Automatically Enhancing the Quality of Android App Bug Reports


Antu Saha
William & Mary


Atish Kumar Dipongkor
Univ. of Central Florida


Sam Bennett
William & Mary


Kevin Moran
Univ. of Central Florida


Andrian Marcus
George Mason University


Oscar Chaparro
William & Mary

Problem & Overview

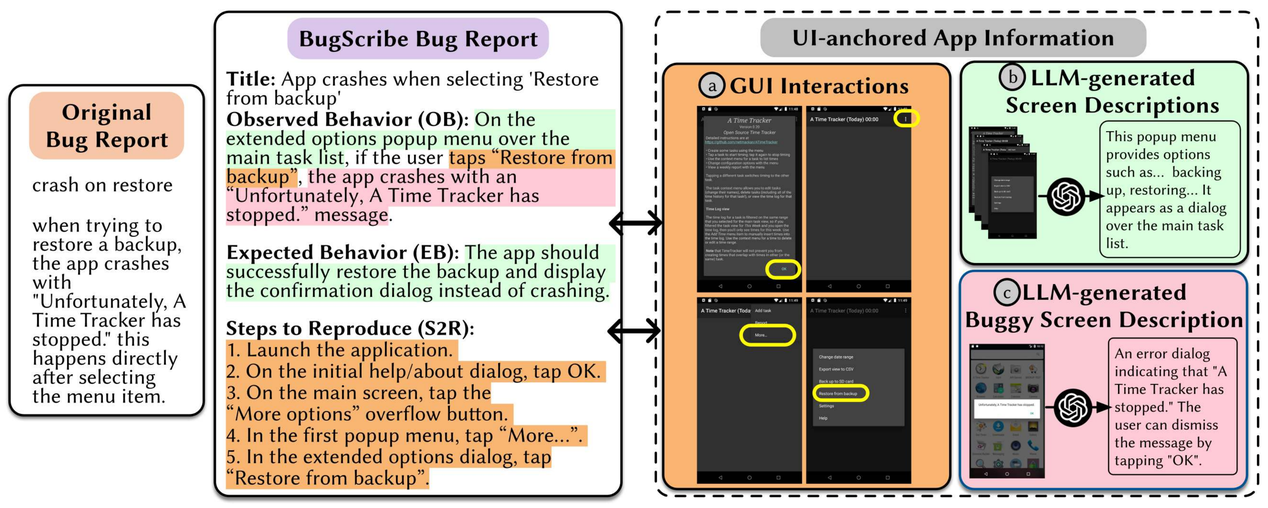
Problem: Low-quality Android bug reports hinder developer understanding, delay resolution, and cause unresolved bugs

- Reports miss essential components: Observed Behavior (OB), Expected Behavior (EB), and Steps to Reproduce (S2Rs)
- Contains incomplete or ambiguous components

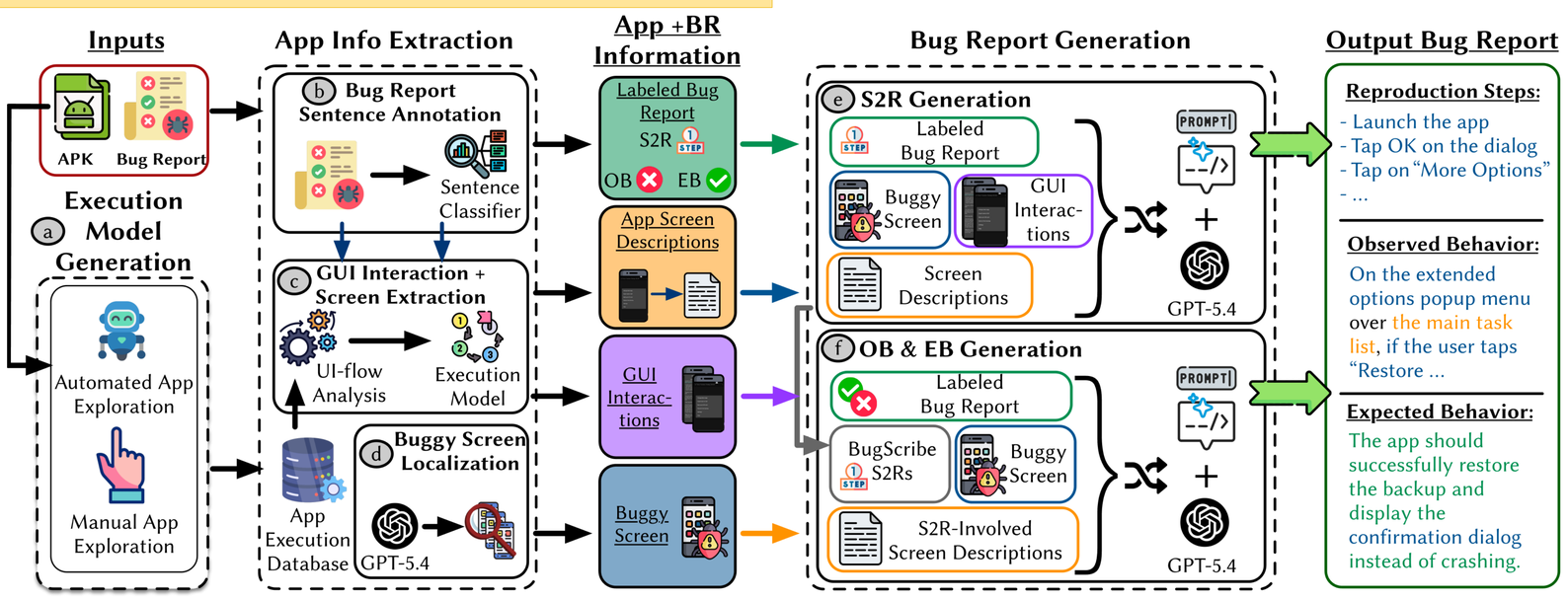
Key Insight: Connect the report with the app information (e.g., screens & GUI interactions) to extract required information

BUGSCRIBE: Android bug LLM-powered approach to enhance user-written reports with complete, correct OB, EB, and S2Rs

Motivating Example



BUGSCRIBE Approach



Which app info helps? (dev set: 10 reports)

S2R Generation

	CS (TP)	ES (FP)	MS (FN)	Precision	Recall	F1 Score
No Info.	80.7	34.3	28.3	70.2	74.0	72.0
Interac.	95.3	14.0	13.7	87.2	87.5	87.3
Interac.+Scrn.	96.7	16.0	12.3	85.8	88.7	87.2
Interac.+Scrn.+BuggyScrn.	96.7	13.7	12.3	87.6	88.7	88.2

OB/EB Generation

	Correct	Incomplete	Ambiguous	Missing	Incorrect
No Info.	30.3	5.3	1.7	0	2.7
Buggy Scrn.	32.3	5.3	0	0	2.3
S2Rs+BuggyScrn.	33.3	5.3	0.3	0	1.0
S2Rs+BuggyScrn.+Scrns.	34.3	4.3	0	0	1.3

BUGSCRIBE requires component-specific app context to generate high-quality OB/EB and S2Rs

S2R → GUI interactions + screens + buggy screen

OB/EB → Generated S2Rs + their screens + buggy screen

Evaluation

48 Bug Reports · 26 Android Apps · Manual Ground Truth · GPT-5.4 & Claude 4.6

Baselines: $U2S_{BR}$ (restructured reports), $GPT/CLAUDE_{NO\text{-}INFO}$ (no app info)

RQ1: S2R Quality

	CS (TP)	ES (FP)	MS (FN)	Precision	Recall	F1 Score
Original BR	193	4	406	98.0	32.2	48.5
$U2S_{BR}$	243	75	356	76.4	40.6	53.0
$GPT_{NO\text{-}INFO}$	345	183	254	65.3	57.6	61.2
$CLAUDE_{NO\text{-}INFO}$	312	135	287	69.8	52.1	59.7
BS_{GPT}	532	57	67	90.3	88.8	89.6
BS_{CLAUDE}	538	82	61	86.8	89.8	88.3

BUGSCRIBE improves the S2R quality by up to 84.7% over the baselines

RQ2: OB/EB Quality

	Correct	Incomplete	Ambiguous	Missing	Incorrect
Original BR	122	8	3	59	0
$U2S_{BR}$	123	14	8	34	13
$GPT_{NO\text{-}INFO}$	143	19	2	12	16
$CLAUDE_{NO\text{-}INFO}$	159	12	6	6	9
BS_{GPT}	159	12	2	0	19
BS_{CLAUDE}	165	11	4	3	9

BUGSCRIBE improves the OB/EB quality by up to 35.2% over the baselines

Application: Bug Reproduction

An LLM agent (MOBILERUN) tries to reproduce 40 bugs automatically from each report version, three executions each

RQ3: Automated Bug Reproduction

	Reproduction Rate (%)	Step Execution Rate (%)
Original BR	43.3 (40.0–47.5)	66.2
$U2S_{BR}$	44.2 (42.5–47.5)	71.9
$GPT_{NO\text{-}INFO}$	53.3 (50.0–60.0)	81.5
BS_{GPT}	68.3 (67.5–70.0)	91.6
Ground Truth	70.8 (67.5–75.0)	96.9

BUGSCRIBE reports improves the automated bug reproduction rate by up to 57.7% compared to the baseline reports

Conclusion & Future Work

- Connecting bug reports to app execution and extracting required information yields higher-quality OB, EB, and S2Rs
- Each report component needs its own app context
- BUGSCRIBE outperforms the original reports and LLM-based baselines that rely on report text alone
- BUGSCRIBE reports help an LLM agent reproduce more bugs

Future Work: We will extend BUGSCRIBE to other platforms (e.g., iOS, web) and integrate it into issue trackers