

대형언어모델 기반
자동디버깅 툴 평가를
위한

**GitHub Recent Bugs
(GHRB) 데이터셋**

이재용, 강성민, 윤주연, 유신

고기잡이, 김홍도

LLM 데이터 ‘오염’ 예시

Matching Text

Found spans are in grey. The longest span is in blue. Hovering over a character highlights the longest span that includes that character (there may be overlapping shorter spans). Clicking shows the component substrings below.

```
@Override
public JsonDeserializer<?> createContextual(DeserializationContext ctxt,
BeanProperty property) throws JsonMappingException
{
    // May need to resolve types for delegate-based creators:
    JsonDeserializer<Object> delegate = null;
    if (_valueInstantiator != null) {
        // [databind#2324]: check both array-delegating and delegating
        AnnotatedWithParams delegateCreator = _valueInstantiator.getDelegateCreator();
        if (delegateCreator != null) {
            JavaType delegateType = _valueInstantiator.getDelegateType(ctxt.getConfig());
            delegate = findDeserializer(ctxt, delegateType, property);
        }
    }
    JsonDeserializer<?> valueDeser = _valueDeserializer;
    final JavaType valueType = _containerType.getContentTypeInfo();
    if (valueDeser == null) {
```

Figure 1. A screenshot of the data inclusion website of Star-Coder. Entire portions of the method are included, and the overlap break (gray to blue) happens at the fix location.

Defects4J는 ‘오염’ 되어있다

Table 1. Compromised bugs by Defects4J v1.0 project

Project	# Bugs	% Comp.	Project	# Bugs	% Comp.
Math	104	67.3%	Lang	61	78.7%
Time	25	92.0%	Chart	25	72.0%
Closure	130	35.4%			


기준: 버그 함수 혹은 버그 재현 테스트가 StarCoder데이터에 90% 이상 포함

그래서 (반자동으로) 새 버그를 모았습니다

TABLE I
PROJECTS AND NUMBER OF REAL BUGS AVAILABLE IN THE PUBLIC RELEASE OF GHRB (AS OF 19 JANUARY 2024)

Project	Bugs	LoC	Test LoC	# Tests	# Stars	Project	Bugs	LoC	Test LoC	# Tests	# Stars
fastjson	1	43.6k	143.4k	>435	25.4k	jackson-dataformat-xml	1	5.9k	9.4k	>2	541
nacos	6	215.9k	8.7k	2.7k	27.4k	gson	12	9.0k	19.6k	1.3k	22.4k
dubbo	1	146.7k	89.4k	3.3k	39.3k	sslcontext	6	3.7k	7.2k	497	406
rocketmq	17	150.3k	54.7k	1.7k	19.8k	jsoup	5	14.3k	12.5k	1.1k	10.3k
assertj	4	45.9k	161.4k	11.8k	2.4k	openapi-generator	5	9.8k	37.6k	1.8k	17.5k
checkstyle	15	41.7k	238.2k	4.5k	7.8k	seata	2	166.4k	30.5k	1.0k	24.2k
jackson-core	3	30.7k	44.8k	>100	2.2k	retrofit	1	3.9k	6.5k	329	42k
jackson-databind	5	73.1k	71.0k	>28	3.3k	Apktool	1	10.7k	3.4k	202	17.6k
Total							85	972	938k	30.8k	263k

(반자동으로) 매달 새 버그가!

 **github-actions** (bot) commented 2 weeks ago · edited by jaeyonglee0205

Data Gathering Statistics

Category	Number of PRs
PRs within relevant period	469
... that also have test files	210
... that also only mention a single issue	206
... that also are in English	168
... that also were merged in the main branch	132
... for which a valid test could also be extracted from the PR	107

Output Summary

Repo	Number of Possible Bugs
apache_rocketmq	38
OpenAPITools_openapi-generator	4
FasterXML_jackson-databind	8
jhy_jsoup	1
alibaba_nacos	2
apache_dubbo	4
Hakky54_sslcontext-kickstart	8

인터페이스도 있어요



🔗 Using GHRB

Some example commands for the command-line interface are as follows:

(Note that all directory paths given to the tool need to be absolute paths within the container.)

1. `info` - View information about a project or a particular bug.
 - Example: `./cli.py info -p gson`
2. `checkout` - checkout the buggy or fixed version of a bug.
 - Example: `./cli.py checkout -p gson -v 1b -w /root/framework/testing`
3. `compile` - compile the code in a directory.
 - Example: `./cli.py compile -w /root/framework/testing`
4. `test` - run the tests for a project.
 - Example: `./cli.py test -w /root/framework/testing`

Information of other commands can be retrieved using the `-h` flag.

