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.

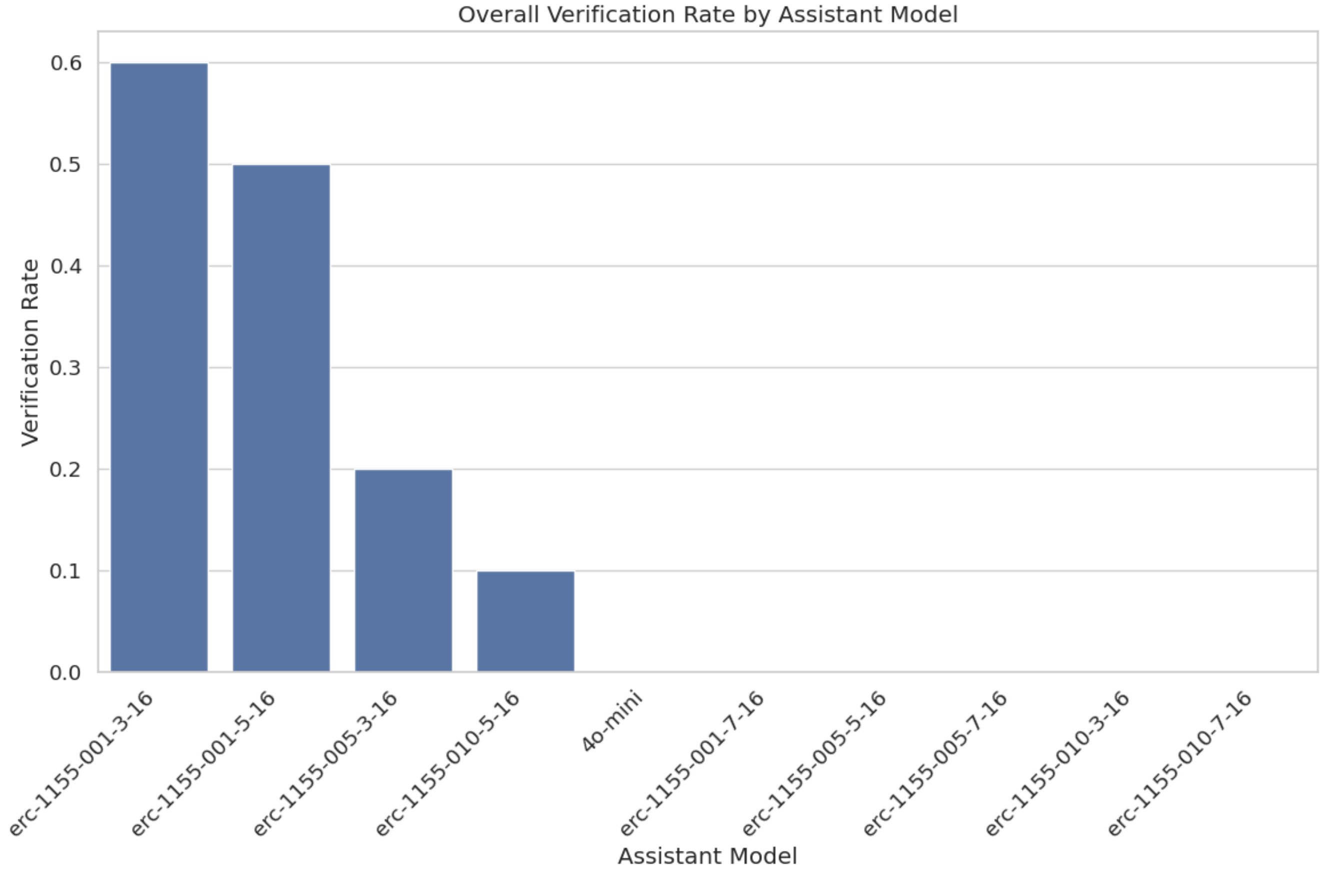
Total Runs Analyzed: 100

Overall Success Rates:

model	verification_rate	verified_count	total_runs
erc-1155-001-3-16	60.00	6	10
erc-1155-001-5-16	50.00	5	10
erc-1155-005-3-16	20.00	2	10
erc-1155-010-5-16	10.00	1	10
4o-mini	0.00	0	10
erc-1155-001-7-16	0.00	0	10
erc-1155-005-5-16	0.00	0	10
erc-1155-005-7-16	0.00	0	10
erc-1155-010-3-16	0.00	0	10
erc-1155-010-7-16	0.00	0	10

Key Observations:

- The 'erc-1155-001-3-16' model achieved the highest overall success rate at 60.00%.
- The average verification rate across all models was 14.00%.
- The 'erc-1155-010-7-16' 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	0.00
erc-1155-010-5-16	10.00
erc-1155-010-3-16	0.00
erc-1155-005-7-16	0.00
erc-1155-005-5-16	0.00
erc-1155-005-3-16	20.00
erc-1155-001-7-16	0.00
erc-1155-001-5-16	50.00
erc-1155-001-3-16	60.00
4o-mini	0.00

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

model	erc1155
erc-1155-010-7-16	0 / 10

model	erc1155
erc-1155-010-5-16	1 / 10
erc-1155-010-3-16	0 / 10
erc-1155-005-7-16	0 / 10
erc-1155-005-5-16	0 / 10
erc-1155-005-3-16	2 / 10
erc-1155-001-7-16	0 / 10
erc-1155-001-5-16	5 / 10
erc-1155-001-3-16	6 / 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:

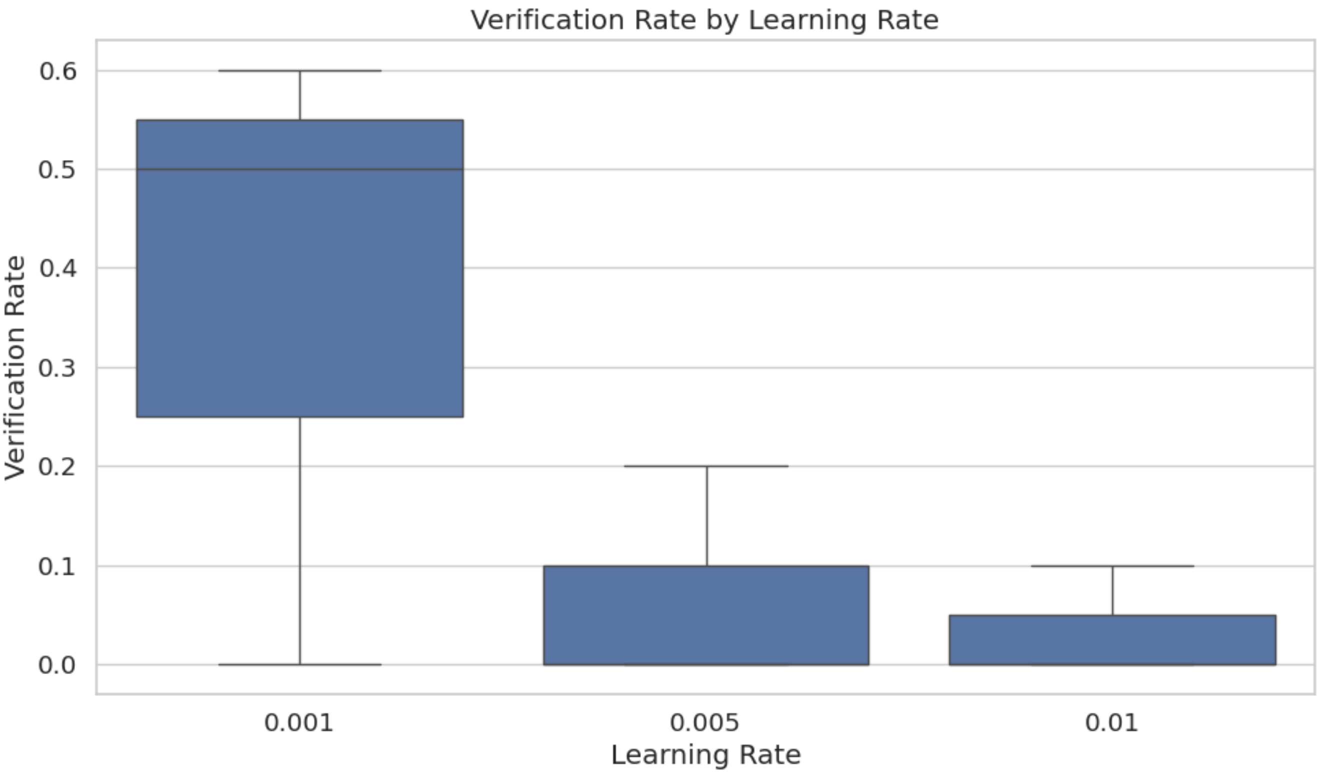
model	avg_fail_iterations	avg_success_iterations	avg_fail_time	avg_success_time	fail_rate
4o-mini	10.0	0.0	319.9581855297089	0.0	100.00
erc-1155-001-7-16	9.3	0.0	1720.0253734350204	0.0	100.00
erc-1155-005-5-16	10.0	0.0	245.0676204919815	0.0	100.00
erc-1155-005-7-16	9.2	0.0	218.4296261548996	0.0	100.00
erc-1155-010-3-16	10.0	0.0	233.13596296310425	0.0	100.00
erc-1155-010-7-16	9.8	0.0	212.90439779758452	0.0	100.00
erc-1155-010-5-16	9.222222222222221	3.0	201.27651551034717	74.02299165725708	90.00

model	avg_fail_iterations	avg_success_iterations	avg_fail_time	avg_success_time	fail_rate
erc-1155-005-3-16	10.0	3.5	3755.226481884718	567.6829489469528	80.00
erc-1155-001-5-16	8.2	4.2	2555.9557731151585	3761.115694856643	50.00
erc-1155-001-3-16	10.0	2.5	281.21035850048065	101.30749650796254	40.00

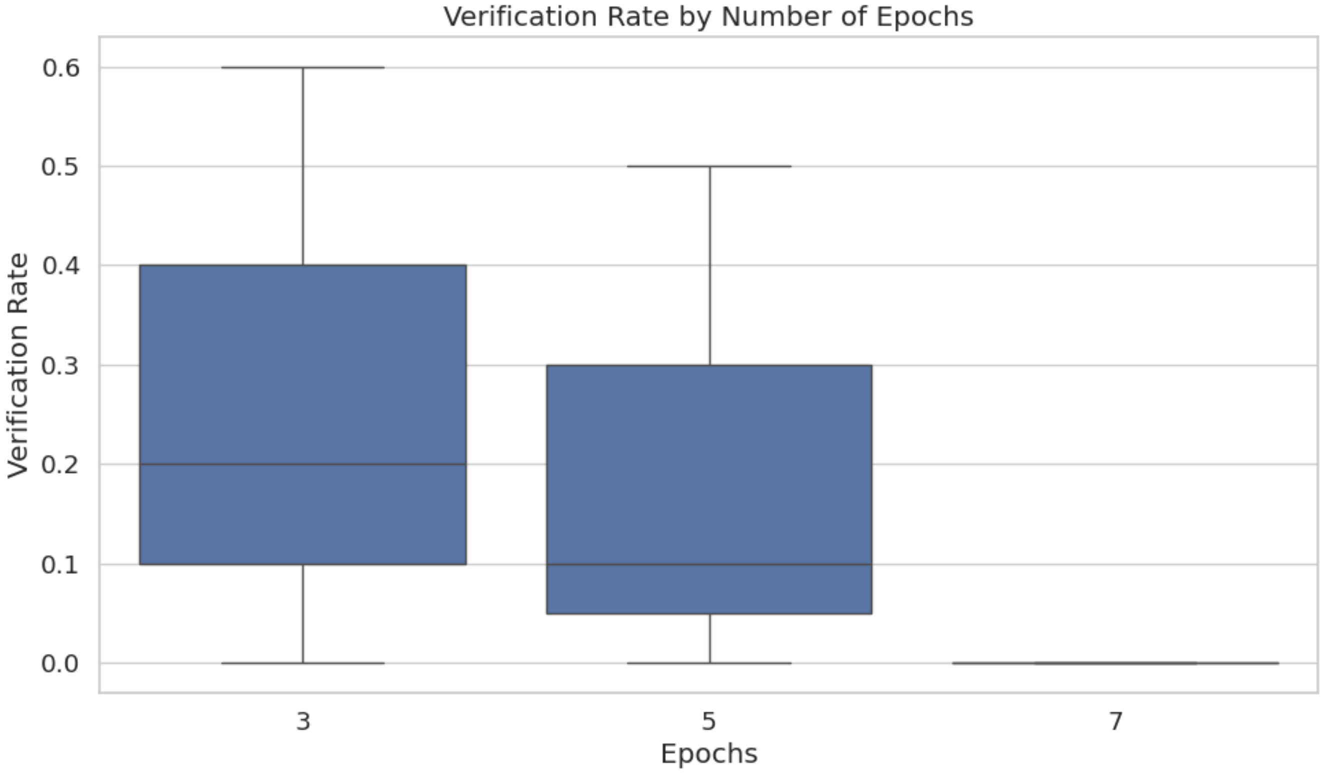
Hyperparameter Analysis

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

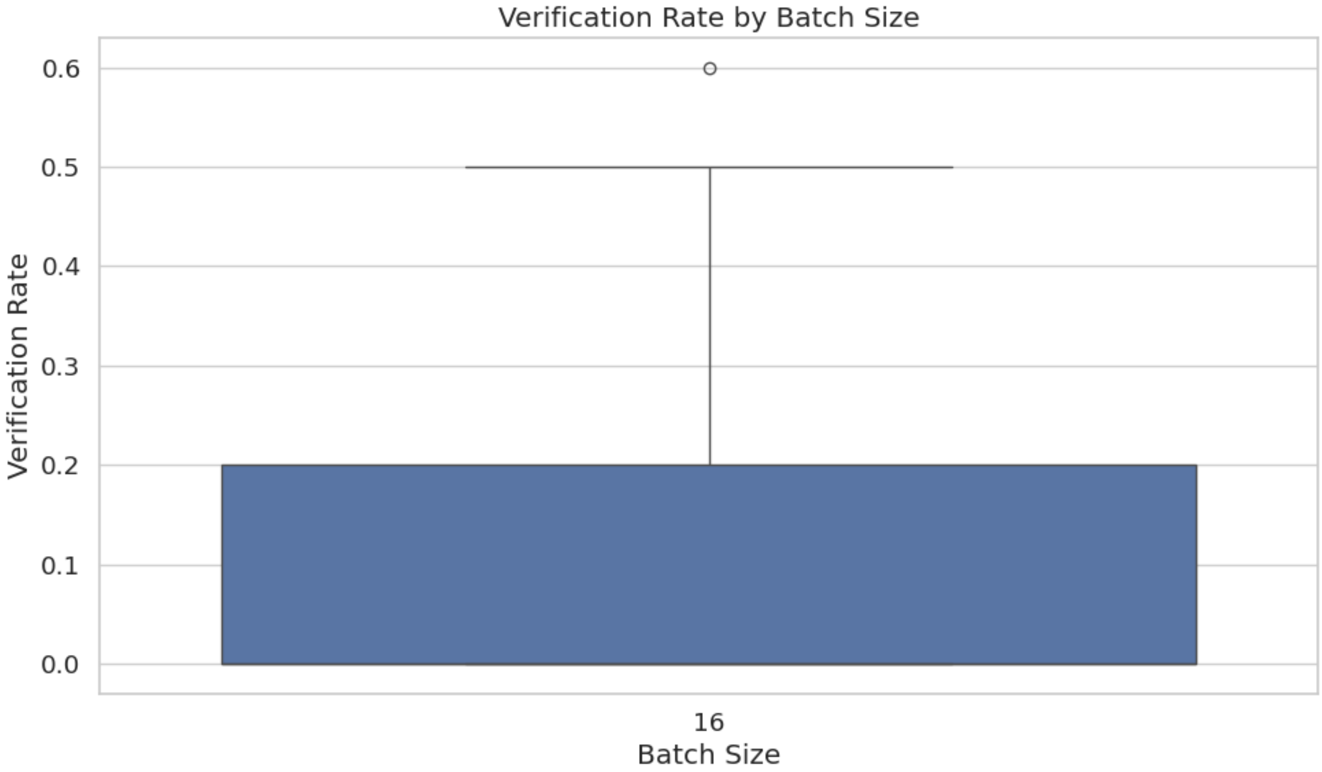
By Learning Rate



By Epochs



By Batch Size



Function-level Verification Analysis

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

Function Verification Rates

Overall Conclusion

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

1. The models `erc-1155-001-3-16`, `erc-1155-001-5-16` and `erc-1155-005-3-16` 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, 3 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 2025-05-21 10:26:37