Test Model on Multiple Datasets

This notebook evaluates a trained model on three test datasets:

- **Natural**: Uniformly random samples from [1, 10^13]
- **Cheat**: Numbers with prime factors only within the first 100 primes
- **Non-cheat**: Numbers with at least one prime factor outside the first 100 primes

It reports per-class performance for each dataset.

Configuration

Specify the model checkpoint and encoding to test:

```
Testing model: ../models/model_interCRT100_with_n_natural/mu/1/checkpoint.pth
Encoding: interCRT100_with_n
Task: mu
Results will be saved to: ../test_results/interCRT100_with_n_mu
```

Helper Functions

```
Helper functions loaded!
```

Check if Model Checkpoint Exists

```
✓ Checkpoint found: ../models/model_interCRT100_with_n_natural/mu/1/checkpoint.pth
```

Check Test Data Files

```
Checking test data files:

✓ natural      : ../input/input_dir_interCRT100_with_n_natural/mu_interCRT100_with_n_n
atural.txt.test (107.67 MB)
✓ cheat        : ../input/input_dir_interCRT100_with_n_cheat/mu_interCRT100_with_n_che
at.txt.test (107.01 MB)
✓ non_cheat    : ../input/input_dir_interCRT100_with_n_non_cheat/mu_interCRT100_with_n
_non_cheat.txt.test (107.67 MB)
```

Run Evaluation on Each Test Dataset

```
Evaluation function ready!
```

Evaluate on Natural Dataset

```
================================================================================
EVALUATING ON NATURAL DATASET
================================================================================

Running command:
python ../Int2Int/train.py --eval_only True --reload_model /mnt/c/Users/ziwen/clair/m
obius_case_study/notebooks/../models/model_interCRT100_with_n_natural/mu/1/checkpoin
t.pth --eval_data /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../input/input
_dir_interCRT100_with_n_natural/mu_interCRT100_with_n_natural.txt.test --eval_size 10
000 --data_types int[201]:range(-1,2) --operation data --cpu True --num_workers 0 --d
ump_path /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../test_results/interCR
T100_with_n_mu/eval_dump

================================================================================
Full output saved to: ../test_results/interCRT100_with_n_mu/eval_natural.log

✓ Evaluation completed successfully!

Overall Accuracy: 59.24%
```

Evaluate on Cheat Dataset

```
================================================================================
EVALUATING ON CHEAT DATASET
================================================================================

Running command:
python ../Int2Int/train.py --eval_only True --reload_model /mnt/c/Users/ziwen/clair/m
obius_case_study/notebooks/../models/model_interCRT100_with_n_natural/mu/1/checkpoin
t.pth --eval_data /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../input/input
_dir_interCRT100_with_n_cheat/mu_interCRT100_with_n_cheat.txt.test --eval_size 10000
--data_types int[201]:range(-1,2) --operation data --cpu True --num_workers 0 --dump_
path /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../test_results/interCRT100
_with_n_mu/eval_dump

================================================================================
Full output saved to: ../test_results/interCRT100_with_n_mu/eval_cheat.log

✓ Evaluation completed successfully!

Overall Accuracy: 36.30%
```

Evaluate on Non-Cheat Dataset

```
================================================================================
EVALUATING ON NON-CHEAT DATASET
================================================================================

Running command:
python ../Int2Int/train.py --eval_only True --reload_model /mnt/c/Users/ziwen/clair/m
obius_case_study/notebooks/../models/model_interCRT100_with_n_natural/mu/1/checkpoin
t.pth --eval_data /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../input/input
_dir_interCRT100_with_n_non_cheat/mu_interCRT100_with_n_non_cheat.txt.test --eval_siz
e 10000 --data_types int[201]:range(-1,2) --operation data --cpu True --num_workers 0
--dump_path /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../test_results/inte
rCRT100_with_n_mu/eval_dump

================================================================================
Full output saved to: ../test_results/interCRT100_with_n_mu/eval_non_cheat.log
```

✓ Evaluation completed successfully!

Overall Accuracy: 59.77%

Compile Results

✓ Successfully evaluated on 3 dataset(s)

Results saved to: ../test_results/interCRT100_with_n_mu/all_results.json

Overall Performance Comparison

```
================================================================================
OVERALL PERFORMANCE SUMMARY
================================================================================
```
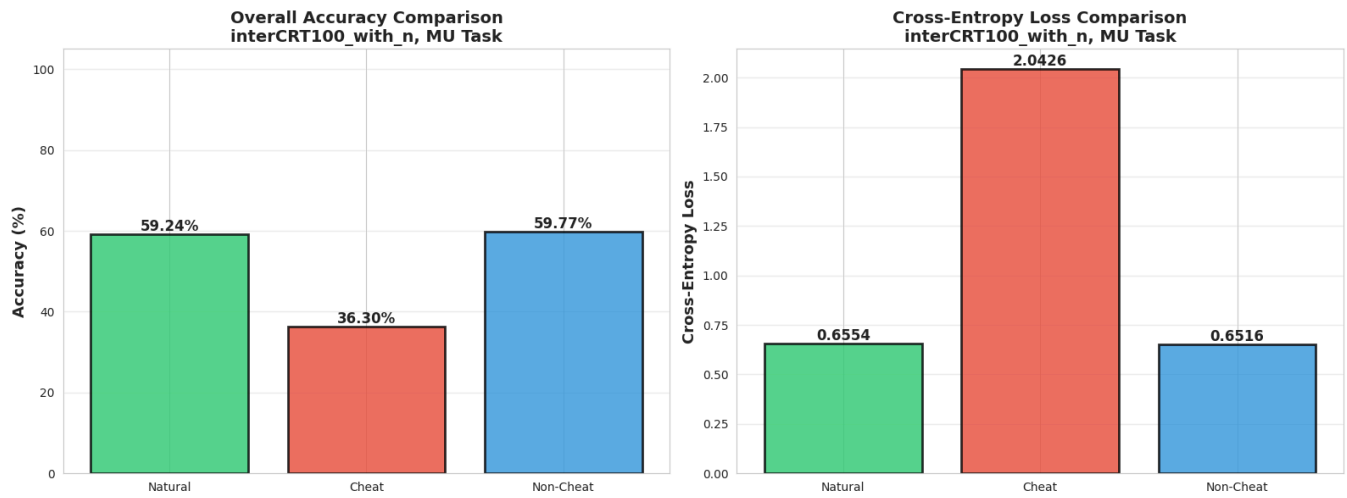
	Dataset	Accuracy (%)	Acc μ=0 (%)	Acc μ=1 (%)	Acc μ=-1 (%)	Perfect (%)	Correct (%)	XE Loss
0	Natural	59.24	86.842105	43.963783	39.110825	59.24	59.24	0.655388
1	Cheat	36.30	38.890907	48.974668	1.315789	36.30	36.30	2.042608
2	Non-Cheat	59.77	86.816802	44.513798	39.647137	59.77	59.77	0.651553

Summary saved to: ../test_results/interCRT100_with_n_mu/overall_summary.csv

Visualization: Overall Performance Comparison

Figure saved to: ../test_results/interCRT100_with_n_mu/overall_comparison.png

Overall Accuracy Comparison
interCRT100_with_n, MU Task

Cross-Entropy Loss Comparison
interCRT100_with_n, MU Task

Visualization: Per-Class Performance Comparison

Figure saved to: ../test_results/interCRT100_with_n_mu/per_class_comparison.png

Per-Class Accuracy Comparison Across Datasets
interCRT100_with_n, MU Task