Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models (2022)

Priyan Vaithilingam, Tianyi Zhang, Elena Glassman

Selina Cheng

Overview

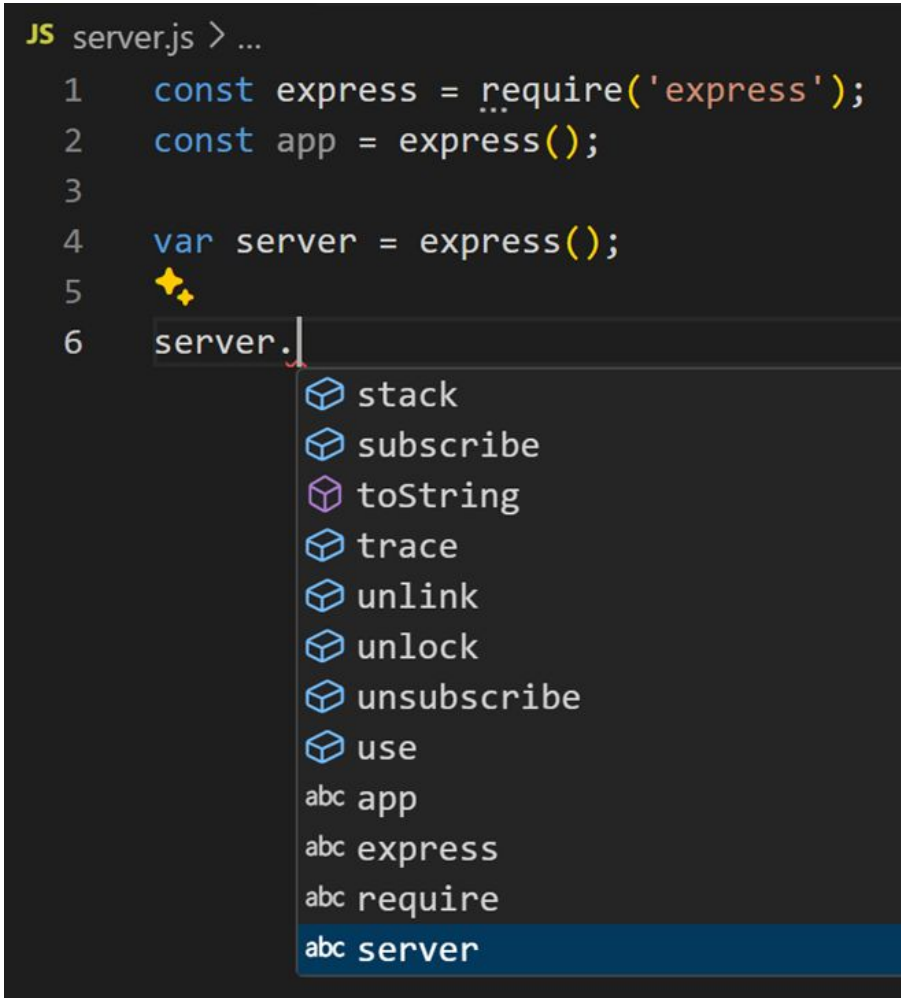
- **Goal:** Study the usability of GitHub Copilot, fit with programming workflow, and programmer perception
- **Why:**
  - Prior work on benchmarks, not usability
  - “Almost accurate yet not perfect code”

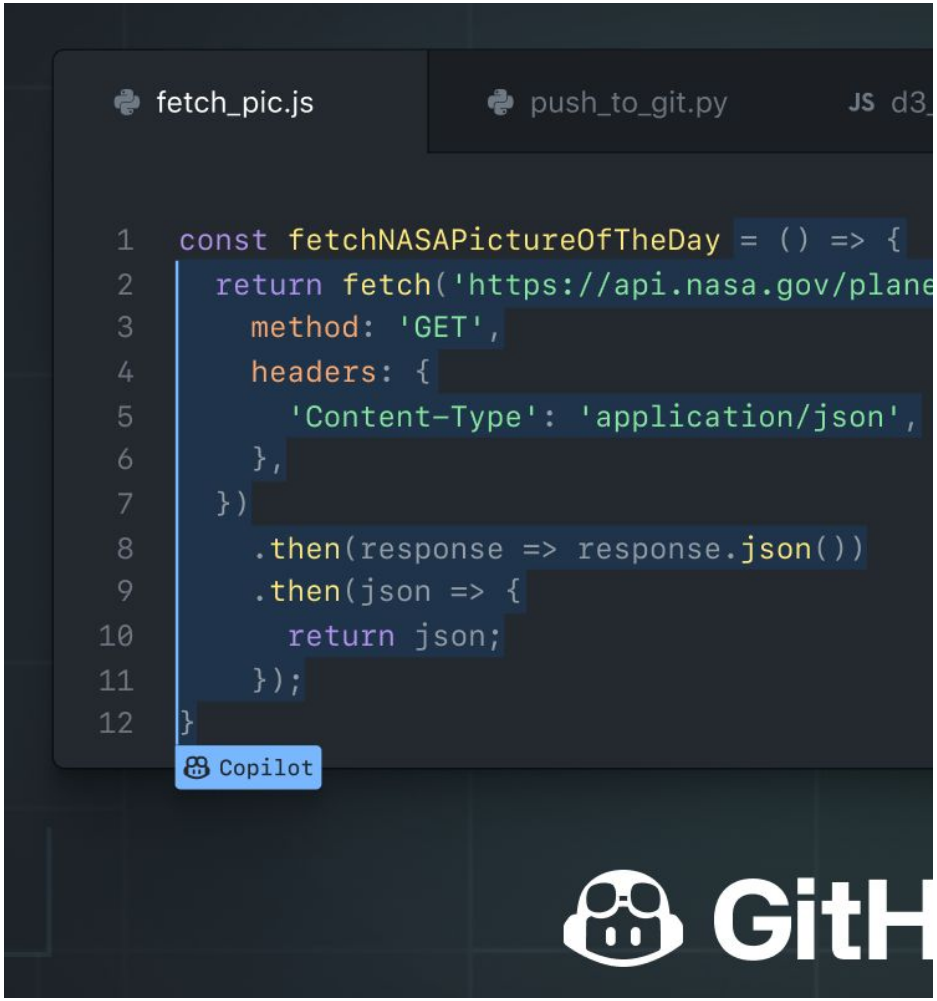
Research Questions

1. How does using Copilot affect the programming experience?
2. How do users recognize errors in code generated by Copilot?
3. What coping mechanisms do users employ when they find errors in code generated by Copilot?
4. What are the obstacles and limitations that can prevent adoption of Copilot?

Methods

- Within-subjects comparative study of students and engineers
- Python programming tasks in VS Code using:
  - Intellisense (default/control)
    - One token
  - Copilot
    - Multiple tokens
    - Prompt (by commenting)


server.js > ...
1 const express = require('express');
2 const app = express();
3
4 var server = express();
5
6 server.
stack
subscribe
toString
trace
unlink
unlock
unsubscribe
use
app
express
require
server


fetch_pic.js
push_to_git.py
JS d3
1 const fetchNASAPictureOfTheDay = () => {
2 return fetch('https://api.nasa.gov/plane
3 method: 'GET',
4 headers: {
5 'Content-Type': 'application/json',
6 },
7 })
8 .then(response => response.json())
9 .then(json => {
10 return json;
11 });
12 }
Copilot
GitH

Experimental Design

- **Tasks**:
  - **Easy:** CSV editing
  - **Medium:** Web scraping
  - **Hard:** Graph plotting
- 24 participants
  - 28 students (undergrad, master's, Ph.D.)
  - 1 software engineer
- One 20-minute "study session" per task
- Binary success/failure (completion)
- After task surveys and final survey

Results

- Successful completion: Copilot < Intellisense
  - Debugging rabbit holes
  - Inexperience with libraries and debugging
  - Done more quickly on average
- **19/24** preferred Copilot
- Useful starting point
  - Saves time/effort from online search
- **Need**: Better ways to understand generated code

Takeaway

To facilitate adoption, tools like Copilot need better explainability and debugging support.