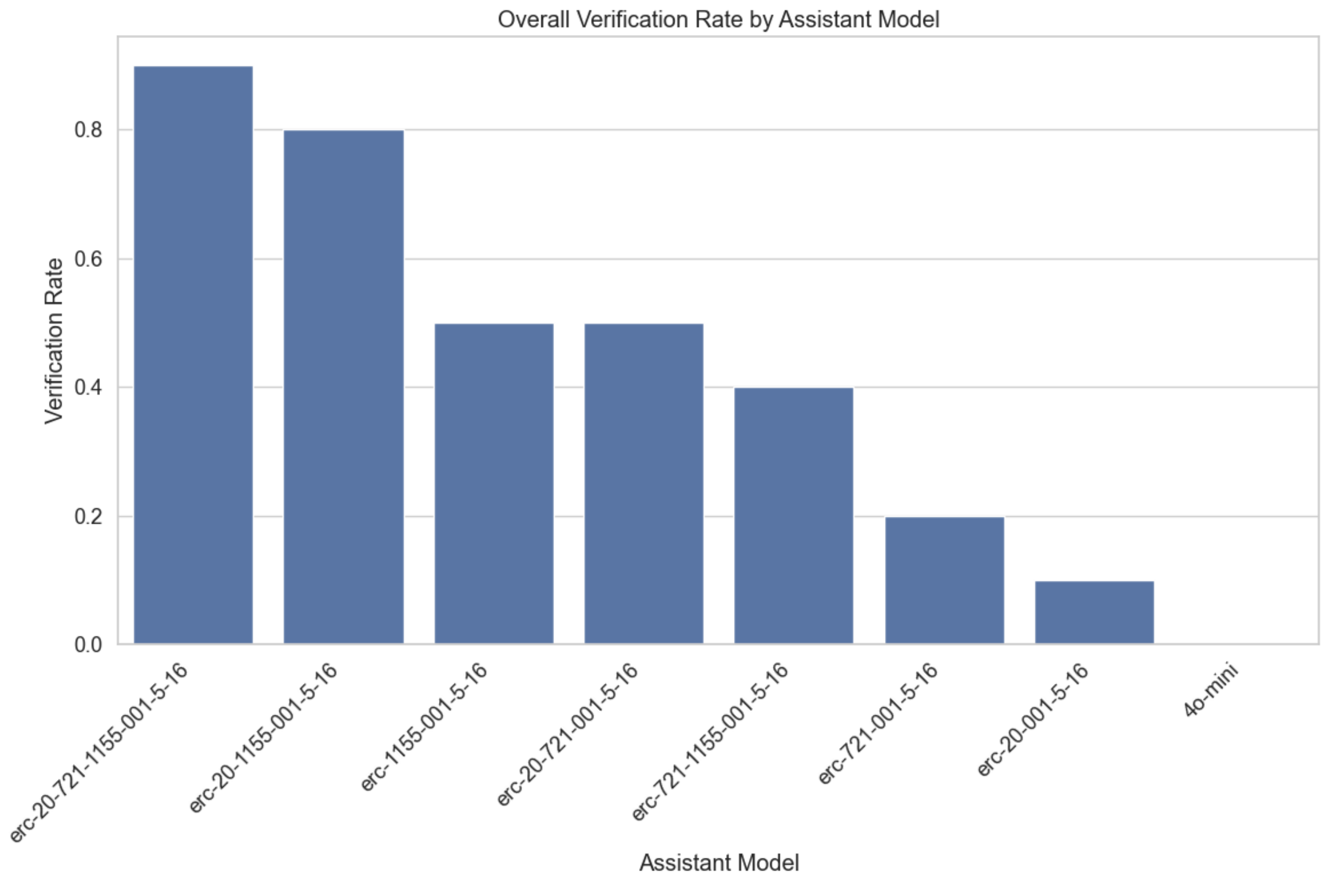
Total Runs Analyzed: 80

Overall Success Rates:

model	verification_rate	verified_count	total_runs
erc-20-721-1155-001-5-16	90.00	9	10
erc-20-1155-001-5-16	80.00	8	10
erc-1155-001-5-16	50.00	5	10
erc-20-721-001-5-16	50.00	5	10
erc-721-1155-001-5-16	40.00	4	10
erc-721-001-5-16	20.00	2	10
erc-20-001-5-16	10.00	1	10
4o-mini	0.00	0	10

Key Observations:

- The 'erc-20-721-1155-001-5-16' model achieved the highest overall success rate at 90.00%.
- The average verification rate across all models was 42.50%.
- The '4o-mini' 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-721-1155-001-5-16	40.00
erc-721-001-5-16	20.00
erc-20-721-1155-001-5-16	90.00
erc-20-721-001-5-16	50.00
erc-20-1155-001-5-16	80.00
erc-20-001-5-16	10.00
erc-1155-001-5-16	50.00
4o-mini	0.00

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

model	erc1155
erc-721-1155-001-5-16	4 / 10
erc-721-001-5-16	2 / 10
erc-20-721-1155-001-5-16	9 / 10
erc-20-721-001-5-16	5 / 10

model	erc1155
erc-20-1155-001-5-16	8 / 10
erc-20-001-5-16	1 / 10
erc-1155-001-5-16	5 / 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:

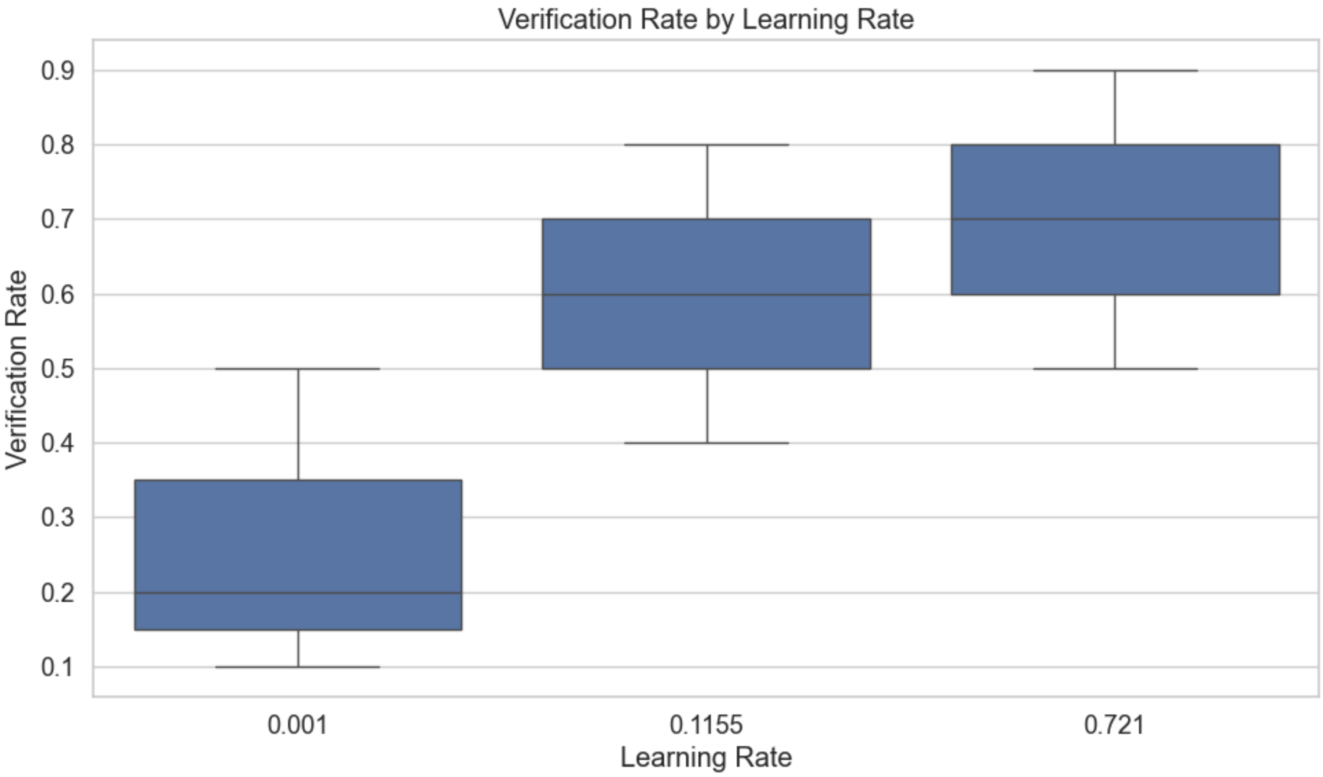
model	avg_fail_iterations	avg_success_iterations	avg_fail_time	avg_success_time	fail_rate
4o-mini	32.1	0.0	911.6306059837341	0.0	100.00
erc-20-001-5-16	21.444444444444443	6.0	437.62032371097143	127.70915937423706	90.00
erc-721-001-5-16	15.0	9.5	329.0432448089123	228.00407946109772	80.00
erc-721-1155-001-5-16	21.833333333333332	16.25	417.5228614807129	299.71189588308334	60.00
erc-1155-001-5-16	17.8	11.8	515.702849817276	249.05493535995484	50.00
erc-20-721-001-5-16	17.8	11.8	515.702849817276	249.05493535995484	50.00
erc-20-1155-001-5-16	18.0	7.875	536.6593418121338	234.57828336954117	20.00

model	avg_fail_iterations	avg_success_iterations	avg_fail_time	avg_success_time	fail_rate
erc-20-721-1155-001-5-16	15.0	8.0	329.86697244644165	176.03217895825705	10.00

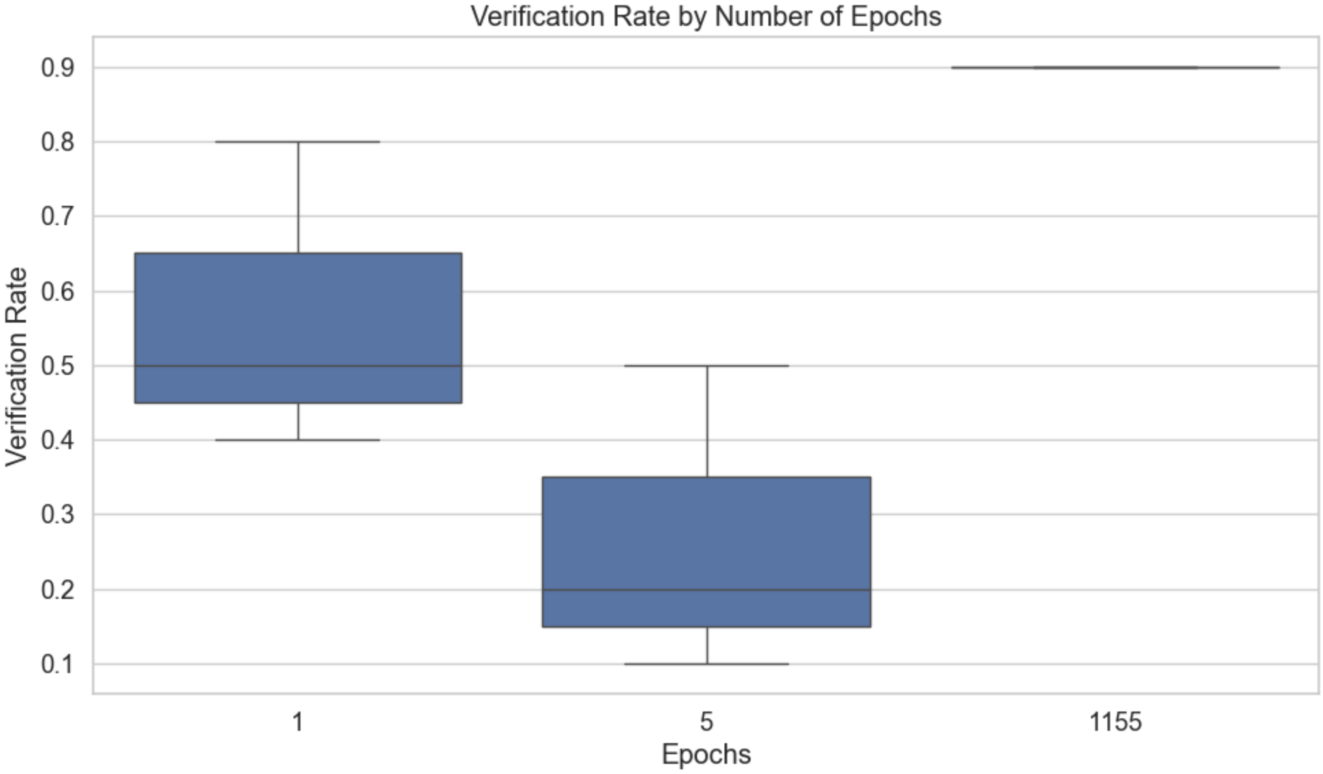
Hyperparameter Analysis

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

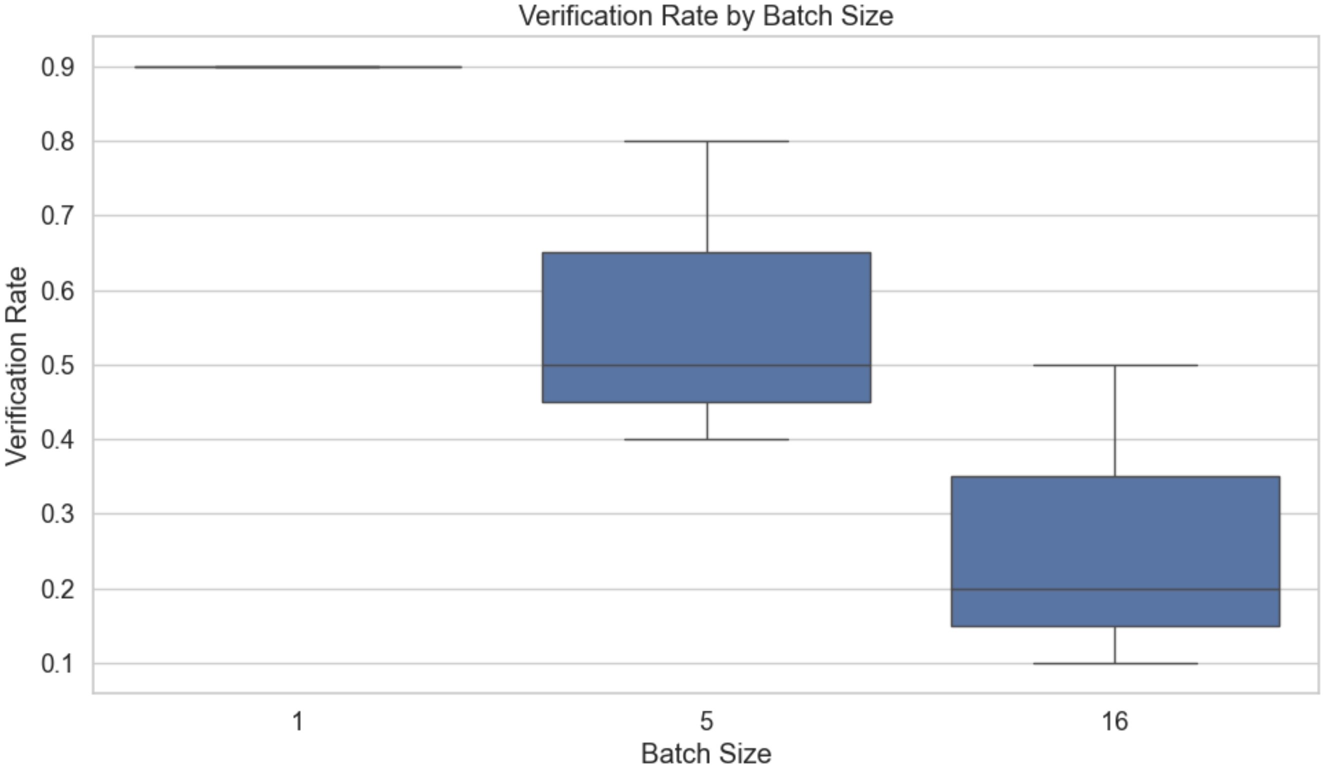
By Learning Rate



By Epochs

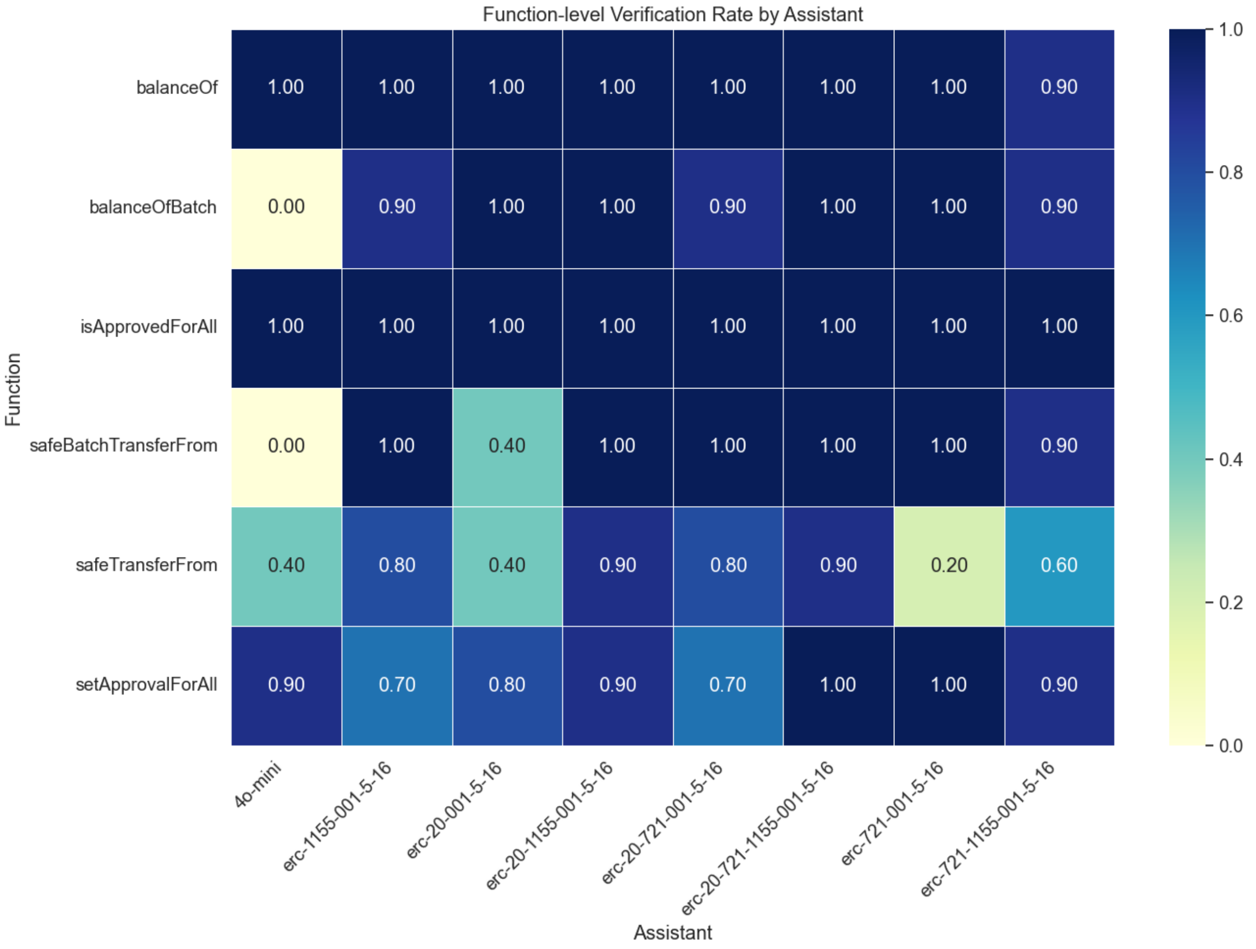


By Batch Size



Function-level Verification Analysis

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



Function	4o-mini	erc-1155-001-5-16	erc-20-001-5-16	erc-20-1155-001-5-16	erc-20-721-001-5-16	erc-20-721-1155-001-5-16	erc-721-001-5-16	erc-721-1155-001-5-16
balanceOf	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90
balanceOfBatch	0.00	0.90	1.00	1.00	0.90	1.00	1.00	0.90
isApprovedForAll	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
safeBatchTransferFrom	0.00	1.00	0.40	1.00	1.00	1.00	1.00	0.90
safeTransferFrom	0.40	0.80	0.40	0.90	0.80	0.90	0.20	0.60
setApprovalForAll	0.90	0.70	0.80	0.90	0.70	1.00	1.00	0.90

Overall Conclusion

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

1. The models `erc-20-721-1155-001-5-16`, `erc-20-1155-001-5-16` 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 0.721, 1155 epochs, and a batch size of 1.
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 2025-05-28 06:22:33