LLM-Supported Natural Language to Bash Translation

Finnian Westenfelder[1,2] , Erik Hemberg[1] , Miguel Tulla[1] , Stephen Moskal[1] , Una-May O'Reilly[1] , Silviu Chiricescu[3]

[1]ALFA Group MIT-CSAIL, [2]Draper Scholar, [3]Charles Stark Draper Laboratory

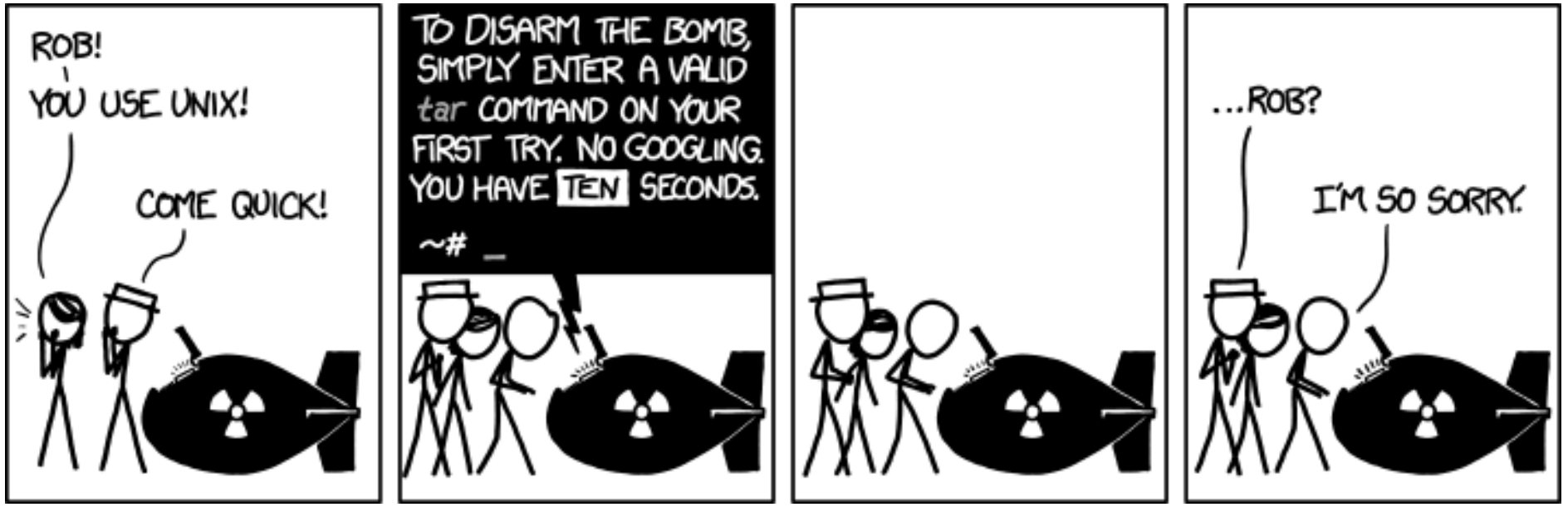
MIT CSAIL

DRAPER

ALFA
ANYSCALE LEARNING FOR ALL

Motivation

- Bash is the default Linux command line interface
- Bash commands are complex and difficult to memorize
- Incorrect commands can cause system failures



https://xkcd.com/1168/

Motivation

- The translation capabilities of Large Language Models (LLMs) can simplify command line interfaces
- Current LLMs are unreliable for translating natural language to Bash commands (NL2SH)
- Improving NL2SH requires
  - (1) clean data
  - (2) reliable benchmarks
  - (3) accurate translation methods

Input: Natural Language
List files in the /workspace directory that were accessed over an hour ago.

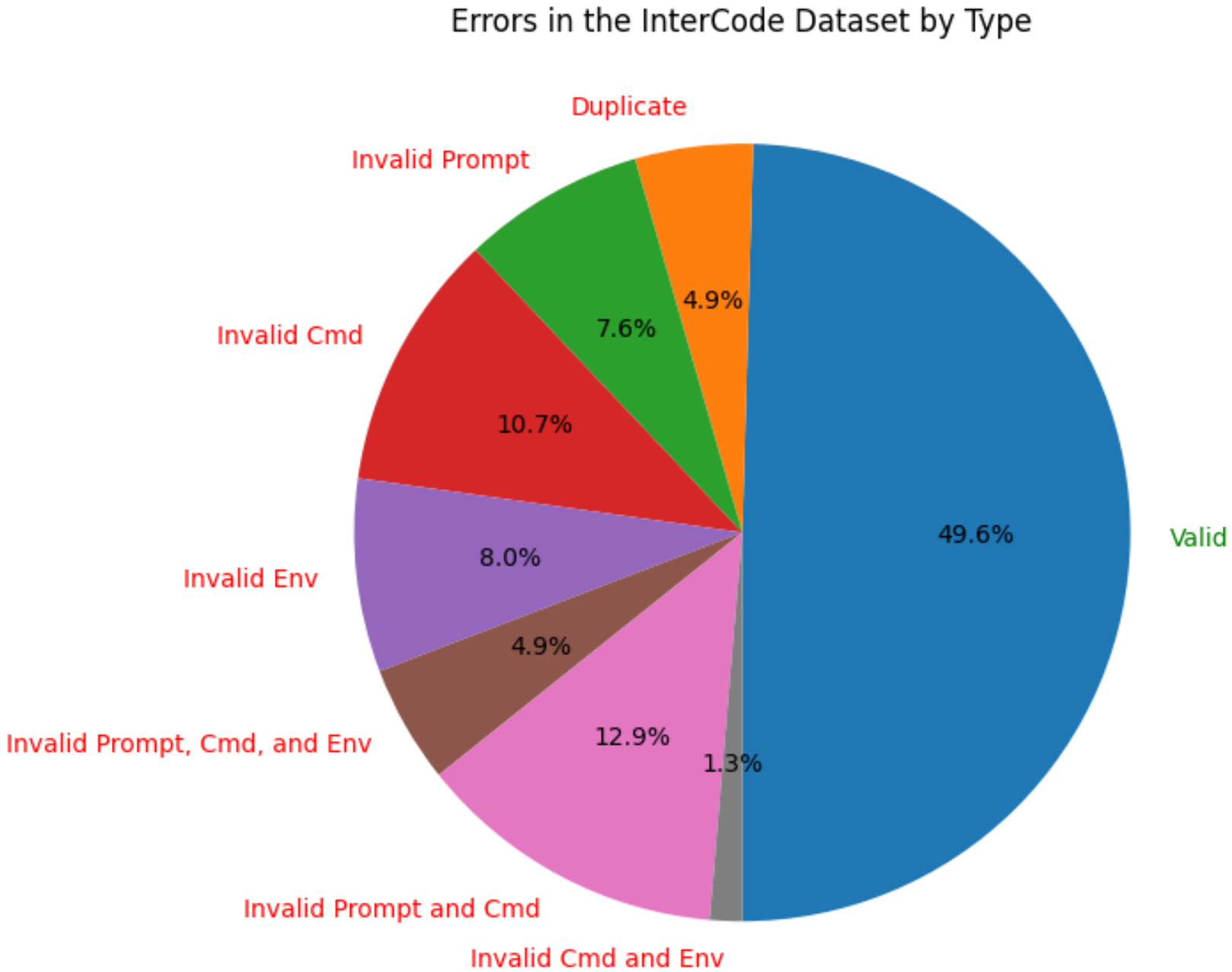
Output: Bash Command
```
find /workspace -type f -amin +60
```

Dataset Overview

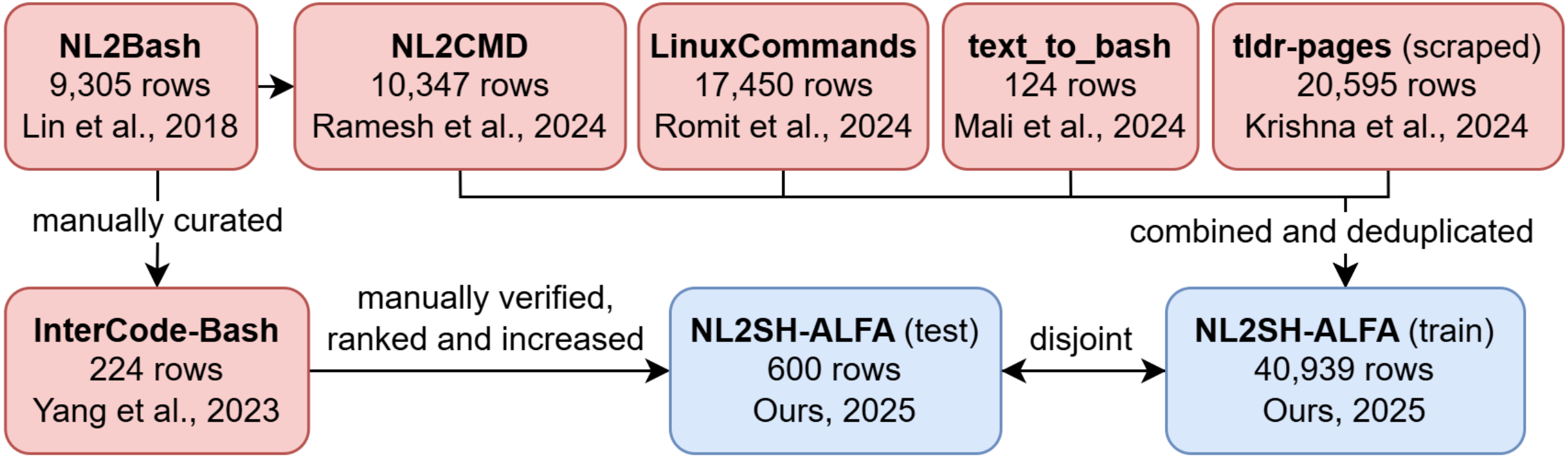
- **Goal:**

Given		Determine
Natural Language	**Bash**	**Correct**
print the system disk usage	`df -h`	True
remove a directory named foo	`rm foo`	False
print the current user's id	`id -u`	True

- **Challenge:** Datasets contain invalid translations.
- **Research Question:** How can we validate NL2SH datasets to ensure models are evaluated using accurate assessments?


Errors in the InterCode Dataset by Type
Duplicate
Invalid Prompt
Invalid Cmd
Invalid Env
Invalid Prompt, Cmd, and Env
Invalid Prompt and Cmd
Invalid Cmd and Env
Valid
4.9%
7.6%
10.7%
8.0%
4.9%
12.9%
1.3%
49.6%

Relationships between NL2SH datasets.



Dataset Findings

- **Research Question:** How can we validate NL2SH datasets to ensure models are evaluated using accurate assessments?
- **Result:** Manual verification of test data improves benchmark reliability. Cleaning and increasing training data improves in-weight learning.
- **Contribution:** We clean and increase the size of NL2SH test and training datasets by over 4x and 2x, respectively.

Benchmark Overview

- **Goal:**

Given			Determine
Natural Language	**GT Command**	**Model Output**	**Equivalent**
print the system boot time	`who -b`	`uptime -s`	True
delete bin\ in the current dir	`rm -r ./bin`	`rm -r /bin`	False
list all groups on the system	`getent group`	`cat /etc/group`	True

- **Challenge:** Benchmarks cannot determine the functional equivalence of Bash commands.
- **Research Question:** How can we design a functional equivalence heuristic that accurately measures the quality of model translations?

NL2SH Translation

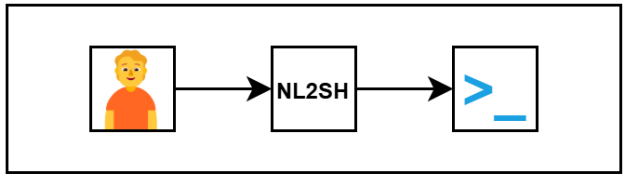
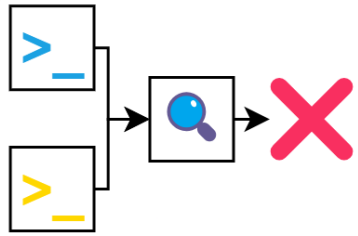


Diagram Key

- **Natural Language Prompt:** Display the current time
- **NL2SH Model:** Fine-tuned LLaMa3.1-8b-Instruct
- **NL2SH Model Command:** timedatectl | grep "Local time"
- **NL2SH Model Command Output:** 2024-06-27 14:18:48
- **Ground Truth Command:** date +"%H:%M:%S"
- **Ground Truth Command Output:** 14:18:48
- **Conventional Evaluation:** BLEU, Edit, TF-IDF
- **Command Execution Environment:** Docker Container
- **Language Model Evaluation:** OpenAI's GPT-4

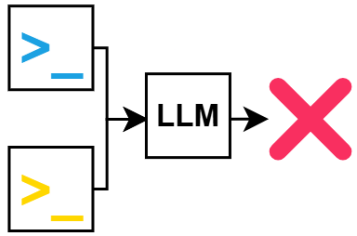
NL2SH Model Benchmarking Methods

Conventional Evaluation (Agarwal et al.)



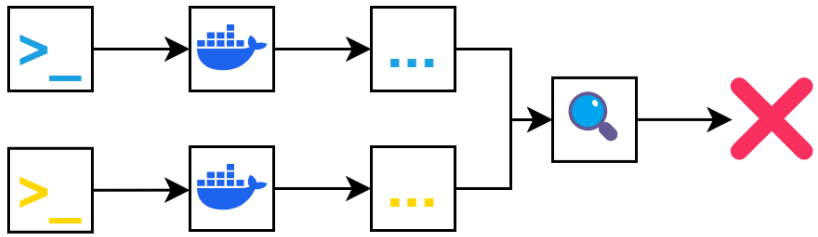
Syntactically different commands may be functionally equivalent

Language Model Evaluation (Song et al.)



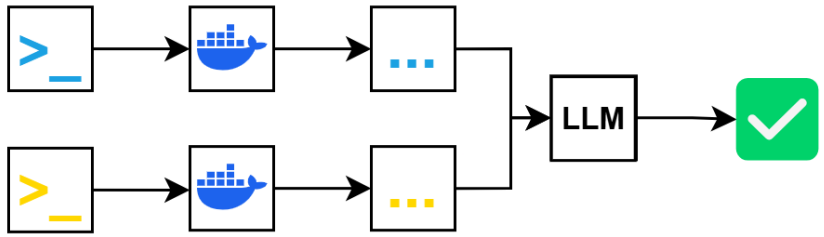
Current language models cannot emulate command execution

Command Execution + Conventional Evaluation (Yang et al.)

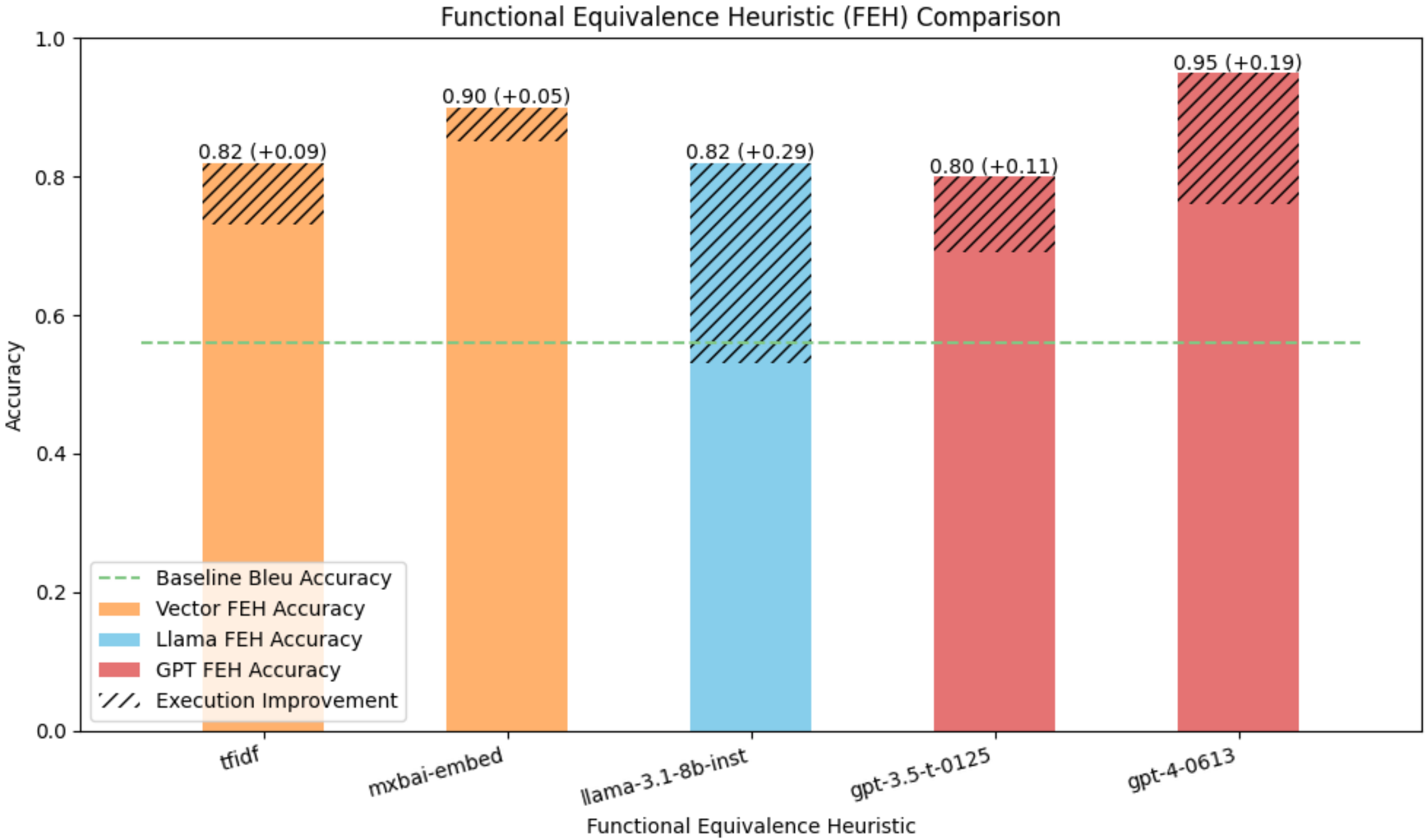


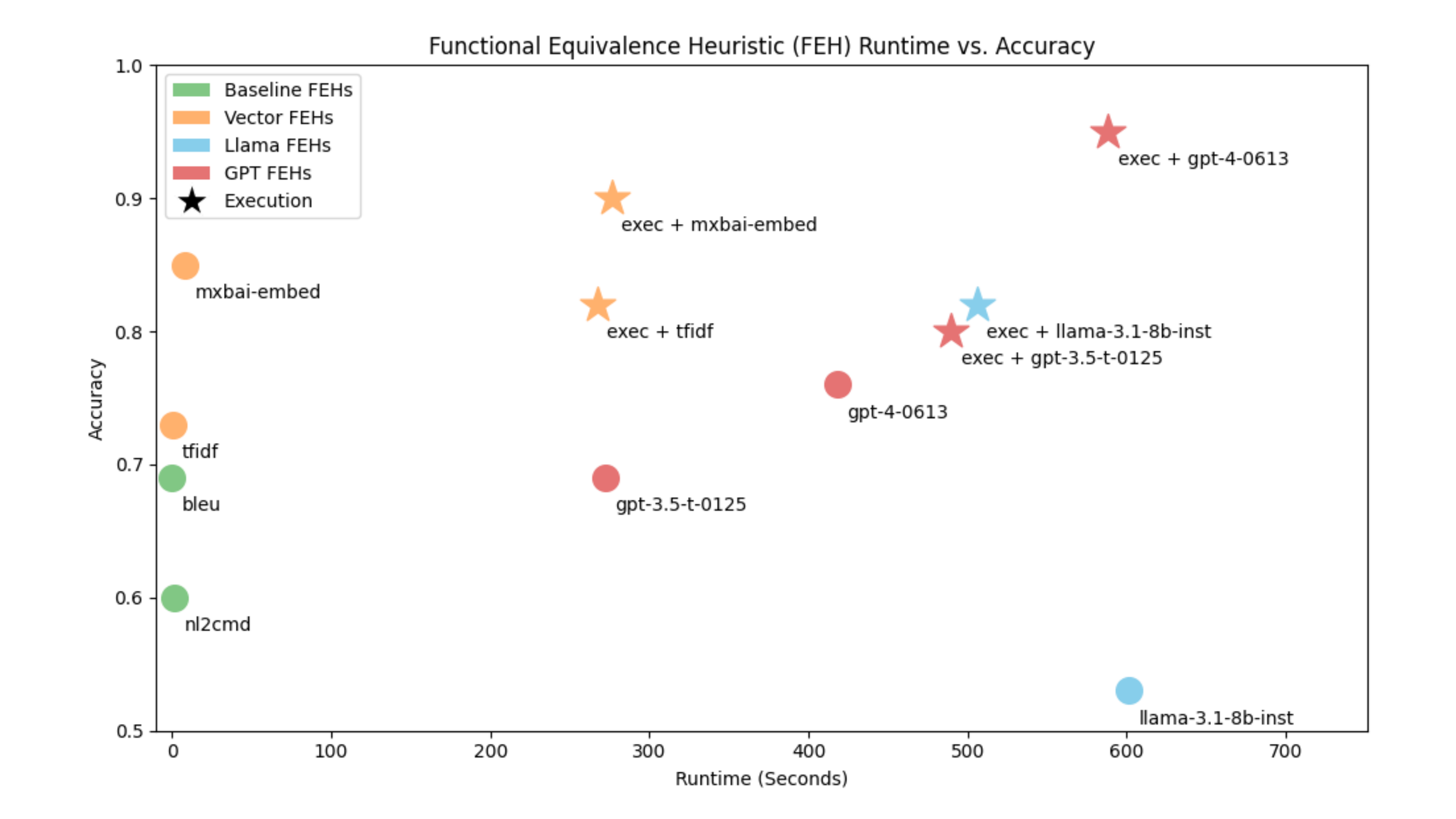
Syntactically different execution results may be functionally equivalent

Command Execution + Language Model Evaluation (Ours)



Command execution then language model evaluation determines functional equivalence


Functional Equivalence Heuristic (FEH) Comparison
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0.82 (+0.09)
0.90 (+0.05)
0.82 (+0.29)
0.80 (+0.11)
0.95 (+0.19)
Baseline Bleu Accuracy
Vector FEH Accuracy
Llama FEH Accuracy
GPT FEH Accuracy
Execution Improvement
tfidf
mxbai-embed
llama-3.1-8b-inst
gpt-3.5-t-0125
gpt-4-0613
Functional Equivalence Heuristic


Functional Equivalence Heuristic (FEH) Runtime vs. Accuracy
Baseline FEHs
Vector FEHs
Llama FEHs
GPT FEHs
Execution
exec + gpt-4-0613
exec + mxbai-embed
mxbai-embed
exec + tfidf
exec + llama-3.1-8b-inst
exec + gpt-3.5-t-0125
gpt-4-0613
tfidf
bleu
gpt-3.5-t-0125
nl2cmd
llama-3.1-8b-inst
Accuracy
Runtime (Seconds)
1.0
0.9
0.8
0.7
0.6
0.5
0
100
200
300
400
500
600
700

Benchmark Findings

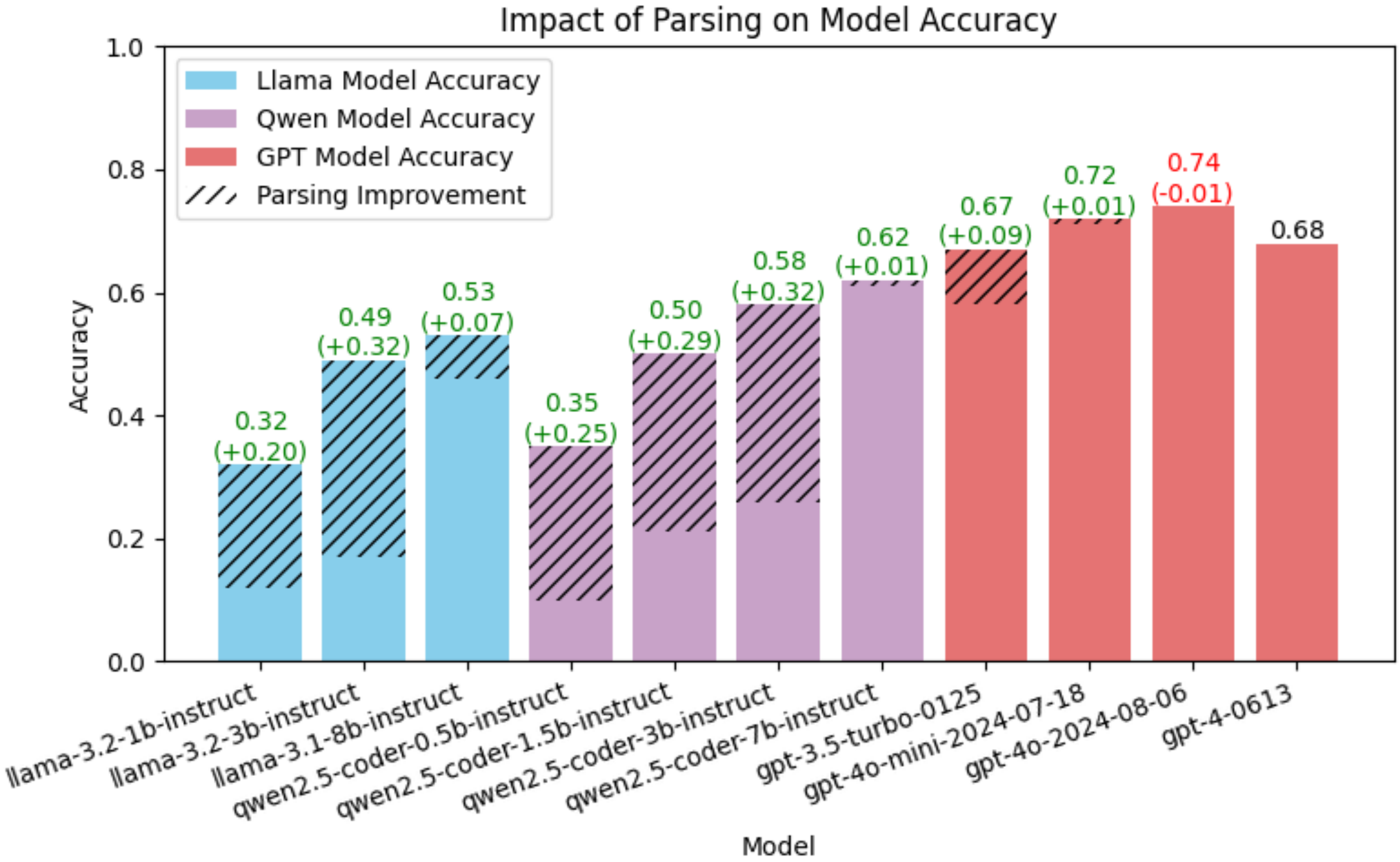
- **Research Question:** How can we design a functional equivalence heuristic that accurately measures the quality of model translations?
- **Result:** Command execution with language model evaluation of command outputs outperforms previous benchmarking methods.
- **Contribution:** We present a state-of-the-art method for determining the functional equivalence of Bash commands.

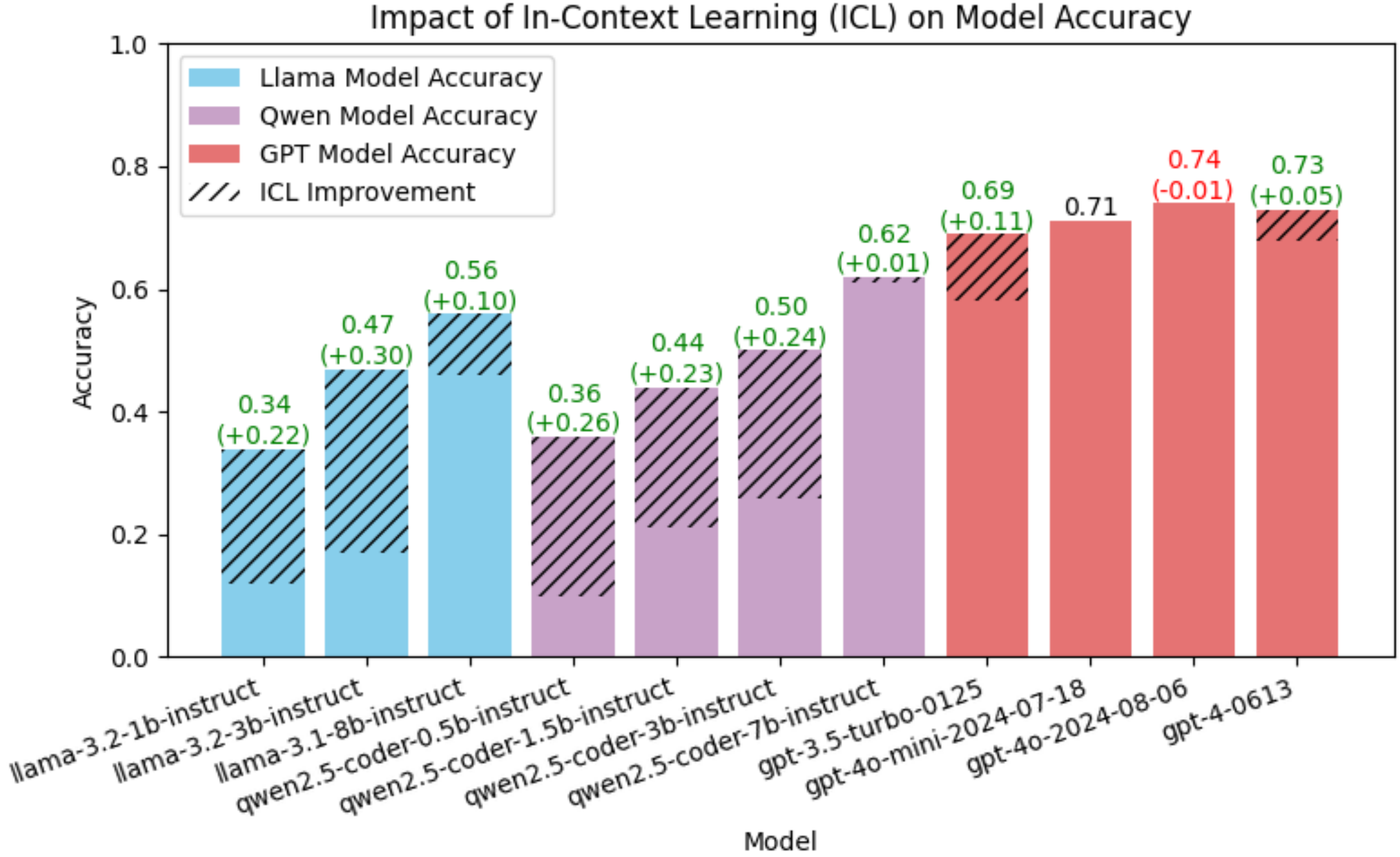
Translation Method Overview

- **Goal:**

Given	Determine
Natural Language	**Bash Command**
List files in /workspace accessed over an hour ago	`find /workspace -type f -amin +60`
base64 decode aGVsbG8	`echo 'aGVsbG8=' | openssl enc -base64 -d`
Sort and print only group names from /etc/group	`cut -d: -f1 /etc/group | sort`

- **Challenge:** LLMs have poor NL2SH performance.
- **Research Question:** How can we improve the NL2SH performance of LLMs as measured by a reliable benchmark?


Impact of Parsing on Model Accuracy
Llama Model Accuracy
Qwen Model Accuracy
GPT Model Accuracy
Parsing Improvement
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.32 (+0.20)
0.49 (+0.32)
0.53 (+0.07)
0.35 (+0.25)
0.50 (+0.29)
0.58 (+0.32)
0.62 (+0.01)
0.67 (+0.09)
0.72 (+0.01)
0.74 (-0.01)
0.68
llama-3.2-1b-instruct
llama-3.2-3b-instruct
llama-3.1-8b-instruct
qwen2.5-coder-0.5b-instruct
qwen2.5-coder-1.5b-instruct
qwen2.5-coder-3b-instruct
qwen2.5-coder-7b-instruct
gpt-3.5-turbo-0125
gpt-4o-mini-2024-07-18
gpt-4o-2024-08-06
gpt-4-0613
Model


Impact of In-Context Learning (ICL) on Model Accuracy
Llama Model Accuracy
Qwen Model Accuracy
GPT Model Accuracy
ICL Improvement
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.34 (+0.22)
0.47 (+0.30)
0.56 (+0.10)
0.36 (+0.26)
0.44 (+0.23)
0.50 (+0.24)
0.62 (+0.01)
0.69 (+0.11)
0.71
0.74 (-0.01)
0.73 (+0.05)
llama-3.2-1b-instruct
llama-3.2-3b-instruct
llama-3.1-8b-instruct
qwen2.5-coder-0.5b-instruct
qwen2.5-coder-1.5b-instruct
qwen2.5-coder-3b-instruct
qwen2.5-coder-7b-instruct
gpt-3.5-turbo-0125
gpt-4o-mini-2024-07-18
gpt-4o-2024-08-06
gpt-4-0613
Model

Translation Method Findings

- **Research Question:** How can we improve the accuracy of NL2SH models as measured by a reliable benchmark?
- **Result:** Parsing, in-context learning, in-weight learning and constrained decoding can improve NL2SH accuracy by up to 32%.
- **Contribution:** We demonstrate methods for improving the NL2SH performance of open and closed-source LLMs.

Conclusion

- Clean datasets are necessary for model training and evaluation
- Bash command execution paired with language model evaluation of command outputs can determine functional equivalence
- Parsing and in-context learning improve NL2SH performance
- NL2SH remains a difficult task and model outputs should be verified before they are used in real-world applications

Future Work

- Automate methods for dataset verification
- Evaluate the correctness of synthetic training data
- Increase speed and scalability of execution-based benchmarking
- Apply execution+LLM benchmarking to other translation tasks
- Evaluate the translation and functional equivalence capabilities of reasoning models
- Evaluate methods for increasing translation speed

Thank You!