Preliminary Results (Pythia and GPT-4)

The overall statistics:

We discovered a hidden bug in the public dataset (UT-10F), and the results are updated when reclassifying UT-10F to buggy (15 buggy and 13 fixed).
The following overall results shows the numbers **before reclassifying UT-10F to buggy -> after reclassifying UT-10F to buggy**.

Analyzers	Pythia	Pythia	GPT-4(UNION)	GPT-4(UNION)	GPT-4(MAJORITY)	GPT-4(MAJORITY)
RunInfo	without	with	without	with	without	with
FN	3 -> 4	2	5 -> 6	2 -> 3	8 -> 9	6 -> 7
FP	2	4 -> 3	3	3	3	3
Precision	0.8571 -> 0.8461	0.7647 -> 0.8125	0.7692 -> 0.75	0.8125 -> 0.8	0.7 -> 0.6667	0.75 -> 0.72727
Recall	0.8 -> 0.7333	0.8666 -> 0.8667	0.6667 -> 0.6	0.8667 -> 0.8	0.4667 -> 0.4	0.6 -> 0.5333
F1 Score	0.8275 -> 0.7857	0.8125 -> 0.8387	0.7143 -> 0.6667	0.8387 -> 0.8	0.56 -> 0.5	0.6667 -> 0.6154

*Should also update the numbers for GPT-4 results (already did for Pythia) when we move UT-10F to buggy category
*It doesn't affect our conclusions, but: with runtime information, Pythia performs **similar (with 1- FN)** to GPT-4(Union) instead of **same.**

1. Pythia

without runinfo: 4 FN, 2FP
with runinfo: 2 FN (-2), 3FP (+1)

Data	Buggy + runinfo	Buggy	Fix + runinfo	Fix
Ut1	warning	warning	/	/
Ut2	error	error	/	/
Ut3	error	error	/	/
Ut4	error	error	/	/
Ut5	FN, warning (FP)	FN, warning (FP)	warning (FP)	1 warning (FP)
Ut6	error	error	/	/
Ut7	/(FN)	/(FN)	/	/
Ut8	error	error	/	/
Ut9	error	error	/	/
Ut10	error	/ (FN)	-	-
Ut10F	error	/ (FN)	-	-
Ut11	error	error	/	/
Ut12	error (points to the crush line), warning	warning	error (FP)	/
Ut13	warning	warning	/	/
Ut15	warning	warning	/	/

2. GPT-4

UNION over 5 runs:
without runinfo: 6FN, 3FP
with runinfo: 3FN (-3), 3FP (-)
MAJORITY over 5 runs (appear >= 3 times):
without runinfo: 9FN, 3FP
with runinfo: 7FN (-2), 3FP (-)

	UNION	UNION	MAJORITY	MAJORITY	UNION	UNION	MAJORITY	MAJORITY
Data	Buggy + runinfo	Buggy	Buggy + runinfo	Buggy	Fix + runinfo	Fix	Fix + runinfo	Fix
Ut1	error	/ (FN)	/ (FN)	/ (FN)	/	/	/	/
Ut2	error	error	error	error	/	/	/	/
Ut3	error	error	error	error	/	/	/	/
Ut4	error (3/3)	error (1/3)	/ (FN)	/ (FN)	/	/	/	/
Ut5	error	error	error	error	/	/	/	/
Ut6	error (FP), FN	error (FP), FN	error (FP), FN	error (FP), FN	error (FP)	error (FP)	error (FP)	error (FP)
Ut7	error	error	/ (FN)	/ (FN)	/	/	/	/
Ut8	error	error	error	error	/	/	/	/
Ut9	error	/ (FN)	error	/ (FN)	error (FP)	error (FP)	error (FP)	error (FP)
Ut10	error	error	/ (FN)	/ (FN)	-	-	-	-
Ut10F	/ (FN)	/ (FN)	/ (FN)	/ (FN)	-	-	-	-
Ut11	error	error	error	error	/	/	/	/
Ut12	/ (FN)	/ (FN)	/ (FN)	/ (FN)	/	/	/	/
Ut13	error	error	error	error	/	/	/	/
Ut15	error	/ (FN)	error	/ (FN)	/	/	/	/

3. Comparison between Pythia and GPT-4(Union) with runinfo (best results from each static analyzer)

FN and FP cases are distinct

	Pythia	GPT-4(Union)	Pythia	GPT-4(Union)
	Buggy + runinfo	Buggy + runinfo	Fix + runinfo	Fix + runinfo
Ut1	warning	error	/	/
Ut2	error	error	/	/
Ut3	error	error	/	/
Ut4	error	error	/	/
Ut5	FN, warning (FP)	error	warning (FP)	/
Ut6	error	error (FP), FN	/	error (FP)
Ut7	/(FN)	error	/	/
Ut8	error	error	/	/
Ut9	error	error	/	error (FP)
Ut10	error	error	-	-
Ut10F	error	/ (FN)	-	-
Ut11	error	error	/	/
Ut12	error, warning	/ (FN)	error (FP)	/
Ut13	warning	error	/	/
Ut15	warning	error	/	/