Benchmarking in Continue.dev

TL;DR: A modular, microkernel based system to benchmark LLMs suitability in the context of software development with Continue.dev. Testing features like unified diff generation, apply button functionality and agentic capabilities. Built-in regression tracking, reporting, custom metrics and secure code execution.

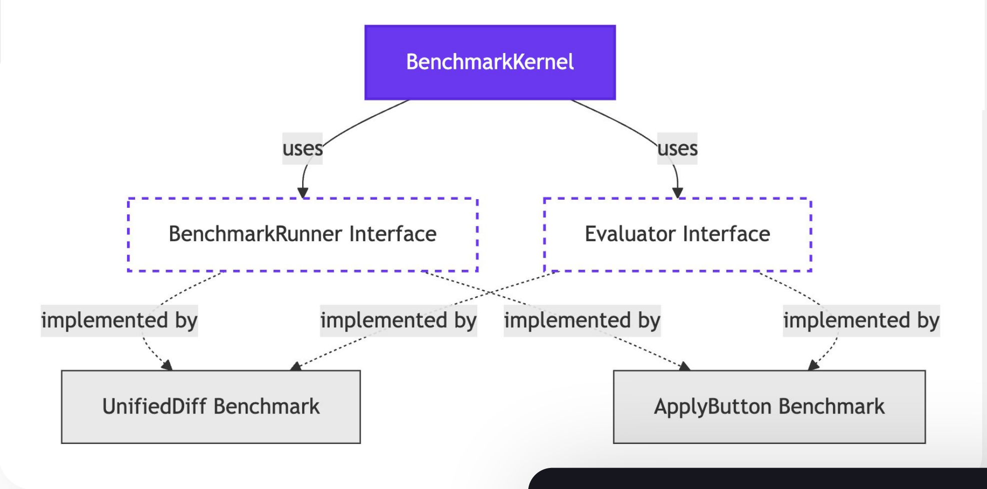
Why Benchmarking?

- Tasks involving LLMs need statistic evaluation over unit tests due to the probabilistic nature [1]

Features?

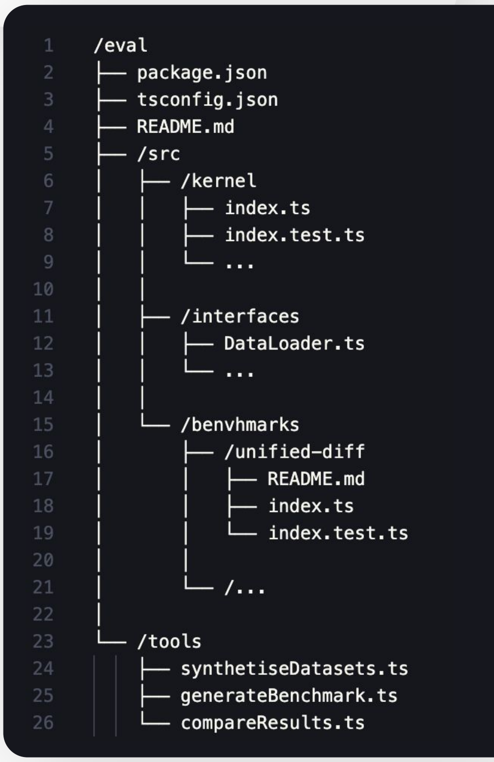
- LLM performance for AI coding tasks
- .continuerules and .promts evaluation
- Test features that involve AI on function, module or system level
- Metrics: latency, cost, accuracy, and custom
- Docker based isolated code execution
- Parallel execution
- Leaderboard
- Terminal and HTML Reports

System Design



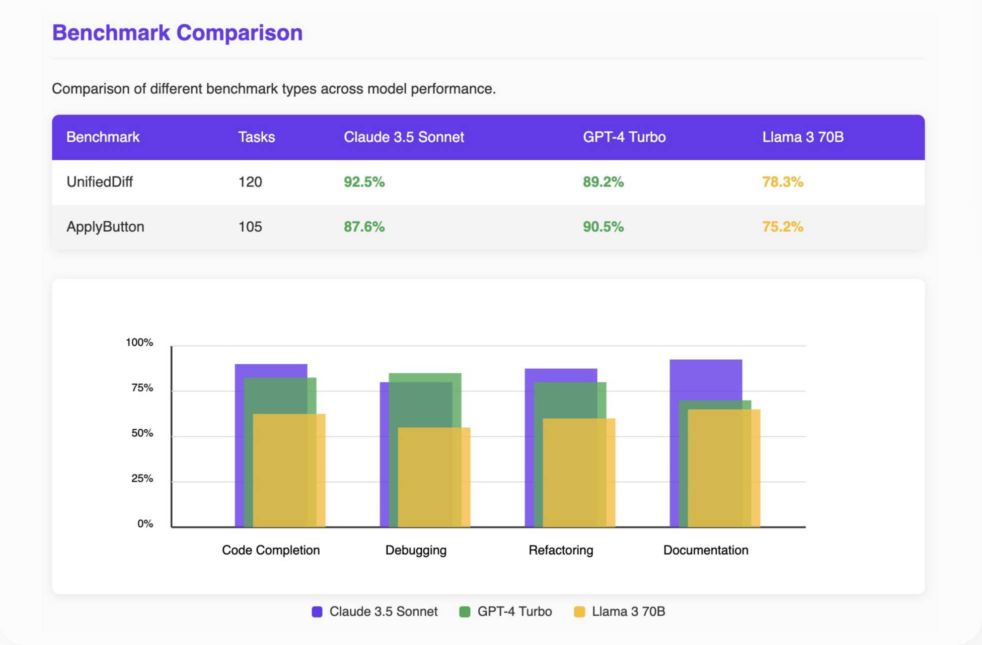
Design Approach

- Microkernel [2]
- Adapter & ports inspired interfaces
- Benchmarks as extendable plugins in ai friendly vertical slices [3]



```
/eval
├── package.json
├── tsconfig.json
├── README.md
├── /src
│   ├── /kernel
│   │   ├── index.ts
│   │   ├── index.test.ts
│   │   └── ...
│   │
│   ├── /interfaces
│   │   ├── DataLoader.ts
│   │   └── ...
│   │
│   └── /benvhmarks
│       ├── /unified-diff
│       │   ├── README.md
│       │   ├── index.ts
│       │   └── index.test.ts
│       │
│       └── /...
│
└── /tools
    ├── synthetiseDatasets.ts
    ├── generateBenchmark.ts
    └── compareResults.ts
```

Report Example



Benchmark Comparison

Comparison of different benchmark types across model performance.

Benchmark	Tasks	Claude 3.5 Sonnet	GPT-4 Turbo	Llama 3 70B
UnifiedDiff	120	92.5%	89.2%	78.3%
ApplyButton	105	87.6%	90.5%	75.2%

Roadmap

- Multimodal
- Dataset and synthetic data generation
- Reinforcement prompt improvement
- A/B Testing
- Human Evaluation Integration

[1] Huyen, C. (2025). *AI Engineering* (p. 160). O'Reilly Media.
[2] Wikipedia. (n.d.). *Microkernel*. from https://en.wikipedia.org/wiki/Microkernel
[3] Bogard, J. (2018). *Vertical slice architecture*. https://www.jimmybogard.com/vertical-slice-architecture/