



# LLM-Supported Natural Language to Bash Translation (NL2SH)

Finnian Westenfelder<sup>1,2</sup>, Erik Hemberg<sup>1</sup>, Miguel Tulla<sup>1</sup>,  
Stephen Moskal<sup>1</sup>, Una-May O'Reilly<sup>1</sup>, Silviu Chiricescu<sup>3</sup>

<sup>1</sup>ALFA Group MIT-CSAIL, <sup>2</sup>Draper Scholar, <sup>3</sup>Charles Stark Draper Laboratory



**Motivation** Bash commands are complex and difficult to memorize. Incorrect commands can cause system failures. Large Language Models (LLMs) are unreliable for translating natural language to Bash (NL2SH). Improving NL2SH requires (1) clean data, (2) reliable benchmarks, and (3) accurate translation methods.

## Dataset

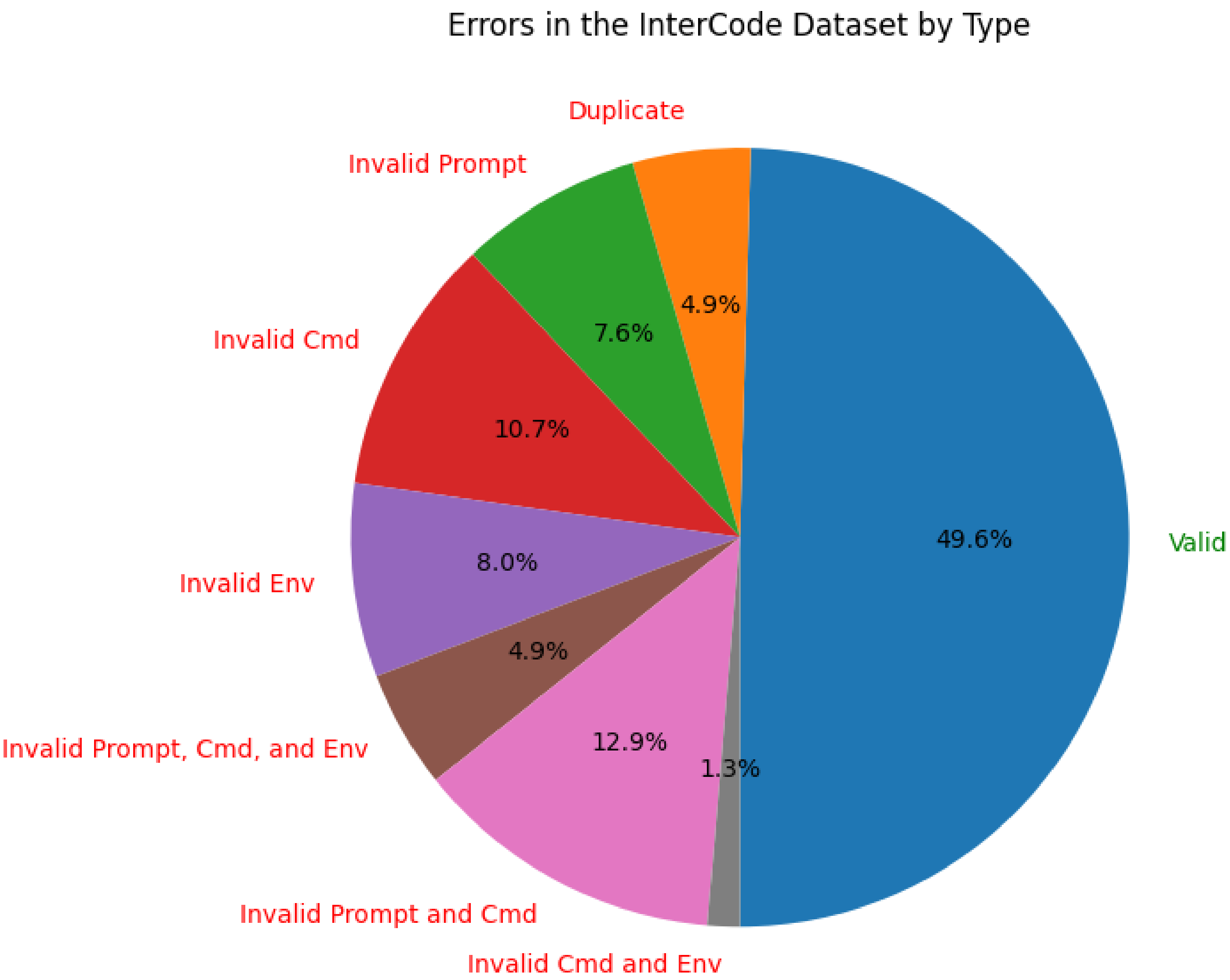
**Goal:**

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?

**Result:** Manual verification of test data improves benchmark reliability. Cleaning and increasing training data improves in-weight learning.



## Benchmark

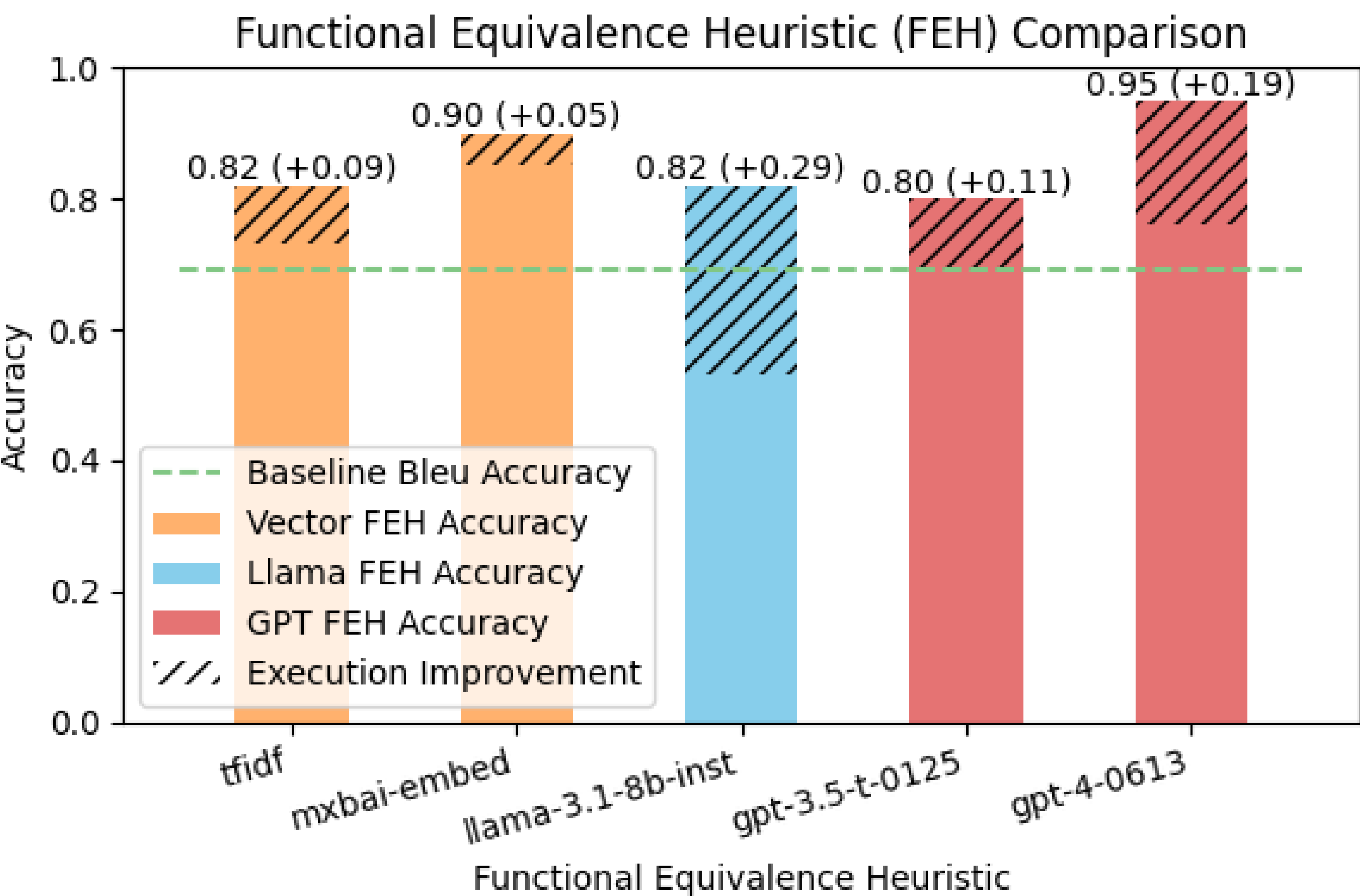
**Goal:**

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?

**Result:** Command execution with language model evaluation of command outputs outperforms previous benchmarking methods.



## Translation Method

**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 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%.

