SWE-bench Multimodal

Do AI Systems Generalize to Visual Software Domains?

John Yang*[1], Carlos E. Jimenez*[2], Alex L. Zhang[2], Kilian Lieret[2],
Joyce Yang[3], Xindi Wu[2], Ori Press[4], Niklas Muennighoff[1],
Gabriel Synnaeve[5], Karthik Narasimhan[2], Diyi Yang[1], Sida I. Wang[5], Ofir Press[2]

[1]Stanford University, [2]Princeton Language & Intelligence, Princeton University,
[3]Cornell University, [4]Tübingen AI Center, University of Tübingen, [5]Meta AI

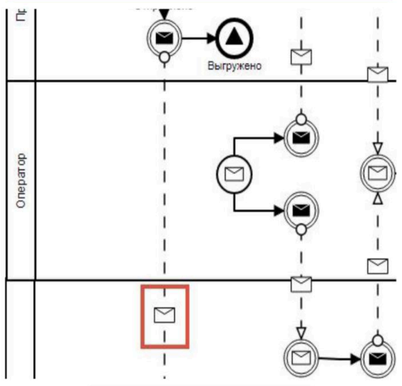
More than text: Many GitHub issues contain images. Do AI SWE agents generalize?

Diagramming

Show message element name

Currently, names of message elements on message flows are not rendered

Given this example diagram [Image] ...



bpmn-js

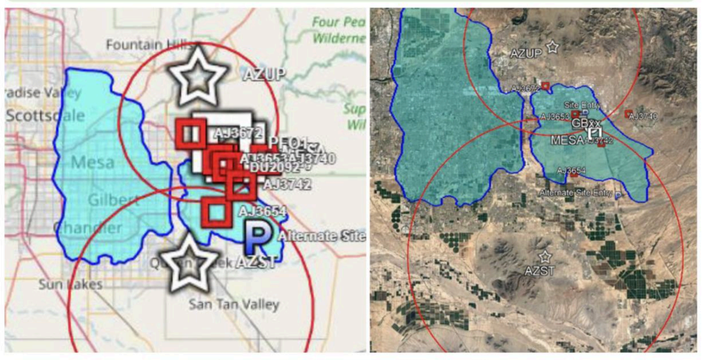
Interactive Mapping

KML Symbol Align/Placement/Size

There is a bug with the anchor point for some symbols

I've attached a screen clipping from Google Earth to show how it is supposed to look.

[Right Image] ...



openlayers

Syntax Highlighting

Bracket highlighted with different color in class inheritance context.

- Reproduced in JSFiddle: https://jsfiddle.net/kkangmj/e7h48w36/7/

[Image] ...

```
open class Tag

class TABLE: Tag {
  fun tr(init: TR.() -> Unit)
}
class TR: Tag {
  fun td(init: TD.() -> Unit)
}
class TD: Tag
```

highlight.js

Web Frameworks

[CascaderSelect]使用虚拟滚动时背景色异常

Component: CascaderSelect

Steps to reproduce

[Image] ...

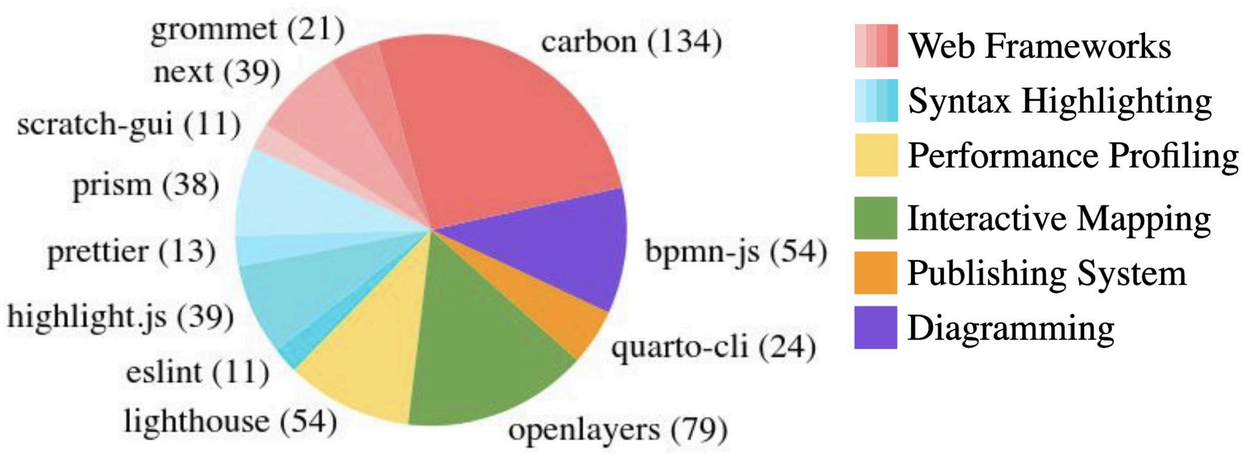
next

Collecting 517 multimodal issues from 12 GitHub Repos

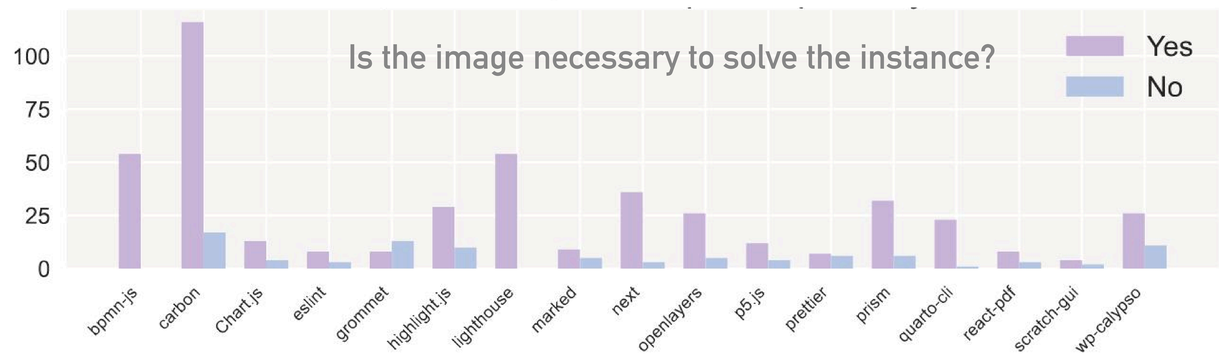
Scrape PRs (1)
- 12 popular repos
- >70% JS

Attribute Filter (2)
- ✓ Resolves issue
- ✓ Adds tests

Execution Filter (3)
- ✓ Installs
- ✓ PR passes tests



Various instance annotations, e.g.,



Multimodal issues are difficult!

System	Model	% Resolved
SWE-agent M	GPT-4o	**12.2**
	Claude 3.5 Sonnet	11.4
SWE-agent JS	GPT-4o	9.2
	Claude 3.5 Sonnet	12.0
SWE-agent Base	GPT-4o	12.0
	Claude 3.5 Sonnet	**12.2**
Agentless JS	GPT-4o	3.1
	Claude 3.5 Sonnet	6.2
RAG	GPT-4o	6.0
	Claude 3.5 Sonnet	5.0

Are the images necessary to solve this task?

System	With Images?	Yes	No
SWE-agent JS	✓	17.6	11.1
	✗	8.8	22.2
RAG	✓	12.4	11.9
	✗	9.1	8.1

Evaluating on SWE-bench (multimodal)

Issue

data leak in GBDT due to warm start (This is about the non-histogram-based version of...

Codebase

- sklearn/
- examples/
- README.rst
- reqs.txt
- setup.cfg
- setup.py

→ **Language Model** →

Generated PR +20 −12

- sklearn
  - gradient_boosting.py
  - helper.py
- utils

→ **Unit Tests**

Pre PR	Post PR	Tests
✗	✓	join_struct_col
✗	✓	vstack_struct_col
✗	✓	dstack_struct_col
✓	✓	matrix_transform
✓	✓	euclidean_diff