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_natural/mu/1/checkpoint.pth
Encoding: interCRT100
Task: mu
Results will be saved to: ../test_results/interCRT100_mu
```

Helper Functions

```
Helper functions loaded!
```

Check if Model Checkpoint Exists

```
✓ Checkpoint found: ../models/model_interCRT100_natural/mu/1/checkpoint.pth
```

Check Test Data Files

```
Checking test data files:

✓ natural     : ../input/input_dir_interCRT100_natural/mu_interCRT100_natural.txt.test (105.83 MB)
✓ cheat       : ../input/input_dir_interCRT100_cheat/mu_interCRT100_cheat.txt.test (105.59 MB)
✓ non_cheat   : ../input/input_dir_interCRT100_non_cheat/mu_interCRT100_non_cheat.txt.test (105.83 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_natural/mu/1/checkpoint.pth --
eval_data /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../input/input_dir_int
erCRT100_natural/mu_interCRT100_natural.txt.test --eval_size 10000 --data_types int[2
00]:range(-1,2) --operation data --cpu True --num_workers 0 --dump_path /mnt/c/Users/
ziwen/clair/mobius_case_study/notebooks/../test_results/interCRT100_mu/eval_dump

================================================================================
Full output saved to: ../test_results/interCRT100_mu/eval_natural.log

✓ Evaluation completed successfully!

Overall Accuracy: 50.45%
```

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_natural/mu/1/checkpoint.pth --
eval_data /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../input/input_dir_int
erCRT100_cheat/mu_interCRT100_cheat.txt.test --eval_size 10000 --data_types int[200]:
range(-1,2) --operation data --cpu True --num_workers 0 --dump_path /mnt/c/Users/ziwe
n/clair/mobius_case_study/notebooks/../test_results/interCRT100_mu/eval_dump

================================================================================
Full output saved to: ../test_results/interCRT100_mu/eval_cheat.log

✓ Evaluation completed successfully!

Overall Accuracy: 41.00%
```

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_natural/mu/1/checkpoint.pth --
eval_data /mnt/c/Users/ziwen/clair/mobius_case_study/notebooks/../input/input_dir_int
erCRT100_non_cheat/mu_interCRT100_non_cheat.txt.test --eval_size 10000 --data_types i
nt[200]:range(-1,2) --operation data --cpu True --num_workers 0 --dump_path /mnt/c/Us
ers/ziwen/clair/mobius_case_study/notebooks/../test_results/interCRT100_mu/eval_dump

================================================================================
Full output saved to: ../test_results/interCRT100_mu/eval_non_cheat.log

✓ Evaluation completed successfully!

Overall Accuracy: 51.62%
```

Compile Results

```
✓ Successfully evaluated on 3 dataset(s)

Results saved to: ../test_results/interCRT100_mu/all_results.json
```

Overall Performance Comparison

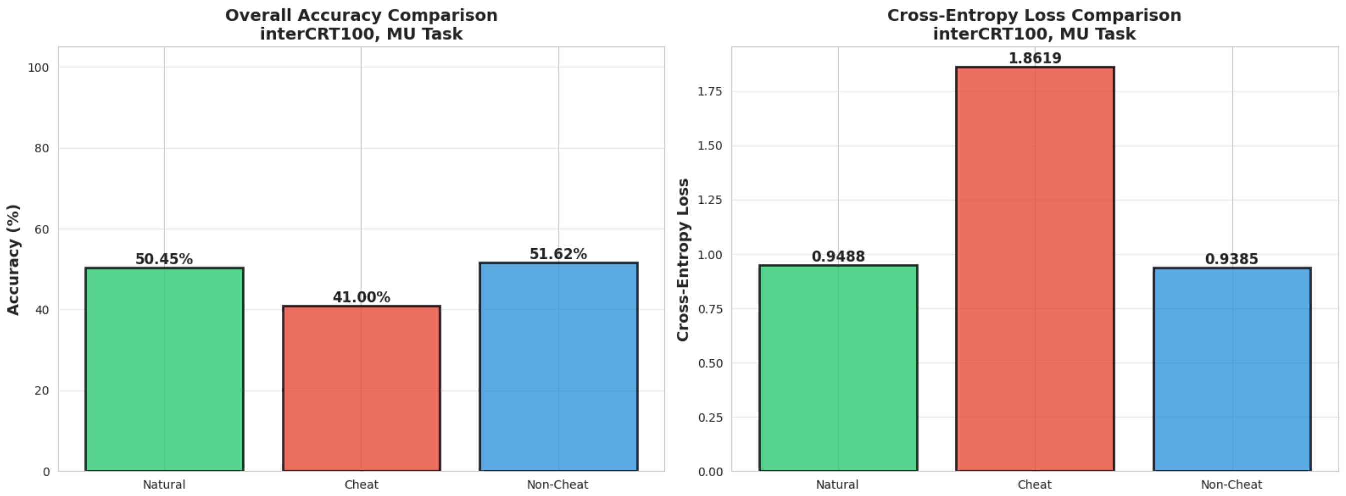
```
================================================================================
OVERALL PERFORMANCE SUMMARY
================================================================================
```

	Dataset	Accuracy (%)	Acc μ=0 (%)	Acc μ=1 (%)	Acc μ=-1 (%)	Perfect (%)	Correct (%)	XE Loss
0	Natural	50.45	86.659878	25.804334	28.255122	50.45	50.45	0.948810
1	Cheat	41.00	44.881390	18.918919	32.900433	41.00	41.00	1.861867
2	Non-Cheat	51.62	88.315977	26.158940	29.411765	51.62	51.62	0.938494

```
Summary saved to: ../test_results/interCRT100_mu/overall_summary.csv
```

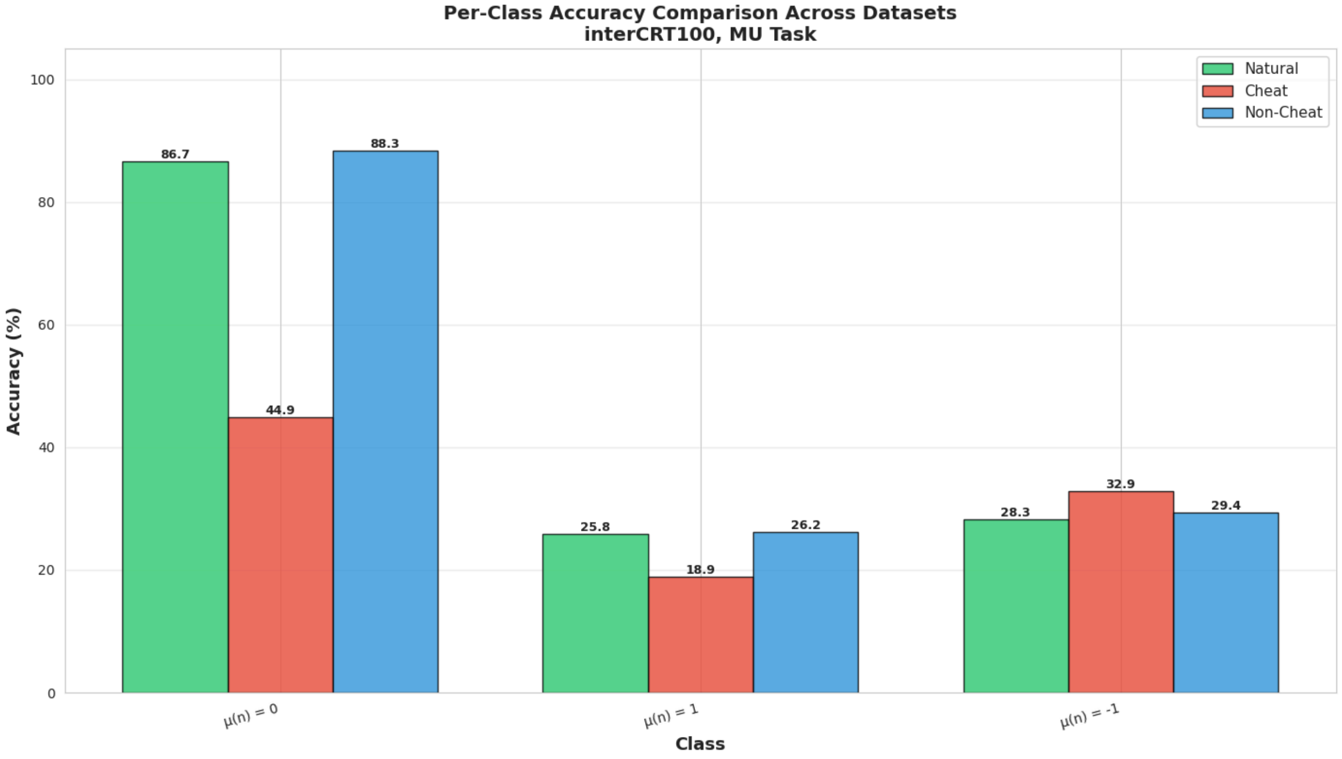
Visualization: Overall Performance Comparison

```
Figure saved to: ../test_results/interCRT100_mu/overall_comparison.png
```



Visualization: Per-Class Performance Comparison

```
Figure saved to: ../test_results/interCRT100_mu/per_class_comparison.png
```


Per-Class Accuracy Comparison Across Datasets
interCRT100, MU Task
Accuracy (%)
🟩 Natural
🟥 Cheat
🟦 Non-Cheat
μ(n) = 0
μ(n) = 1
μ(n) = -1
Class