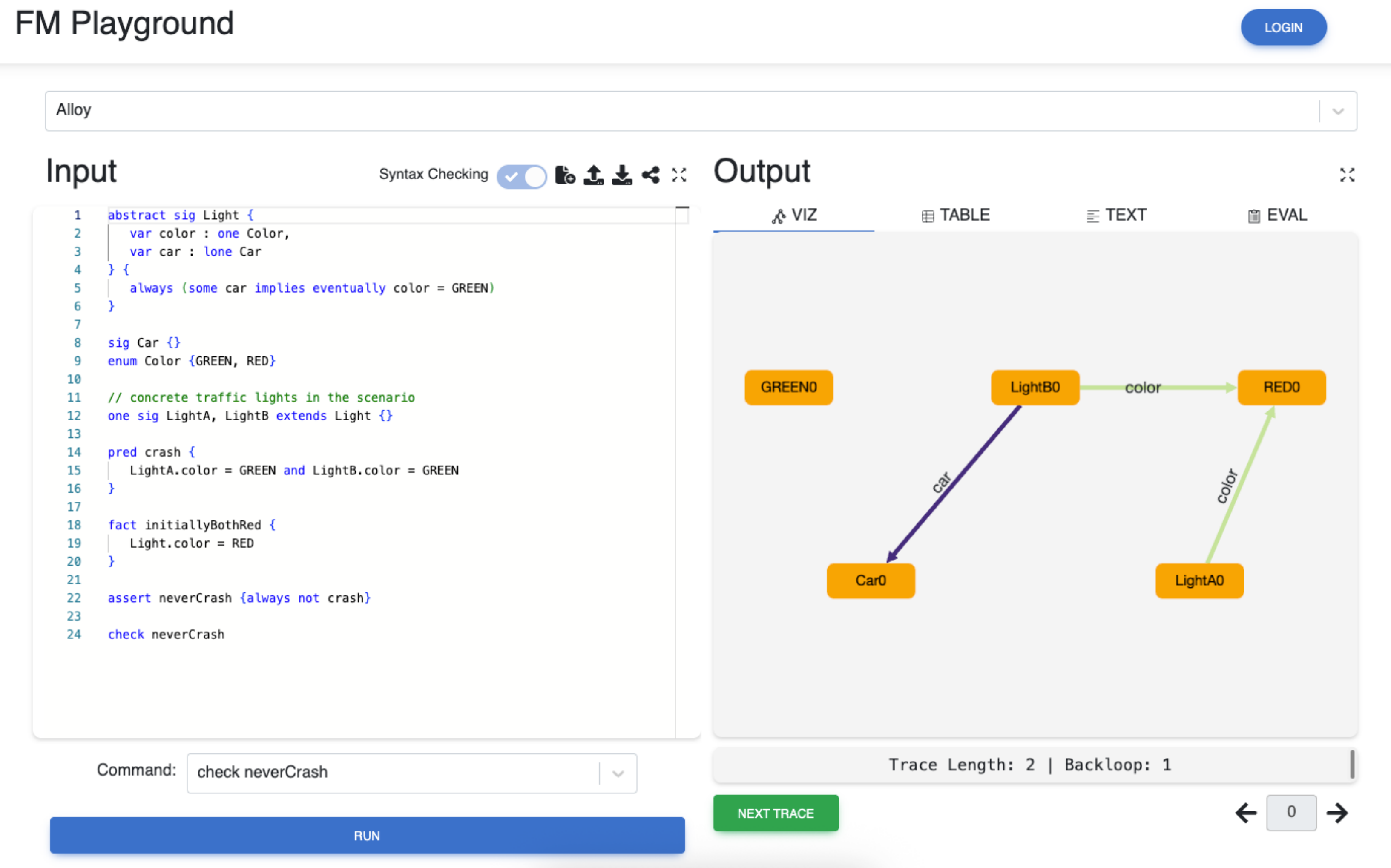
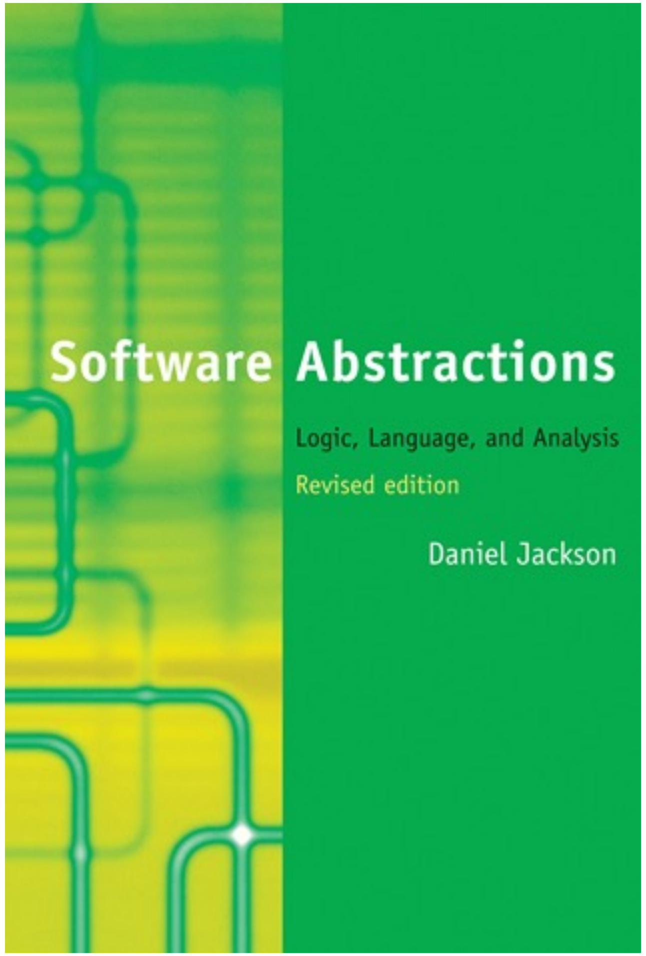
On Writing Alloy Models: Metrics and a new Dataset

Soaibuzzaman, Salar Kalantari, and Jan Oliver Ringert



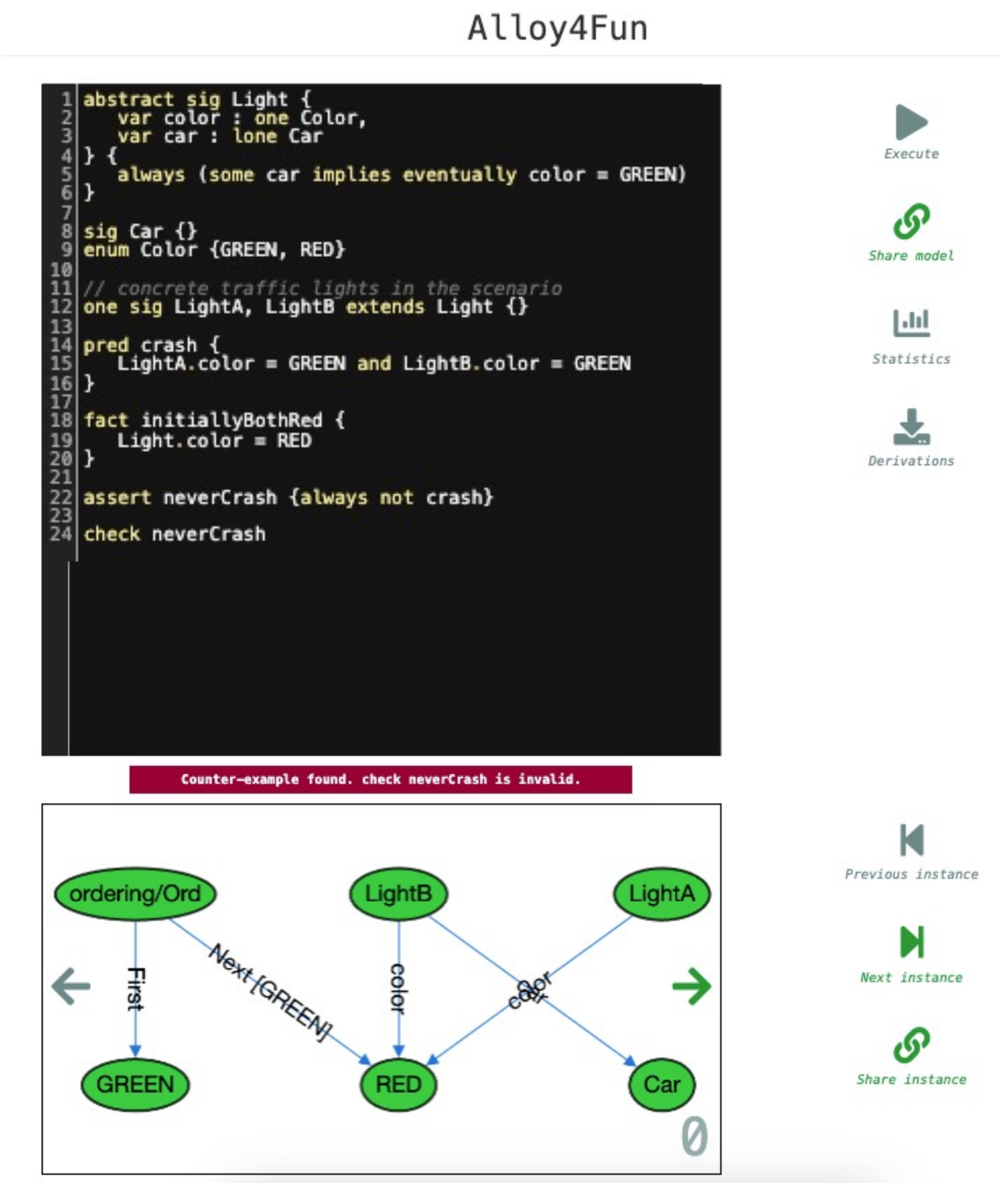
Alloy Language and Alloy Analyzer

- Alloy: Specification language based on relational first-order logic
  - Everything is a relation
- Alloy Analyzer: explore models and instances, check assertions
  - Quick feedback
  - Interactive specification development
- Applications: software design models, API design, protocol and security analyses, software synthesis, …



Existing: Alloy4Fun & A4F Dataset

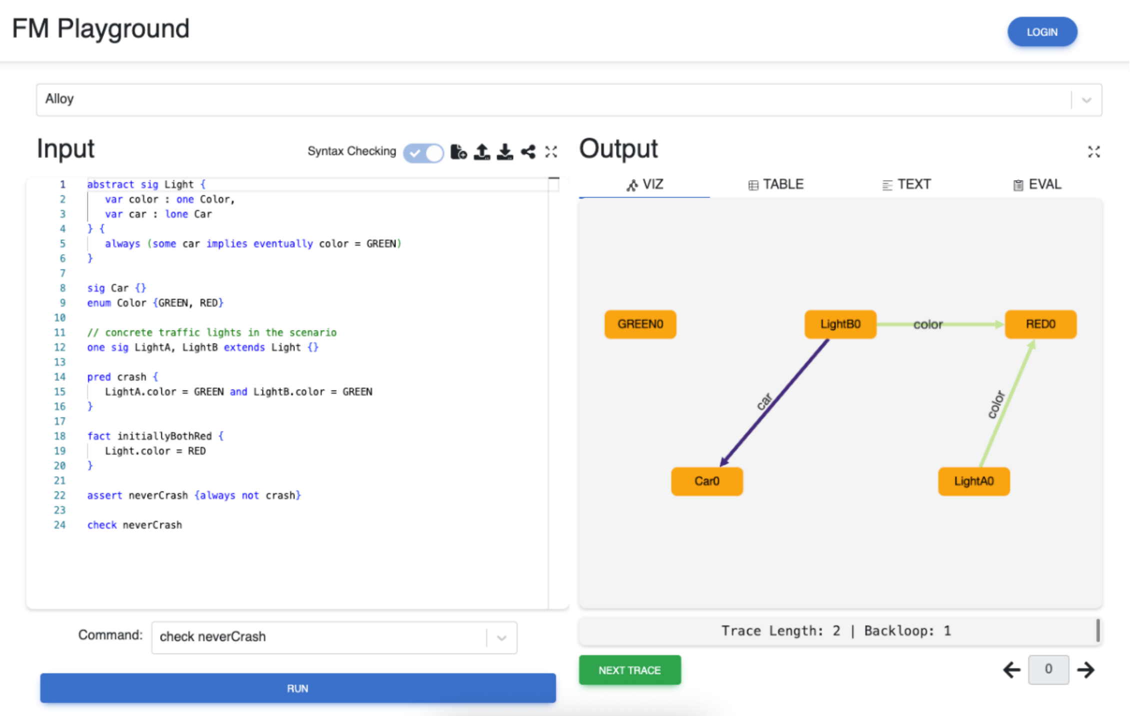
- A web application for writing and analyzing Alloy models intended for teaching Alloy
- Offers automated assessment and feedback by predefined predicates
- Dataset captures fine-grained editing histories.
- Focuses on predicate completion, not the full spectrum of Alloy modeling (signatures, fields, facts, commands)
- Helps to understand the process of writing Alloy models



Nuno Macedo, Alcino Cunha, José Pereira, Renato Carvalho, Ricardo Silva, Ana C. R. Paiva, Miguel Sozinho Ramalho, Daniel Castro Silva: *Experiences on teaching alloy with an automated assessment platform*. Sci. Comput. Program. 211: 102690 (2021)

New: FM Playground & FMP_{als} Dataset

- A web app for writing and analyzing models in various modeling and specification languages
- Provides different visualizations, i.e., graph, table, text, and an alloy evaluator
- Offers storage of permalinks, histories etc
- Try it at: https://play.formal-methods.net



Our Contribution

- FMP_{als} Dataset: A new, complementary dataset from the Formal Methods Playground.
  - More diverse: Users develop signatures, fields, facts, commands, etc.
  - Often starts from a blank canvas
- Comparative Analysis: Compared model evolution and metrics across FMP_{als} and Alloy4Fun
- Halstead Metrics for Alloy: Defined and applied a Halstead-based difficulty metric

Meet the Datasets: A4F vs. FMP_{als}

Alloy4Fun

- 96,397 models (after filtering)
- 5,268 unique edit paths (sequences of user submissions)
- Derived from 19 distinct starter models with multiple tasks
- A4FpT (per Task): Partitioned A4F paths for task-specific analysis (24,592 paths)

FMP_{als} (FM Playground Alloy)

- 8,219 Alloy models. (~22,000 now)
- 747 unique edit paths
- 392 unique initial models (many start from scratch)

RQ1: Dataset Characteristics - Errors & Similarity

- Error Types & Location

Dataset		Type	Syntax	sig	pred	fact	assert	fun	run	check
A4F	#	13657	13734	72	27202	26	1	52	22	11
	%	49.9	50.1	0.002	99.3	≈0	≈0	≈0.001	≈0	≈0
FMP_{als}	#	566	1962	376	625	769	101	97	378	87
	%	22.4	77.6	15.5	25.7	31.6	4.2	4	15.5	3.6

  – **Insight:** FMP_{als} shows users struggle with a broader range of Alloy constructs

- Submission Similarity

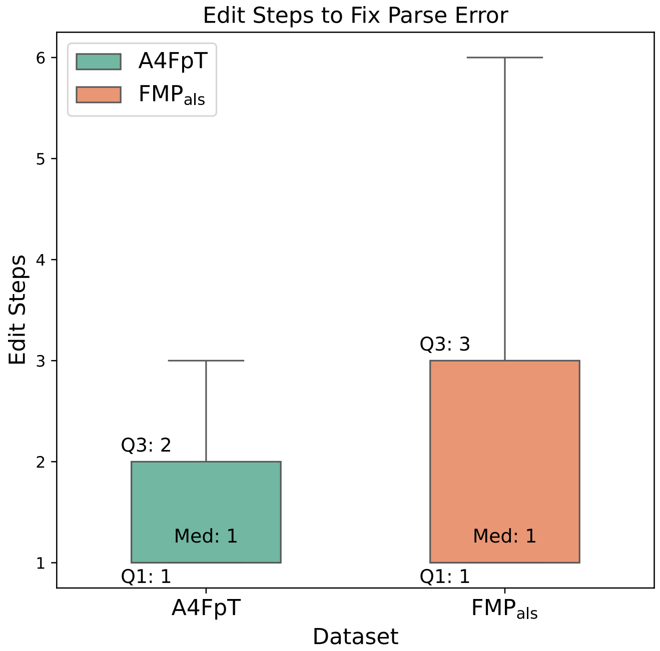
	A4FpT		FMP_{als}	
	#	%	#	%
Syntactically Unique Models	57777	59.9	3513	42.7
Syntactically Correct Models (in unique models)	37024	64.1	1880	53.5
Syntax Errors (in unique models)	20753	35.9	1633	46.5
Models within single edit paths:				
Consecutive Identical Models	4664	4.64	3174	25.58
Non-Consecutive Identical Models	5758	5.73	667	5.38

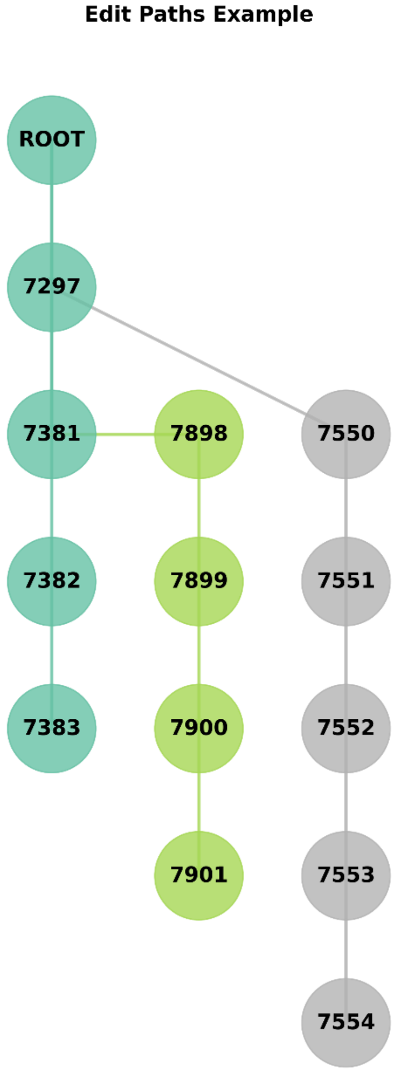
RQ1: Dataset Characteristics - Fixing Errors

- Error Presence in Edit Paths:

	A4FpT	FMP_{als}
Edit Paths (#)	24592	747
With Invalid Models (%)	39.24	54.08
Without Valid Models (%)	3.8	6.55
Edit Path Length ≥ 5 (%)	25.93	64.79
Max Edit Path Length	107	211

- Steps to Fix Errors:





Halstead Metrics for Alloy

$$D = \frac{\eta_1}{2} \times \frac{N_2}{\eta_2}, \frac{\#\,unique\ operators}{2} \times \frac{\#\,occurrences\ of\ operands}{\#\,unique\ operands}$$

$$\text{Difficulty, } D = \frac{5}{2} \times \frac{11}{7} = 3.93$$

```
sig File { link : set File }
sig Trash in File {}
sig Protected in File {}

pred inv1 { /* The trash is empty. */ no Trash}
pred inv2 { /* All files are deleted. */ }
pred inv3 { /* Some file is deleted. */ }
```

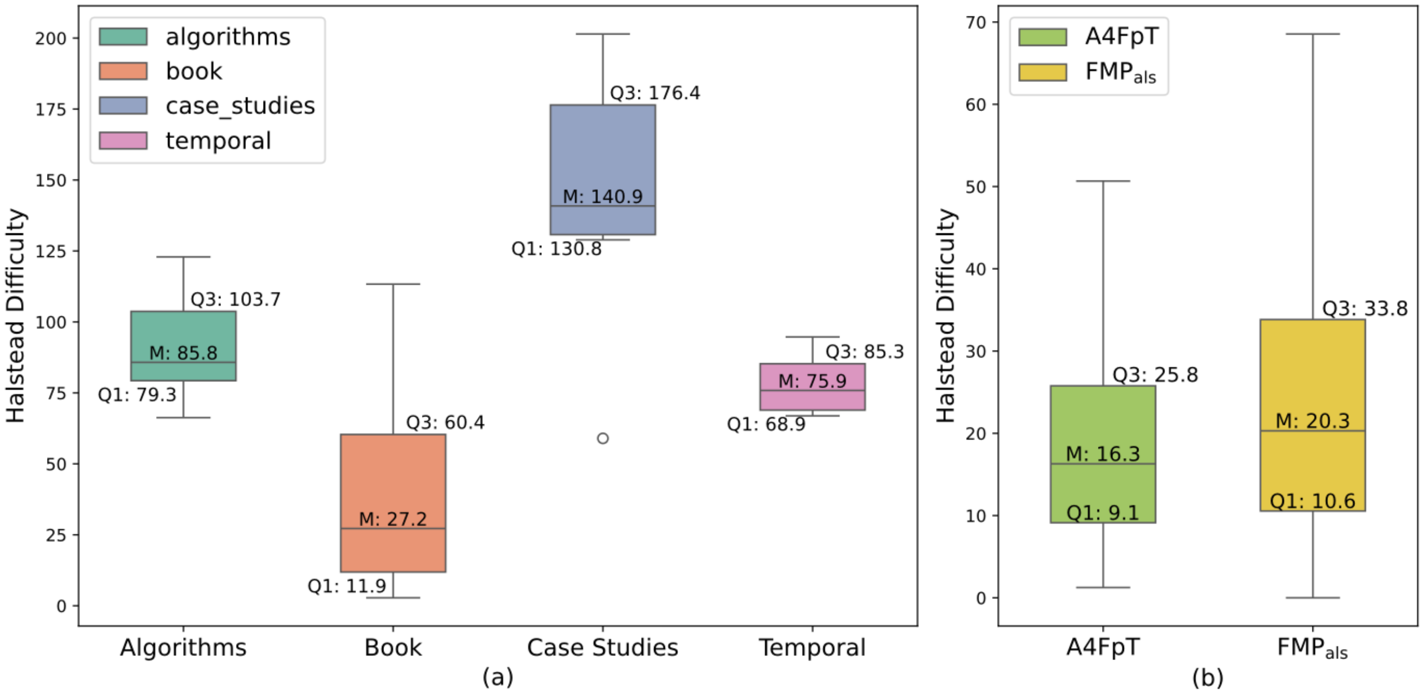
Operator	Count	Operand	Count
sig	3	inv1	1
no	1	inv2	1
set	1	inv3	1
in	2	Protected	1
pred	3	File	4
		link	1
		Trash	2
Total	**10**	**Total**	**11**

Our Counting Strategy for Alloy:

- Operators: Keywords (sig, pred, run), multiplicity (some, all), logical/arithmetic operators
- Operands: Names of modules, signatures, fields, variables, and literals

RQ2: Halstead Difficulty Comparison

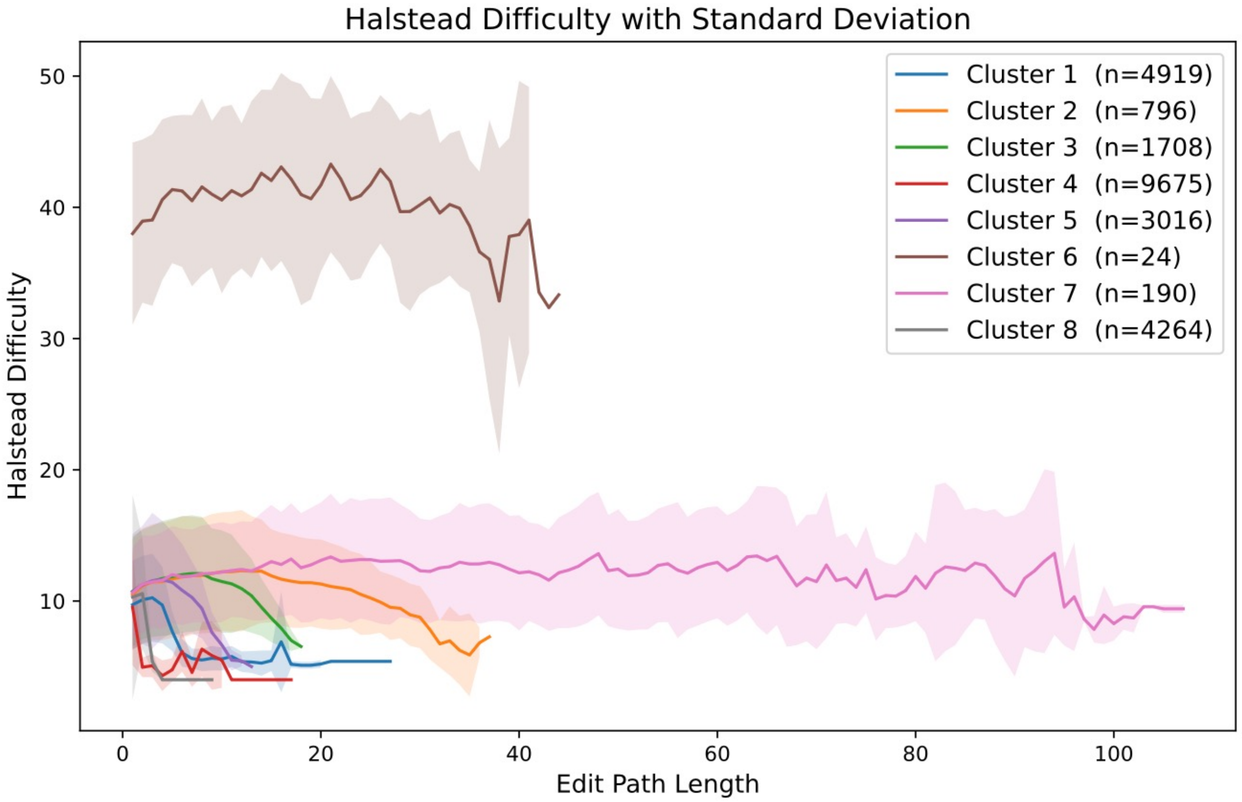
- Baseline: Alloy Analyzer Sample Models
  - "Book" examples: Median Difficulty ~27
  - "Case Studies": Much higher, Median Difficulty ~140
- Comparing Datasets - Final Submissions
  - A4FpT Median: ~16
  - FMP_{als} Median: ~20



Bauhaus-Universität Weimar

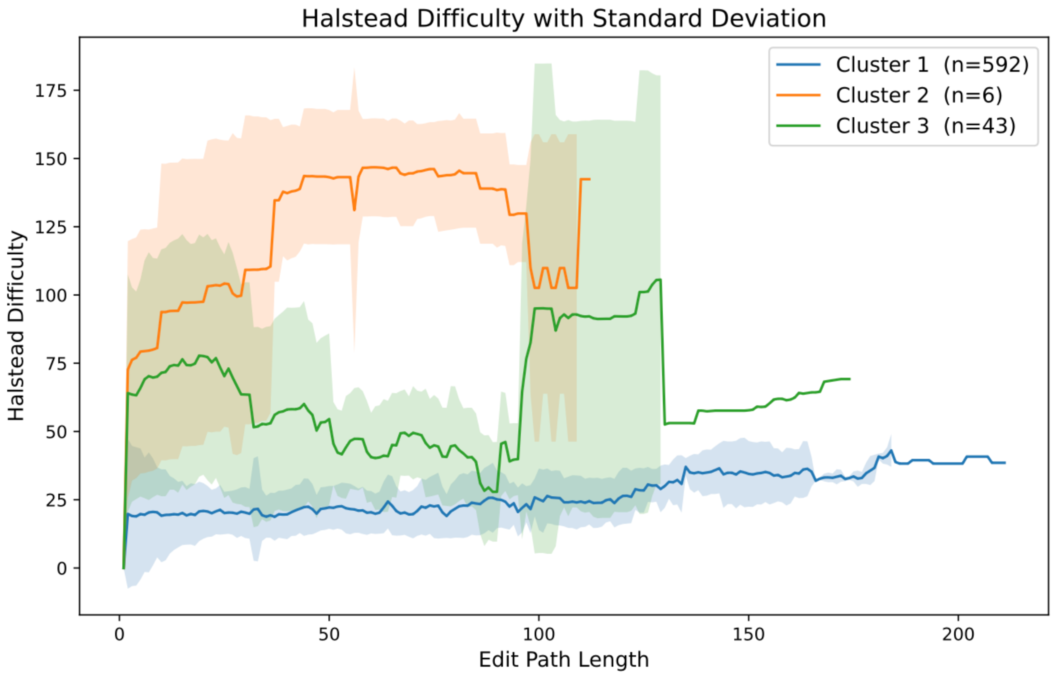
RQ3: How Does Difficulty Evolve?

- A4FpT (Alloy4Fun per Task)
  - 8 clusters of Halstead difficulty evolution
  - Most clusters show low standard deviation (consistent difficulty within cluster)
  - **Insight:** Many clusters show a decrease in difficulty towards the end of edit paths



RQ3: How Does Difficulty Evolve?

- FMP_{als} (Formal Methods Playground Alloy)
  - 3 main clusters.
  - Cluster 1 (largest): Broad range of difficulty, increasing or stable. Reflects iterative development
  - Cluster 2 (small, high difficulty): Mostly auto-generated models (student project)
  - Cluster 3 (diverse): Wide spectrum of complexity
  - **Insight:** FMP_{als} often shows more iterative growth in complexity, as users build from scratch

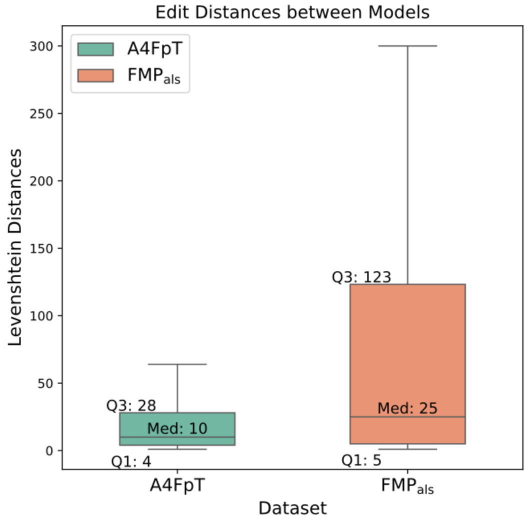


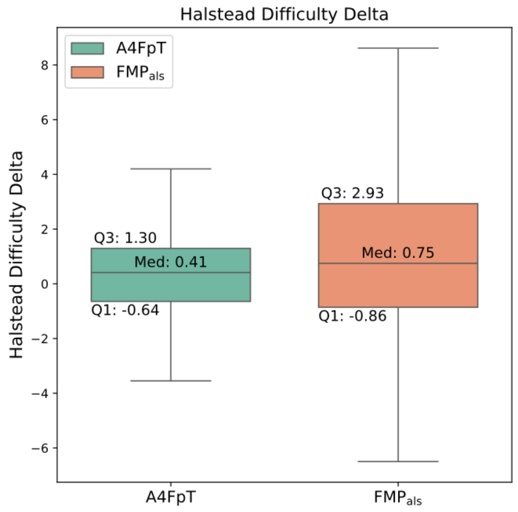
RQ4: Halstead Difficulty, Errors & Fixing Times

- Correlation: Difficulty vs. Time to Fix Errors:
  - A4FpT: Weak negative correlation (-0.032)
  - FMP_{als}: Weak positive correlation (0.236)
- Correlation: Difficulty vs. Error Occurrence (Logistic Regression):
  - A4FpT: Weak negative correlation
  - FMP_{als}: Weak positive correlation
- **Key Takeaway:** Halstead difficulty shows only weak correlations with error occurrence or fixing times. Other factors (user expertise, error nature) are likely more dominan

RQ5: Edit Sizes - Levenshtein & Difficulty Delta

- Levenshtein Distance (character changes between edits)
  - A4FpT: Smaller edits (Median: 10 chars)
  - FMP_{als} : Larger edits (Median: 25 chars)
- Halstead Difficulty Delta (difficulty change between edits):
  - Both: More than 25% of edits decrease difficulty
  - FMP_{als} : Larger median changes in difficulty
- **Insight:** Users on FMP_{als} make larger changes per edit step





Bauhaus-Universität Weimar

Key Findings & Implications

- FMP_{als} is a valuable, complementary dataset:
  - Shows challenges beyond predicate writing
  - Supports evaluation for various purposes: model repair, incremental solving, teaching materials
- Different Evolution Patterns:
  - FMP_{als} : More iterative growth, larger edits
  - A4FpT: Often ends with simplification/refinement, smaller, focused edits
- Halstead Difficulty:
  - Weak correlation with error rates/fixing times – not a sole indicator of "difficulty to get right”
- Tool Interaction: Repeated analyses in FMP_{als} might suggest different user interactions with instances or tool features

Conclusion & Future Work

- Presented FMP_{als} dataset, highlighting its unique characteristics
- Analyzed model evolution using Halstead difficulty, revealing distinct patterns
- Halstead difficulty is descriptive but not strongly predictive of error-fixing effort
- Future Work:
  - Updates to FMP_{als} as usage grows
  - Deeper analysis of error types and fixing strategies
  - Investigating user interaction with Alloy Analyzer instances

- Data availability:
  - Formal Methods Playground (public, open source)
  - Dataset updated on Zenodo

Questions?