GitHub: Building Your Coding Profile

Teaching Assistants

Boston University

Mar. 24th, 2025

GitHub

"collaborative, open sharing in the programming world"

Robert Shaw
robertgshaw2-redhat

Unfollow

The Future of AI is Open

129 followers · **82** following

@redhat
Boston
@robertshaw21
in/robert-shaw-1a01399a

Sponsoring

Achievements

x4 x3 x2

Block or Report

Pinned

vllm-project/vllm Public
A high-throughput and memory-efficient inference and serving engine for LLMs
Python 42.5k 6.4k

vllm-project/llm-compressor Public
Transformers-compatible library for applying various compression algorithms to LLMs for optimized deployment with vLLM
Python 1.1k 102

Open-source Projects

1,172 contributions in the last year

Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar
Mon
Wed
Fri

Learn how we count contributions
Less More

2025
2024
2023
2022

Contributions

@vllm-project @neuralmagic @librosa More

Activity overview

Contributed to vllm-project/vllm, neuralmagic/nm-vllm, vllm-project/llm-compressor and 46 other repositories

48% Code review
22% Commits
4% Issues
26% Pull requests

Commits, Code Review, Issues, Pull Request

Contribution activity

March 2025

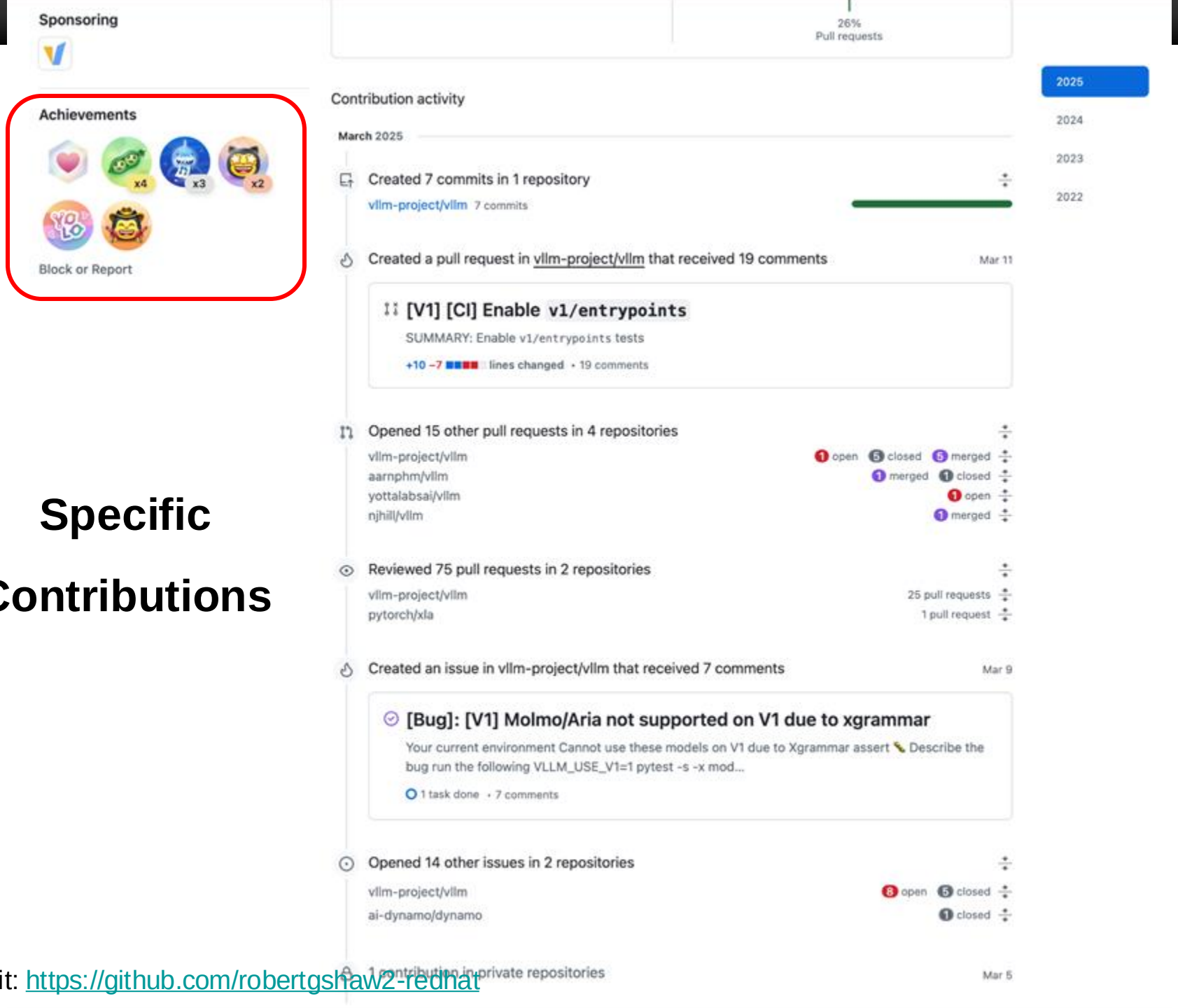
Created 7 commits in 1 repository
vllm-project/vllm 7 commits

Created a pull request in vllm-project/vllm that received 19 comments Mar 11

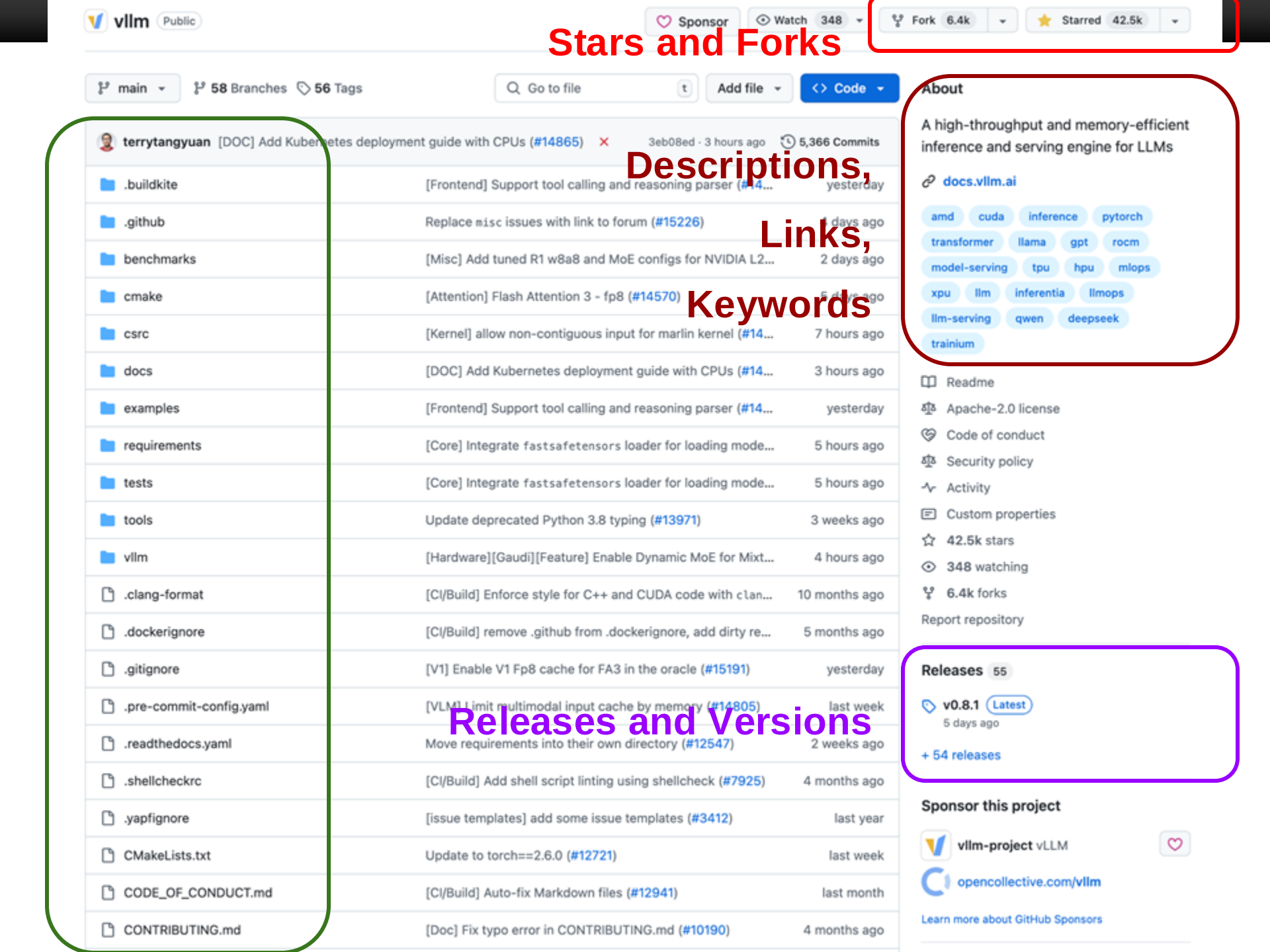
[V1] [CI] Enable `v1/entrypoints`
SUMMARY: Enable `v1/entrypoints` tests

Credit: https://github.com/robertgshaw2-redhat

Specific Contributions



Credit: https://github.com/robertgshaw2-redhat


vllm Public
Sponsor
Watch 348
Fork 6.4k
Starred 42.5k
Stars and Forks
main
58 Branches
56 Tags
Go to file
Add file
Code
terrytangyuan [DOC] Add Kubernetes deployment guide with CPUs (#14865)
3eb08ed · 3 hours ago
5,366 Commits
.buildkite [Frontend] Support tool calling and reasoning parser (#14... yesterday
.github Replace misc issues with link to forum (#15226) 1 days ago
benchmarks [Misc] Add tuned R1 w8a8 and MoE configs for NVIDIA L2... 2 days ago
cmake [Attention] Flash Attention 3 - fp8 (#14570) 5 days ago
csrc [Kernel] allow non-contiguous input for marlin kernel (#14... 7 hours ago
docs [DOC] Add Kubernetes deployment guide with CPUs (#14... 3 hours ago
examples [Frontend] Support tool calling and reasoning parser (#14... yesterday
requirements [Core] Integrate fastsafetensors loader for loading mode... 5 hours ago
tests [Core] Integrate fastsafetensors loader for loading mode... 5 hours ago
tools Update deprecated Python 3.8 typing (#13971) 3 weeks ago
vllm [Hardware][Gaudi][Feature] Enable Dynamic MoE for Mixt... 4 hours ago
.clang-format [CI/Build] Enforce style for C++ and CUDA code with clan... 10 months ago
.dockerignore [CI/Build] remove .github from .dockerignore, add dirty re... 5 months ago
.gitignore [V1] Enable V1 Fp8 cache for FA3 in the oracle (#15191) yesterday
.pre-commit-config.yaml [VLM] Limit multimodal input cache by memory (#14805) last week
.readthedocs.yaml Move requirements into their own directory (#12547) 2 weeks ago
.shellcheckrc [CI/Build] Add shell script linting using shellcheck (#7925) 4 months ago
.yapfignore [issue templates] add some issue templates (#3412) last year
CMakeLists.txt Update to torch==2.6.0 (#12721) last week
CODE_OF_CONDUCT.md [CI/Build] Auto-fix Markdown files (#12941) last month
CONTRIBUTING.md [Doc] Fix typo error in CONTRIBUTING.md (#10190) 4 months ago
Descriptions,
Links,
Keywords
About
A high-throughput and memory-efficient inference and serving engine for LLMs
docs.vllm.ai
amd cuda inference pytorch
transformer llama gpt rocm
model-serving tpu hpu mlops
xpu llm inferentia llmops
llm-serving qwen deepseek
trainium
Readme
Apache-2.0 license
Code of conduct
Security policy
Activity
Custom properties
42.5k stars
348 watching
6.4k forks
Report repository
Releases 55
v0.8.1 Latest
5 days ago
+ 54 releases
Releases and Versions
Sponsor this project
vllm-project vLLM
opencollective.com/vllm
Learn more about GitHub Sponsors

README | Code of conduct | Apache-2.0 license | Security

Easy, fast, and cheap LLM serving for everyone

| Documentation | Blog | Paper | Twitter/X | User Forum | Developer Slack |

[2025/03] We are collaborating with Ollama to host an Inference Night at Y Combinator in San Francisco on Thursday, March 27, at 6 PM. Discuss all things inference local or data center!

[2025/04] We're hosting our first-ever *vLLM Asia Developer Day* in Singapore on *April 3rd*! This is a full-day event (9 AM - 9 PM SGT) in partnership with SGInnovate, AMD, and Embedded LLM. Meet the vLLM team and learn about LLM inference for RL, MI300X, and more! Register Now

---

Latest News 🔥

- [2025/03] We hosted the first vLLM China Meetup! Please find the meetup slides from vLLM team here.
- [2025/03] We hosted the East Coast vLLM Meetup! Please find the meetup slides here.
- [2025/02] We hosted the ninth vLLM meetup with Meta! Please find the meetup slides from vLLM team here and AMD here. The slides from Meta will not be posted.
- [2025/01] We are excited to announce the alpha release of vLLM V1: A major architectural upgrade with 1.7x speedup! Clean code, optimized execution loop, zero-overhead prefix caching, enhanced multimodal support, and more. Please check out our blog post here.
- [2025/01] We hosted the eighth vLLM meetup with Google Cloud! Please find the meetup slides from vLLM team here, and Google Cloud team here.
- [2024/12] vLLM joins pytorch ecosystem! Easy, Fast, and Cheap LLM Serving for Everyone!

▶ Previous News

About

vLLM is a fast and easy-to-use library for LLM inference and serving.

Originally developed in the Sky Computing Lab at UC Berkeley, vLLM has evolved into a community-driven project with contributions from both academia and industry.

vLLM is fast with:

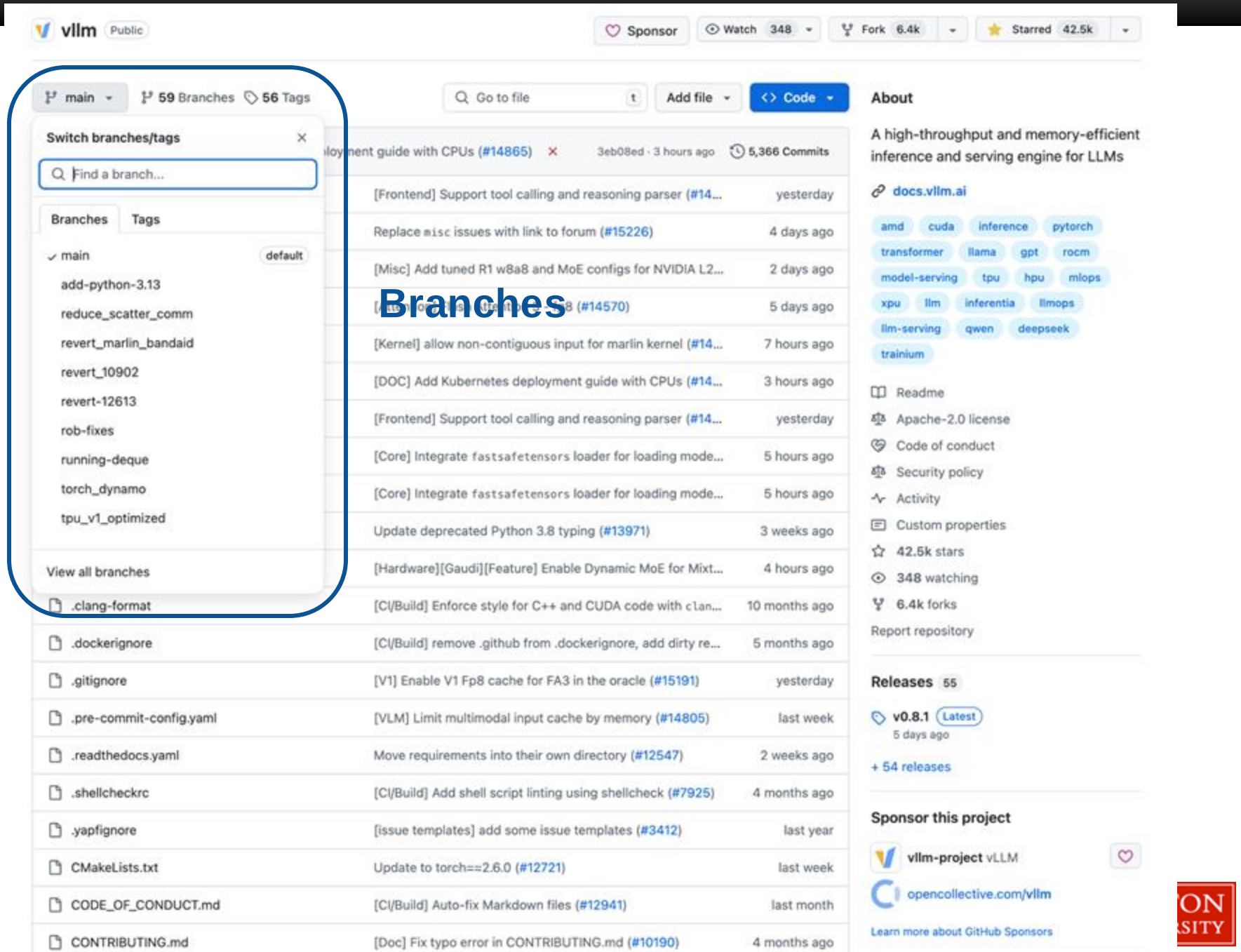
README.MD

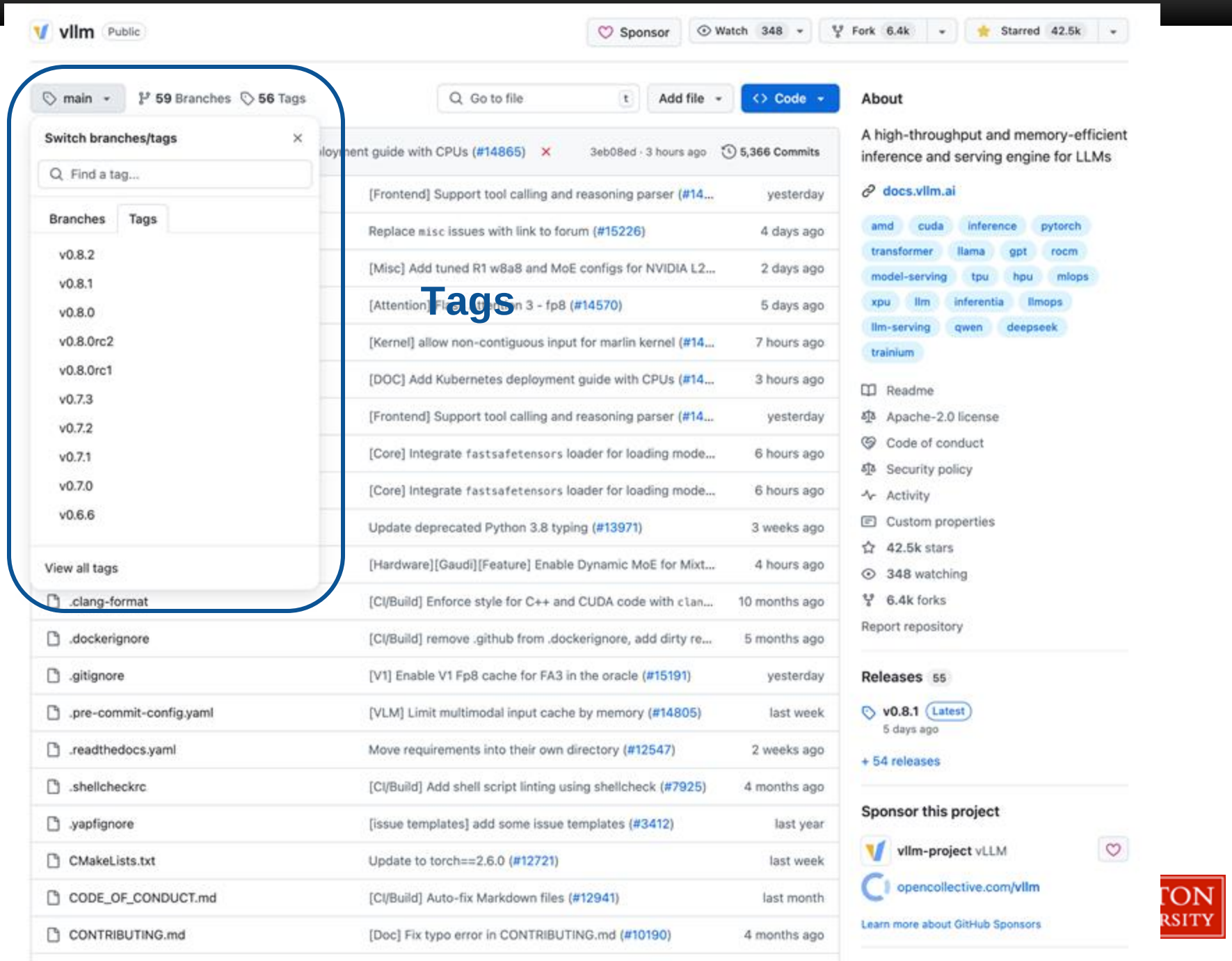
Markdown Syntax

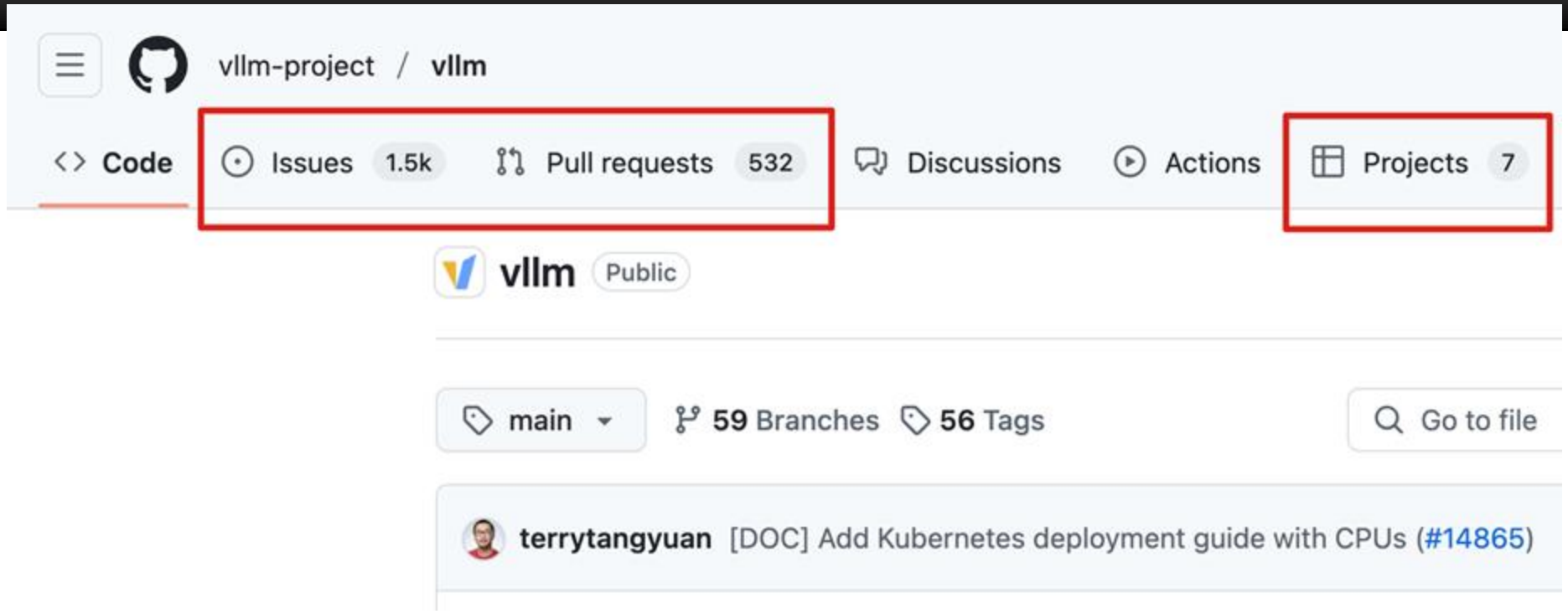
g

BOSTON UNIVERSITY

Markdown: https://www.markdownguide.org/basic-syntax/

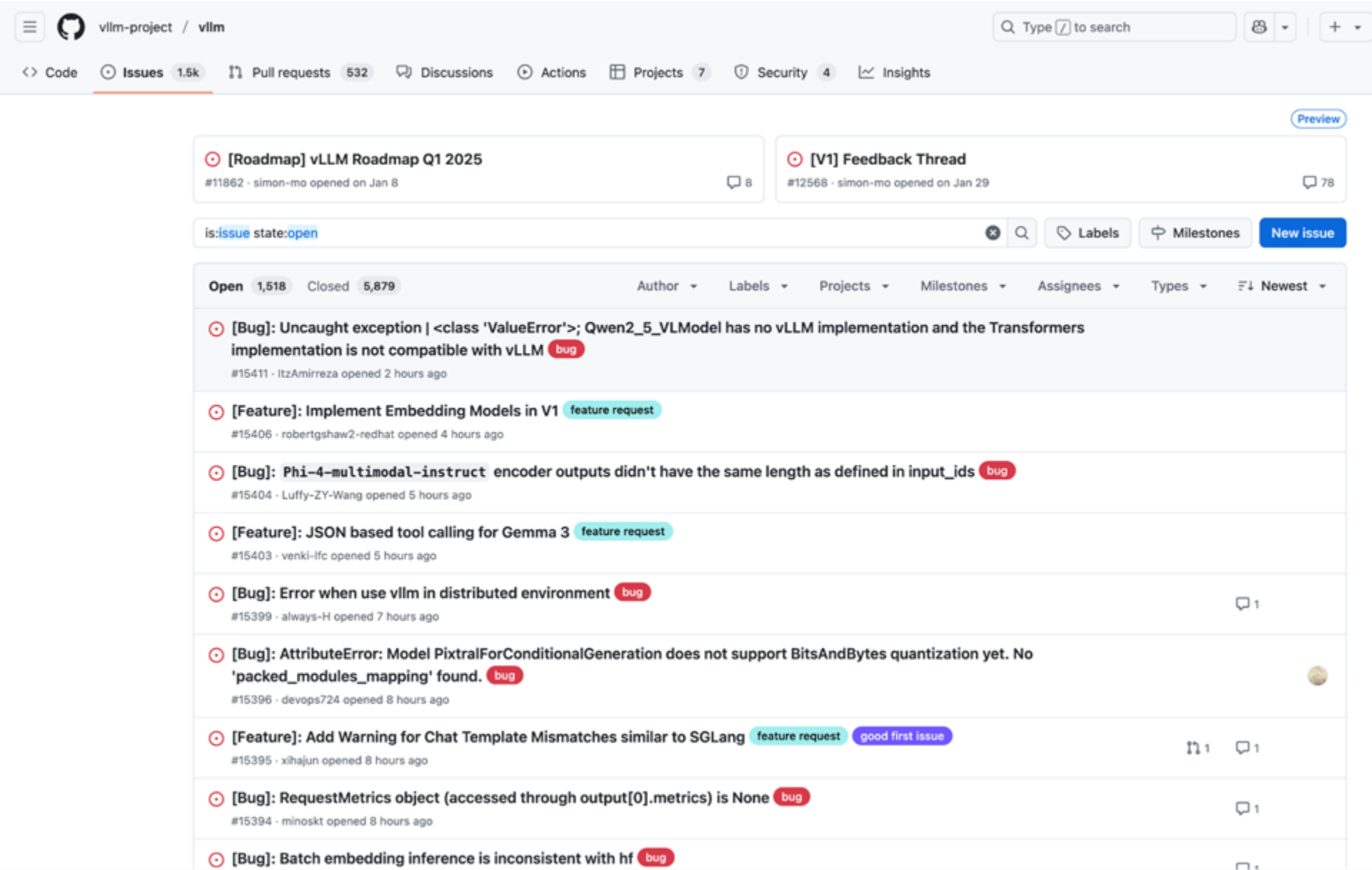

vllm Public
Sponsor
Watch 348
Fork 6.4k
Starred 42.5k
main
59 Branches
56 Tags
Go to file
Add file
Code
Switch branches/tags
Find a branch...
Branches
Tags
main default
add-python-3.13
reduce_scatter_comm
revert_marlin_bandaid
revert_10902
revert-12613
rob-fixes
running-deque
torch_dynamo
tpu_v1_optimized
View all branches
Branches
ployment guide with CPUs (#14865)
3eb08ed · 3 hours ago
5,366 Commits
[Frontend] Support tool calling and reasoning parser (#14... yesterday
Replace misc issues with link to forum (#15226) 4 days ago
[Misc] Add tuned R1 w8a8 and MoE configs for NVIDIA L2... 2 days ago
5 days ago
[Kernel] allow non-contiguous input for marlin kernel (#14... 7 hours ago
[DOC] Add Kubernetes deployment guide with CPUs (#14... 3 hours ago
[Frontend] Support tool calling and reasoning parser (#14... yesterday
[Core] Integrate fastsafetensors loader for loading mode... 5 hours ago
[Core] Integrate fastsafetensors loader for loading mode... 5 hours ago
Update deprecated Python 3.8 typing (#13971) 3 weeks ago
[Hardware][Gaudi][Feature] Enable Dynamic MoE for Mixt... 4 hours ago
.clang-format [CI/Build] Enforce style for C++ and CUDA code with clan... 10 months ago
.dockerignore [CI/Build] remove .github from .dockerignore, add dirty re... 5 months ago
.gitignore [V1] Enable V1 Fp8 cache for FA3 in the oracle (#15191) yesterday
.pre-commit-config.yaml [VLM] Limit multimodal input cache by memory (#14805) last week
.readthedocs.yaml Move requirements into their own directory (#12547) 2 weeks ago
.shellcheckrc [CI/Build] Add shell script linting using shellcheck (#7925) 4 months ago
.yapfignore [issue templates] add some issue templates (#3412) last year
CMakeLists.txt Update to torch==2.6.0 (#12721) last week
CODE_OF_CONDUCT.md [CI/Build] Auto-fix Markdown files (#12941) last month
CONTRIBUTING.md [Doc] Fix typo error in CONTRIBUTING.md (#10190) 4 months ago
About
A high-throughput and memory-efficient inference and serving engine for LLMs
docs.vllm.ai
amd cuda inference pytorch transformer llama gpt rocm model-serving tpu hpu mlops xpu llm inferentia llmops llm-serving qwen deepseek trainium
Readme
Apache-2.0 license
Code of conduct
Security policy
Activity
Custom properties
42.5k stars
348 watching
6.4k forks
Report repository
Releases 55
v0.8.1 Latest
5 days ago
+ 54 releases
Sponsor this project
vllm-project vLLM
opencollective.com/vllm
Learn more about GitHub Sponsors


vllm Public
Sponsor
Watch 348
Fork 6.4k
Starred 42.5k
main
59 Branches
56 Tags
Go to file
Add file
Code
Switch branches/tags
Find a tag...
Branches
Tags
v0.8.2
v0.8.1
v0.8.0
v0.8.0rc2
v0.8.0rc1
v0.7.3
v0.7.2
v0.7.1
v0.7.0
v0.6.6
View all tags
Tags
3eb08ed · 3 hours ago
5,366 Commits
.clang-format
.dockerignore
.gitignore
.pre-commit-config.yaml
.readthedocs.yaml
.shellcheckrc
.yapfignore
CMakeLists.txt
CODE_OF_CONDUCT.md
CONTRIBUTING.md
About
A high-throughput and memory-efficient inference and serving engine for LLMs
docs.vllm.ai
amd
cuda
inference
pytorch
transformer
llama
gpt
rocm
model-serving
tpu
hpu
mlops
xpu
llm
inferentia
llmops
llm-serving
qwen
deepseek
trainium
Readme
Apache-2.0 license
Code of conduct
Security policy
Activity
Custom properties
42.5k stars
348 watching
6.4k forks
Report repository
Releases 55
v0.8.1 Latest
5 days ago
+ 54 releases
Sponsor this project
vllm-project vLLM
opencollective.com/vllm
Learn more about GitHub Sponsors

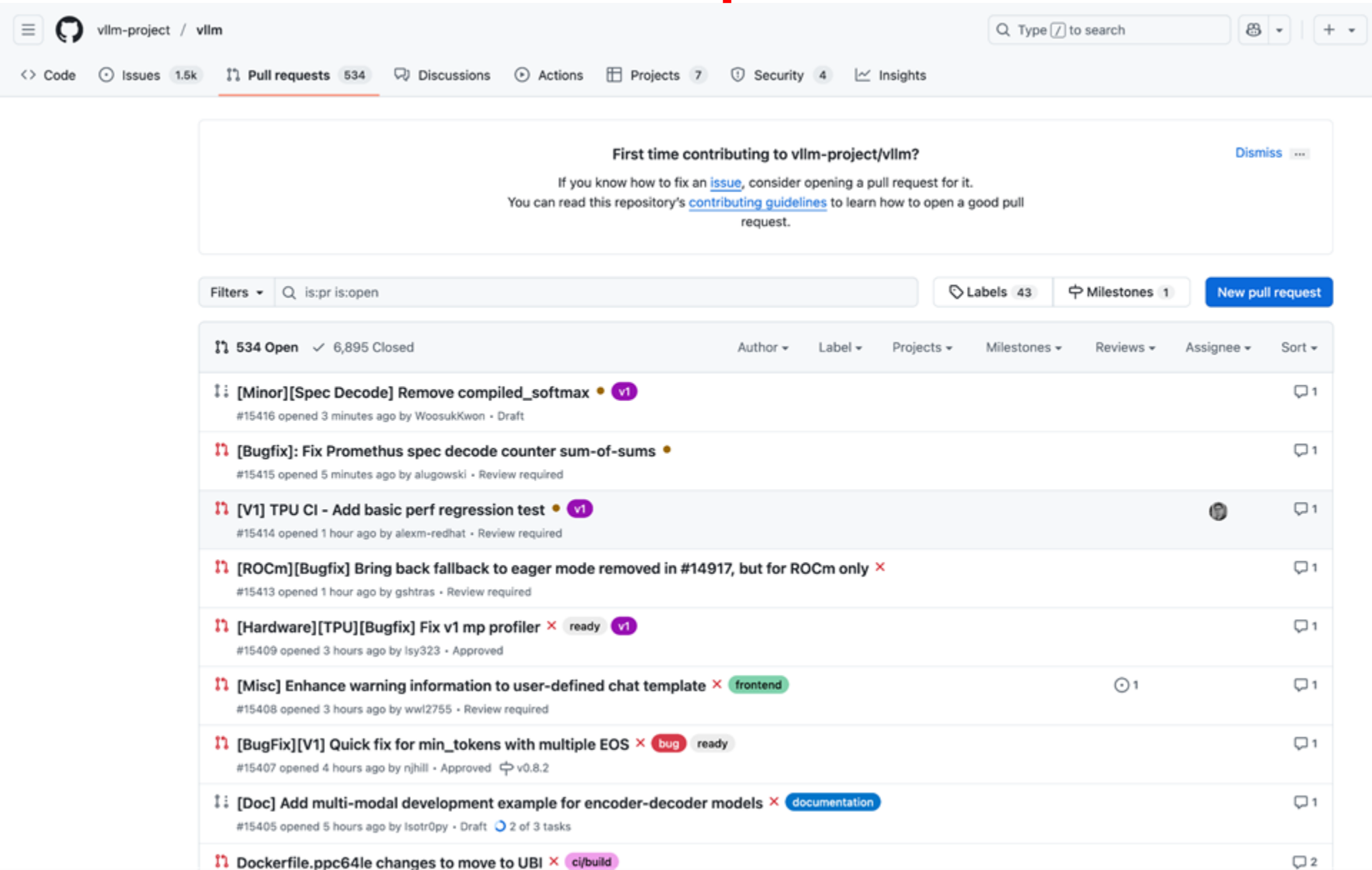


Issues, Pull Requests, Projects

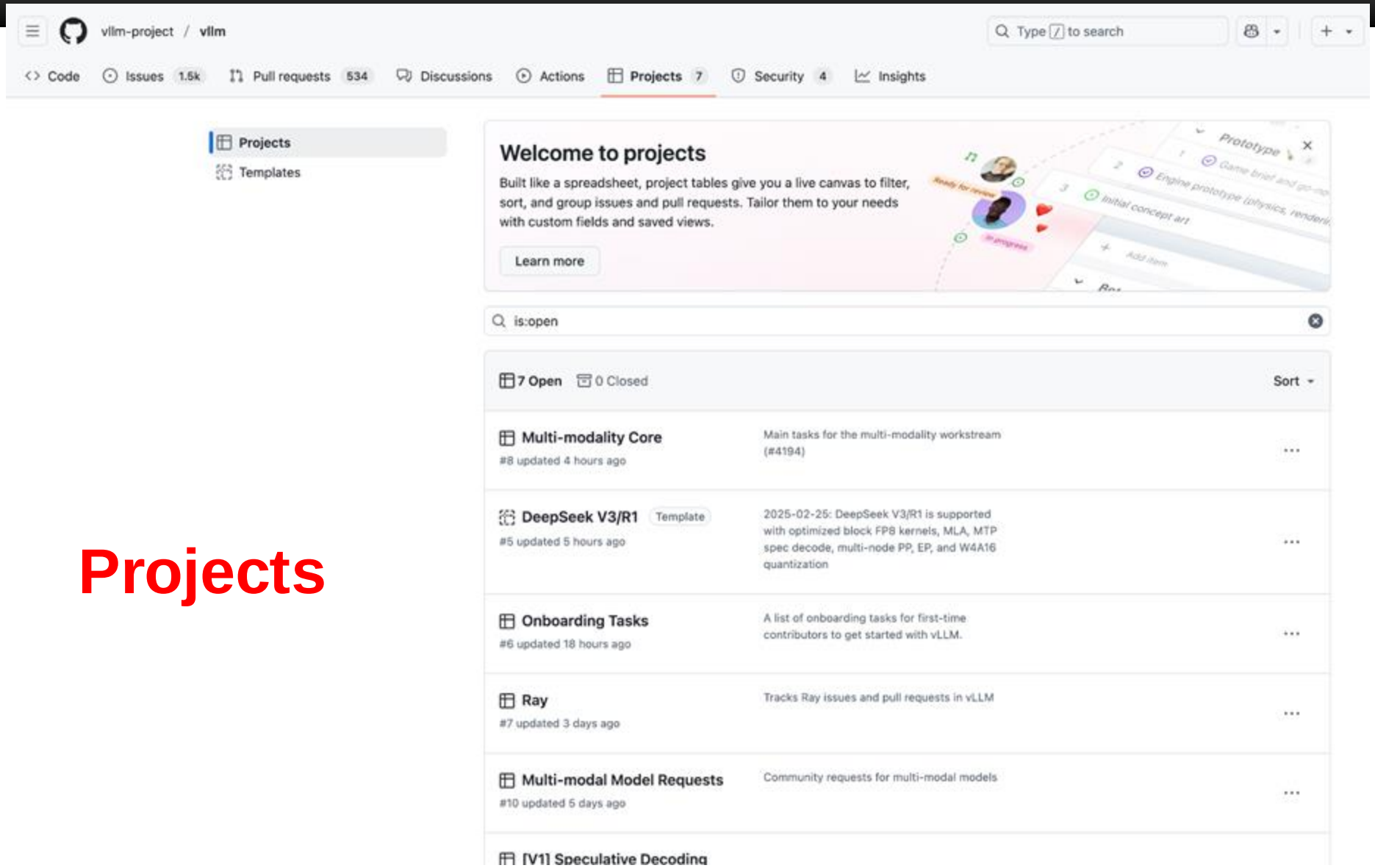
Issues


vllm-project / vllm
Type / to search
Code
Issues 1.5k
Pull requests 532
Discussions
Actions
Projects 7
Security 4
Insights
Preview
[Roadmap] vLLM Roadmap Q1 2025
#11862 · simon-mo opened on Jan 8
8
[V1] Feedback Thread
#12568 · simon-mo opened on Jan 29
78
is:issue state:open
Labels
Milestones
New issue
Open 1,518
Closed 5,879
Author
Labels
Projects
Milestones
Assignees
Types
Newest
[Bug]: Uncaught exception | <class 'ValueError'>; Qwen2_5_VLModel has no vLLM implementation and the Transformers implementation is not compatible with vLLM
bug
#15411 · ItzAmirreza opened 2 hours ago
[Feature]: Implement Embedding Models in V1
feature request
#15406 · robertgshaw2-redhat opened 4 hours ago
[Bug]: Phi-4-multimodal-instruct encoder outputs didn't have the same length as defined in input_ids
bug
#15404 · Luffy-ZY-Wang opened 5 hours ago
[Feature]: JSON based tool calling for Gemma 3
feature request
#15403 · venki-lfc opened 5 hours ago
[Bug]: Error when use vllm in distributed environment
bug
#15399 · always-H opened 7 hours ago
1
[Bug]: AttributeError: Model PixtralForConditionalGeneration does not support BitsAndBytes quantization yet. No 'packed_modules_mapping' found.
bug
#15396 · devops724 opened 8 hours ago
[Feature]: Add Warning for Chat Template Mismatches similar to SGLang
feature request
good first issue
#15395 · xihajun opened 8 hours ago
1
1
[Bug]: RequestMetrics object (accessed through output[0].metrics) is None
bug
#15394 · minoskt opened 8 hours ago
1
[Bug]: Batch embedding inference is inconsistent with hf
bug

Pull request



Projects


vllm-project / vllm
Type / to search
Code
Issues 1.5k
Pull requests 534
Discussions
Actions
Projects 7
Security 4
Insights
Projects
Templates
Welcome to projects
Built like a spreadsheet, project tables give you a live canvas to filter, sort, and group issues and pull requests. Tailor them to your needs with custom fields and saved views.
Learn more
is:open
7 Open
0 Closed
Sort
Multi-modality Core
#8 updated 4 hours ago
Main tasks for the multi-modality workstream (#4194)
DeepSeek V3/R1
Template
#5 updated 5 hours ago
2025-02-25: DeepSeek V3/R1 is supported with optimized block FP8 kernels, MLA, MTP spec decode, multi-node PP, EP, and W4A16 quantization
Onboarding Tasks
#6 updated 18 hours ago
A list of onboarding tasks for first-time contributors to get started with vLLM.
Ray
#7 updated 3 days ago
Tracks Ray issues and pull requests in vLLM
Multi-modal Model Requests
#10 updated 5 days ago
Community requests for multi-modal models
[V1] Speculative Decoding

vllm-project / Projects / DeepSeek V3/R1

DeepSeek V3/R1

Increased items preview | Feedback | Public template

2025-02-25: DeepSeek V3/R1 is supported with optimized block FP8 kernels, MLA, MTP spec decode, multi-node PP, EP, and W4A16 quantization

Status board | My items

Filter by keyword or by field

Roadmap/Status board

Backlog 48

This item hasn't been started

vllm #14939
[Bug]: AssertionError with Speculative Decoding in vLLM Using DeepSeek R1 Distill Qwen Models
bug

vllm #12860
[Bug]: Can not load DeepSeek-R1-Distill-Llama-70B with VLLM
bug

vllm #14266
[Bug]: weight_loader of fp8 weights are wrongly set to None. [Deepseek V3/R1]
bug

vllm #15044
[Bug]: CPU infrencing won't work for DeepSeek-R1
bug

vllm #15333
[Bug]: Can't deserialize object: ObjectRef, DeepSeek R1, H20*16, pp2, tp8, v1 engine
bug

In progress 3

This is actively being worked on

vllm #13282
[Bug]: DeepSeek-R1-Distill-Llama-70B max_model_len can not larger than 8192
bug

vllm #13074
[Usage]: Context Size Limitation and CUDA OOM with DeepSeek R1 on 2 Nodes (TP8 PP2, 16 GPUs with 141GB VRAM Each)
usage

vllm #13630
[WIP][Kernel] Flashinfer MLA support
ci/build | documentation | v1

Done 48

This has been completed

vllm #13833
DeepSeek V2/V3/R1 only place `lm_head` on last pp rank
ready

vllm #12597
Implement MLA for deepseek v3/r1

vllm #13839
Improve pipeline partitioning
ready

vllm #14028
Using vllm to deploy DeepSeek- R1-distil-qwen-32B-q4_k_M. gguf does not have <think> when output, how to solve this problem
usage

vllm #13867
[Attention] Flash MLA for V1
ci/build | ready | v1

vllm #13789
[Attention] MLA support for V1

Outline

- Git & GitHub
- Basic Git Commands
- Pull Requests & Code Review
- Branches and Merging
- Actions & Automating Tests
- README Examples

Part 1 - Git and GitHub

- **Git**
  - **Developer:** Linus Torvalds
  - Initially for Linux projects
  - Extend to collaborative & open-source projects
  - **Version control** system
  - Tracks changes in files
  - **Multiple branches** support
  - **Multiple user** support
  - **Codes merging**
  - In 2023, 93% developers use GitHub

Download: https://git-scm.com/downloads

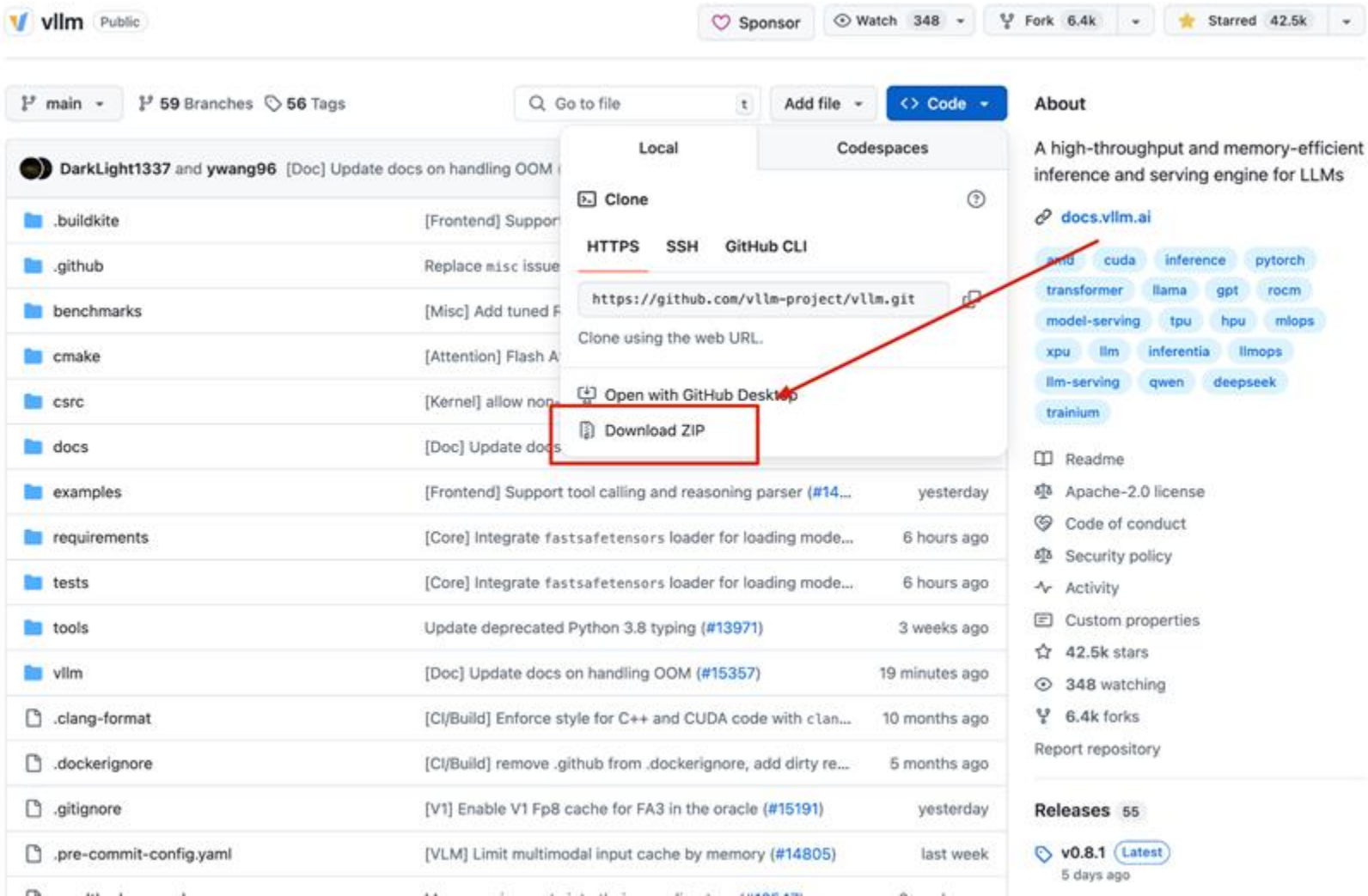
Credit: https://www.britannica.com/technology/GitHub

Part 1 - Git and GitHub

- **GitHub**
  - **Cloud-based platform**
  - **GitHub copilot**
  - **create, store, manage, and share codes**
  - **Create and store** your code in a repository
  - **Share** your work
  - Let others **review** your code and make suggestions
  - **Collaborate** on a shared project

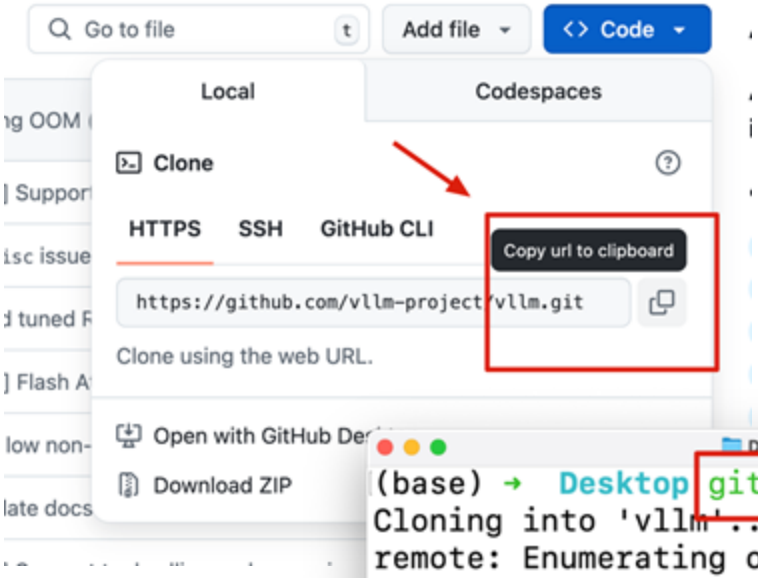
Part 2 - Basic Git Commands

- Download Codes (manually)



Part 2 - Basic Git Commands

- **Download Codes**
  - Copy the link of this repo
  - `git clone https://github.com/vllm-project/vllm.git`

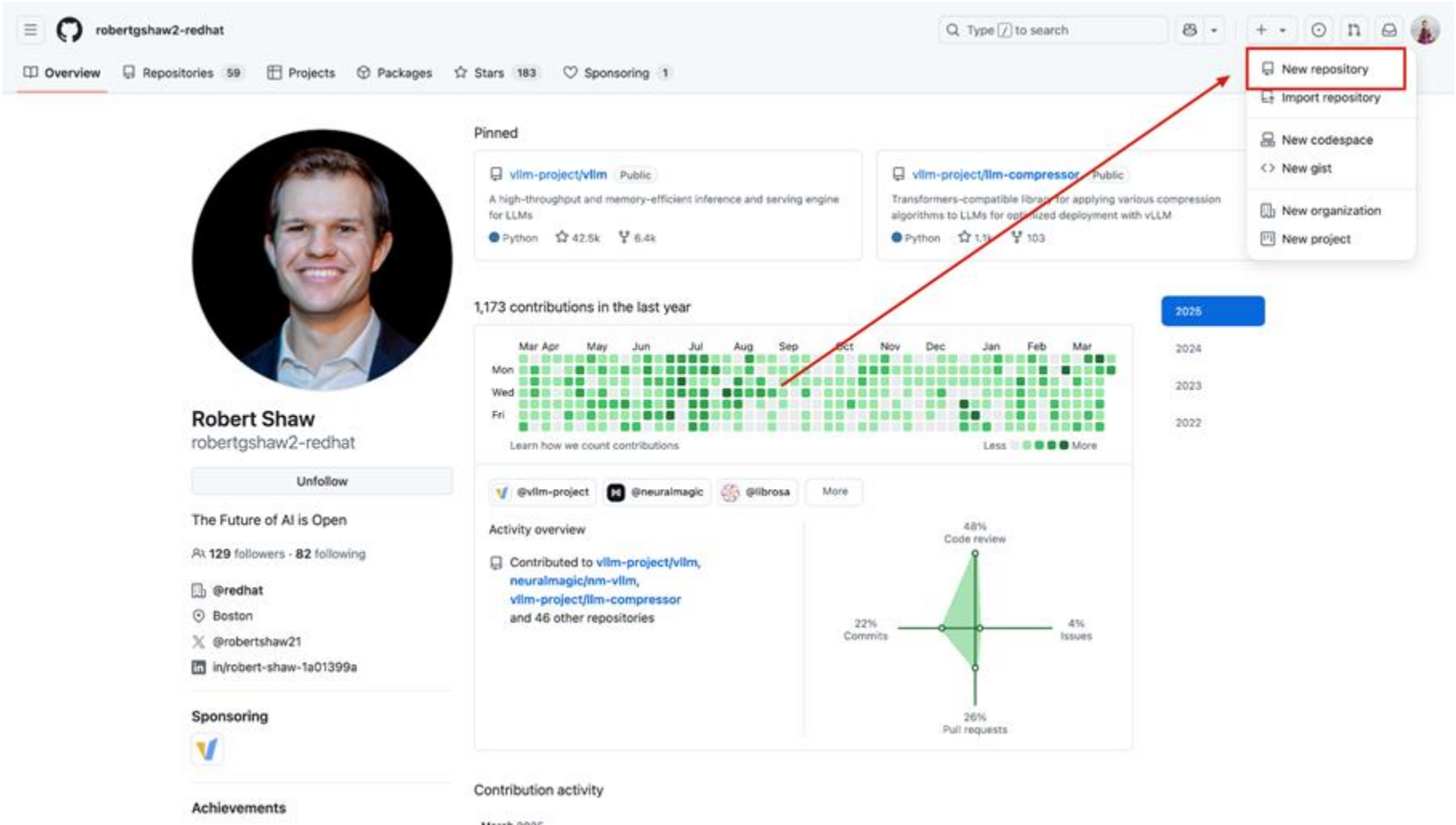


Step 1: copy the link

Step 2: `git clone link`

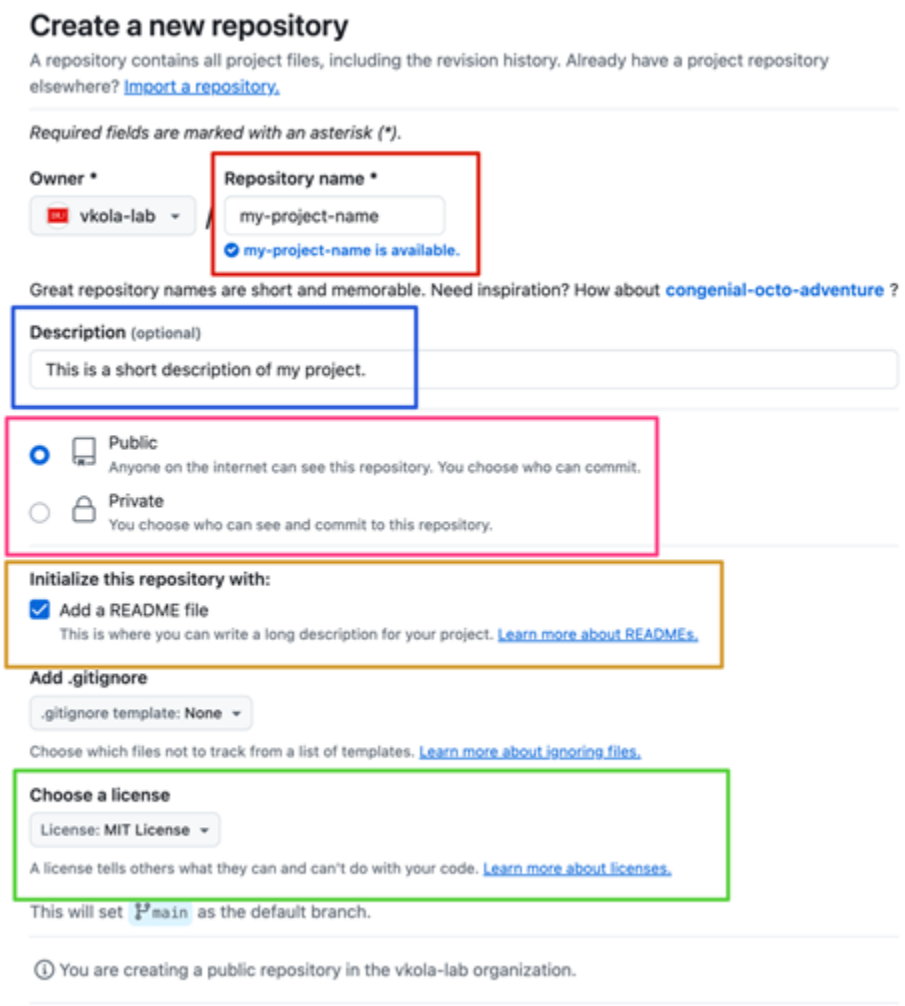
Part 2 - Basic Git Commands

- **Upload Codes → First create a repo on your profile**



Part 2 - Basic Git Commands

- **Upload Codes**



1: Repo name

2: Description

3: Public/Private

4: README.MD

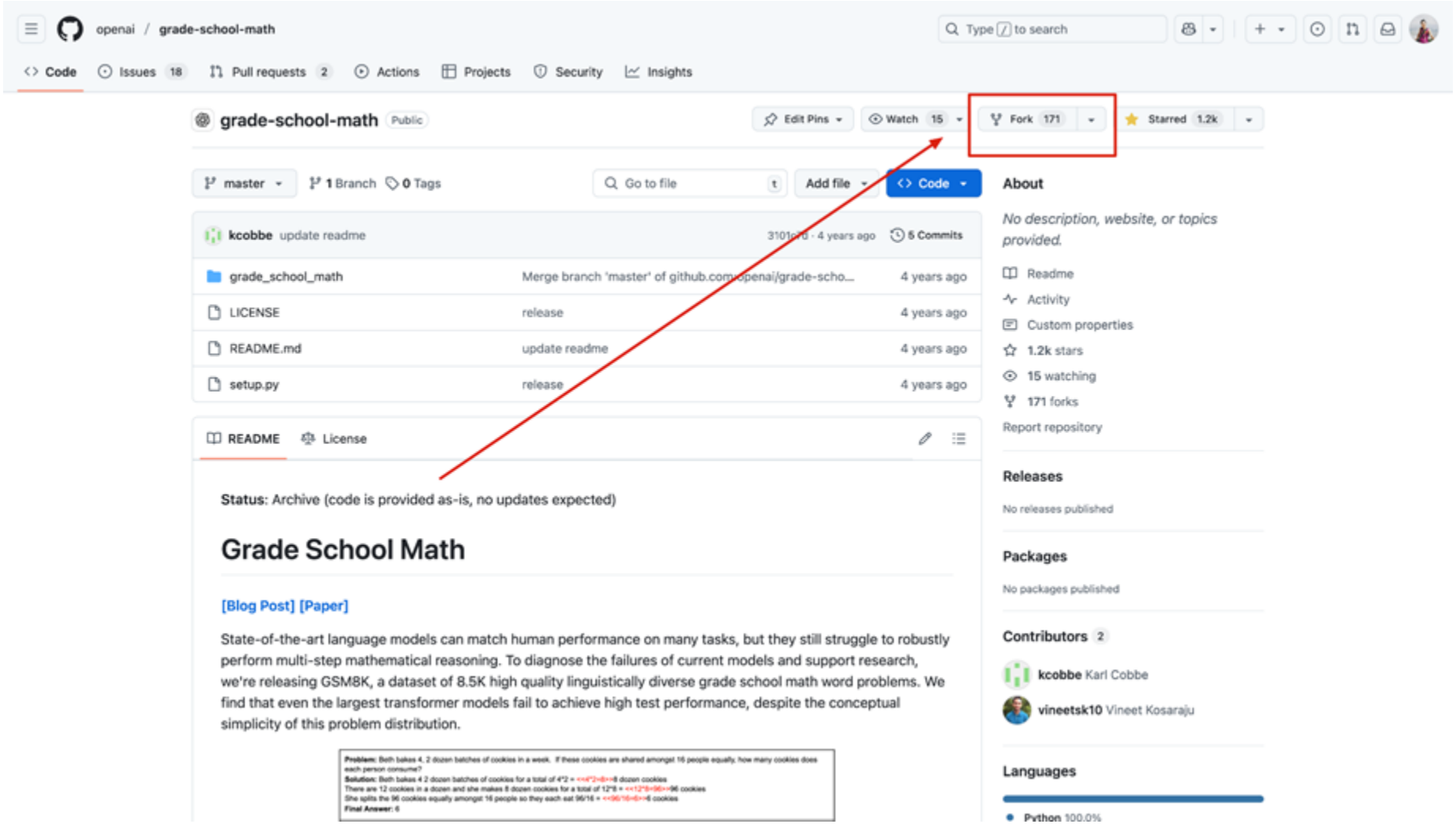
5: License

Part 2 - Basic Git Commands

- **Commit and Upload Codes**
- **First way:** drag and drop the code files
- **Second way:** terminal
  - **Add the codes** in the current folder to git
    - `git add .`
  - **Commit the codes**
    - `git commit -m "Update my lovely codes!"`
  - **Push the codes**
    - `git push`

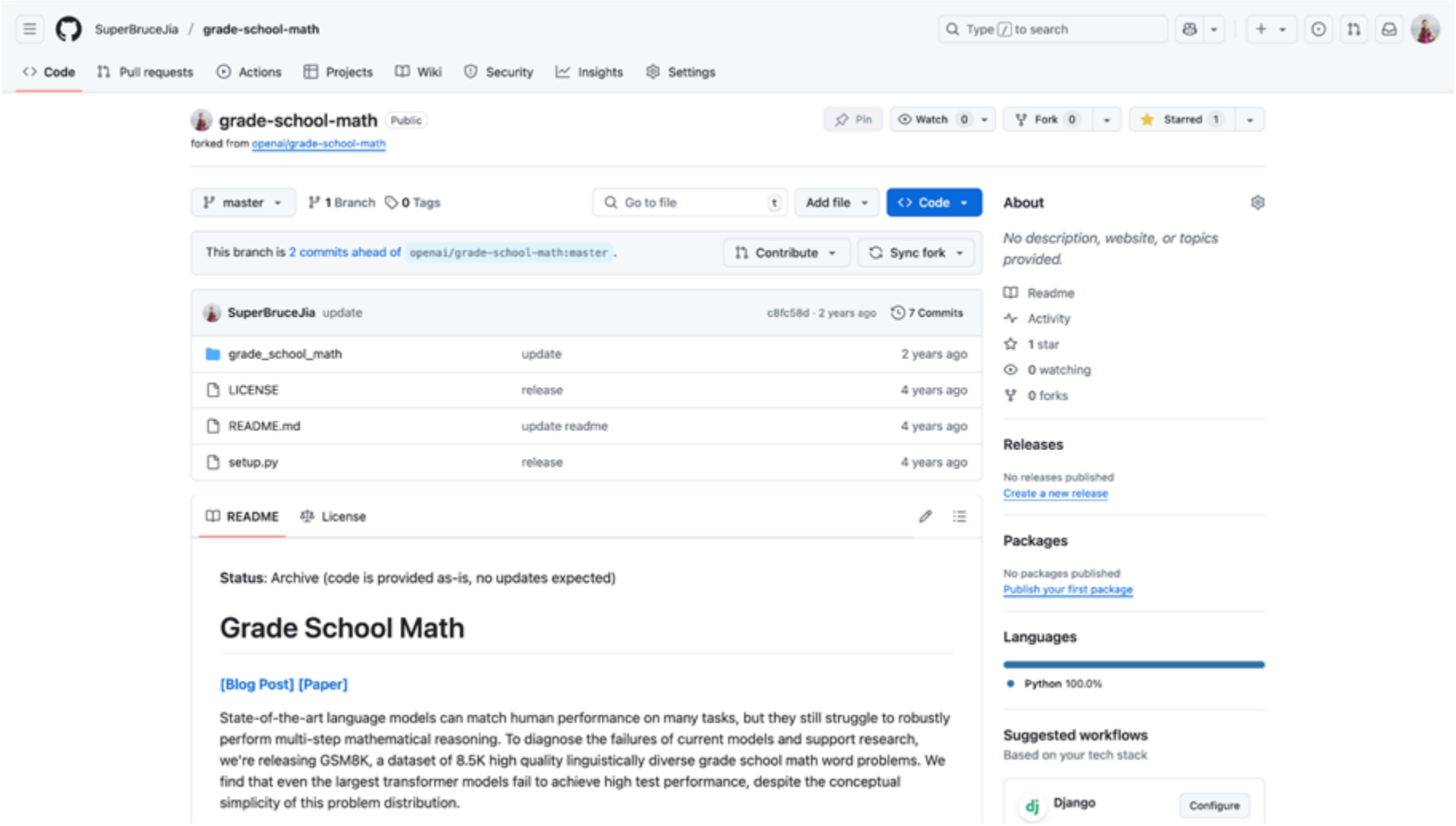
Part 3 - Pull Requests & Code Review

- **Fork the Codes if you haven't made the changes**



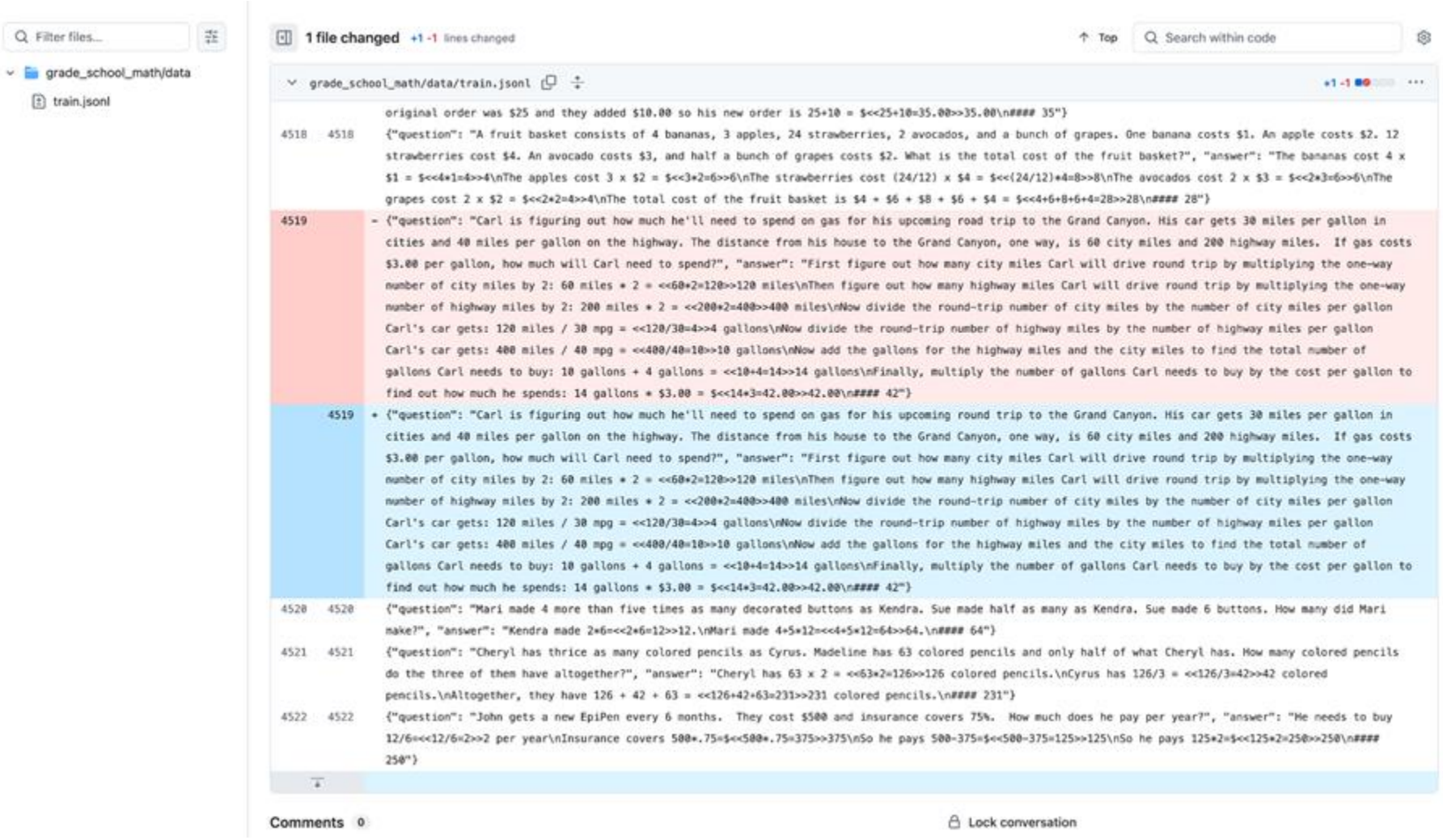
Part 3 - Pull Requests & Code Review

- **Fork the Codes if you haven't made the changes**



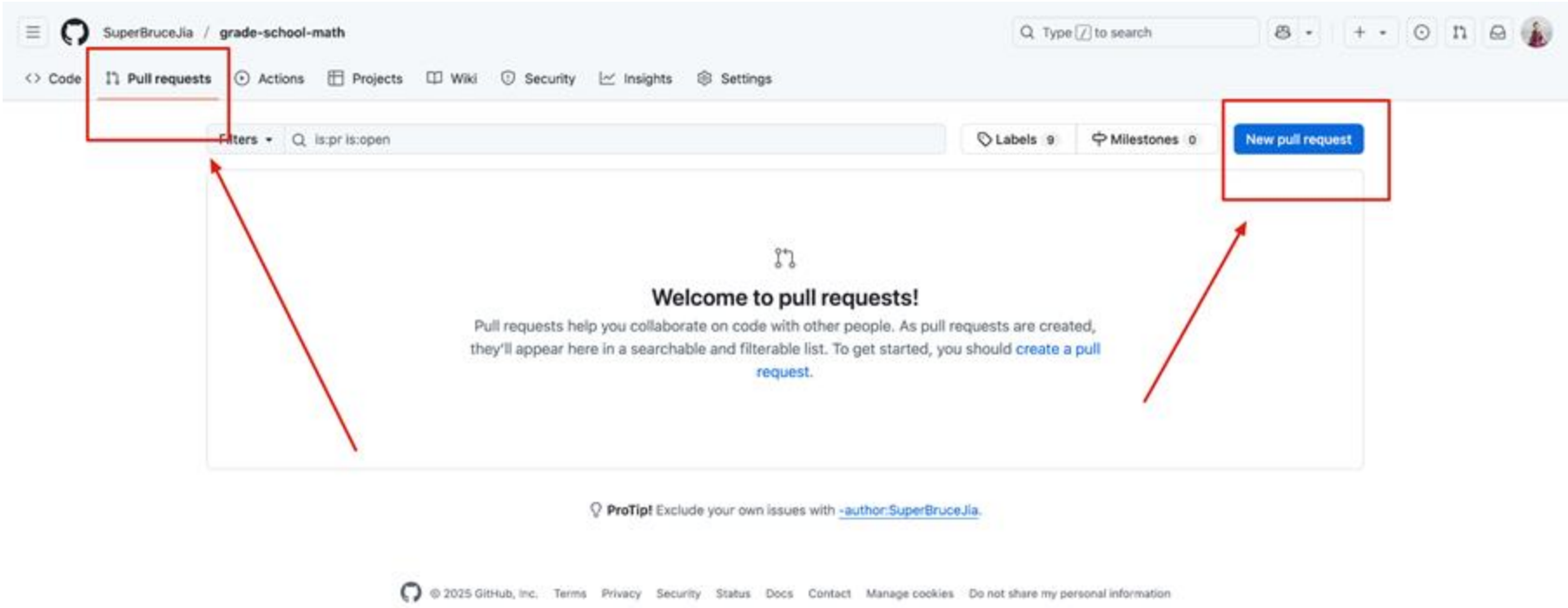
Part 3 - Pull Requests & Code Review

- **Make any changes to the codes → Commit**



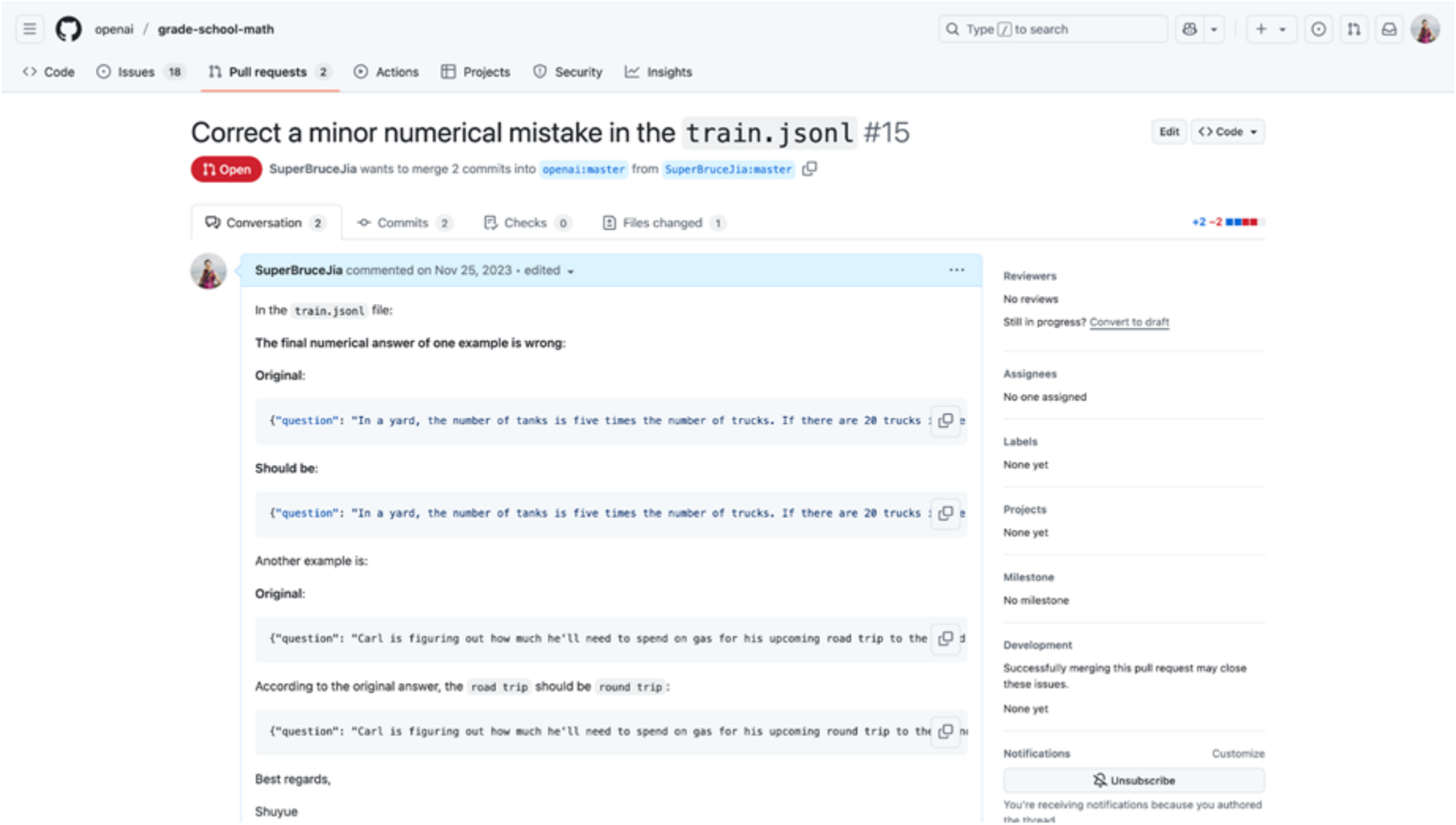
Part 3 - Pull Requests & Code Review

- **New pull request**



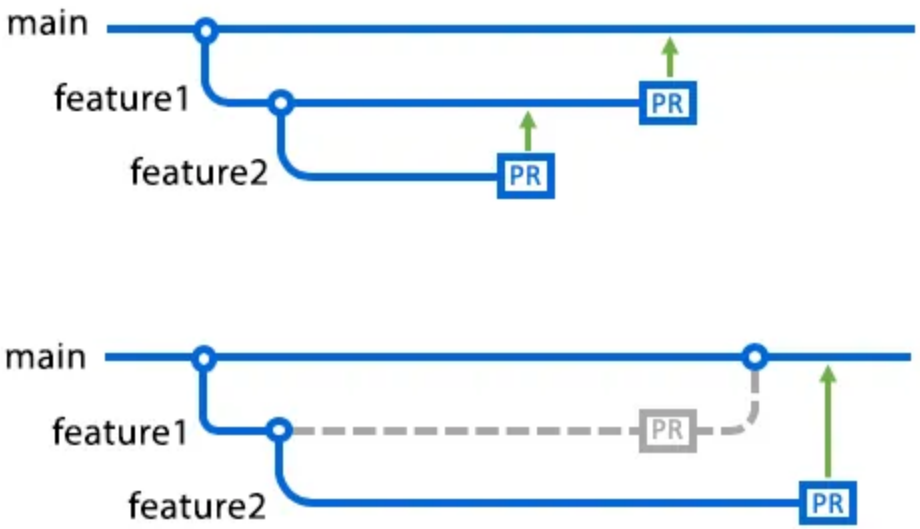
Part 3 - Pull Requests & Code Review

- **New pull request → Code review & Comment → Merge**



Part 4 - Branches & Merging

Potential conflict issue



1: `main` branch

2: `feature 1` branch

3: `feature 2` branch

Credit: https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/about-branches

Part 4 - Branches & Merging

- Create a new branch and Upload Codes
- Create a new branch called `new-feature`
  - `git checkout -b new-feature`
- Add changes to your new branch
  - `git add .`
- Commit your changes
  - `git commit -m "Change method in lovely file."`
- Push your branch to a remote
  - `git push -u origin new-feature`

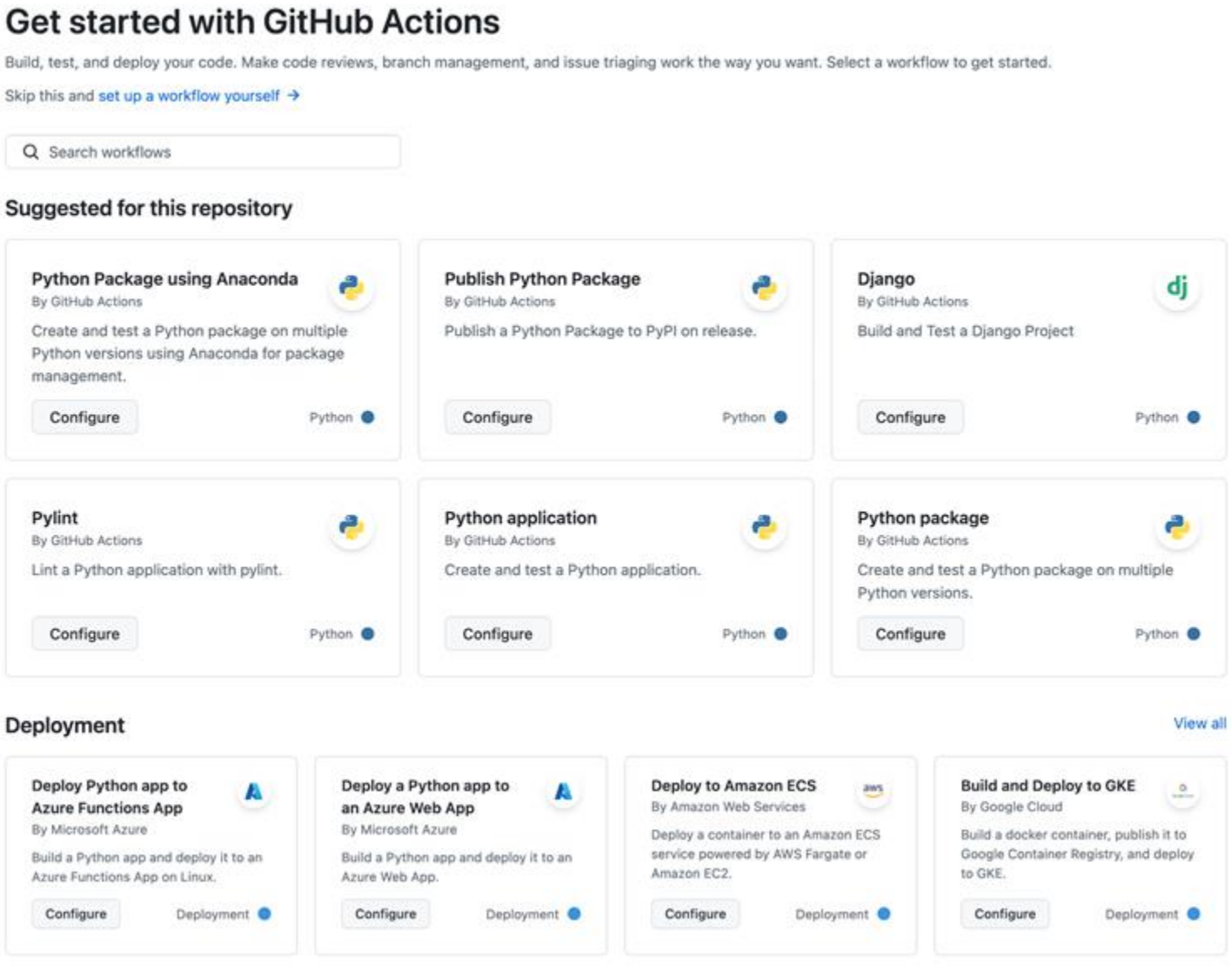
Credit: https://graphite.dev/guides/git-commit-to-new-branch

Part 4 - Branches & Merging

- If you have made changes in `new-feature` branch
- Merge codes to the main branch
  - `git checkout main`
- Add `new-feature` branch changes to `main` branch
  - `git merge new-feature`
- `Delete new-feature` branch if you don't need it
  - `git branch -d new-feature`

Credit: https://www.w3schools.com/git/git_branch_merge.asp

Part 5 - Actions and Automating Tests



Part 5 - Actions and Automating Tests

- **Continuous integration and continuous delivery (CI/CD)**
- **Automate your build, test, and deployment pipeline**
- **Workflow**

  `.github/workflows/github-actions-demo.yml`

  1. `Name`
  2. `Jobs`
  3. `Steps`
  4. `Actions/Runs`

- **Deploy your homepage using GitHub Action (`GitHub Page`)**

  https://pages.github.com

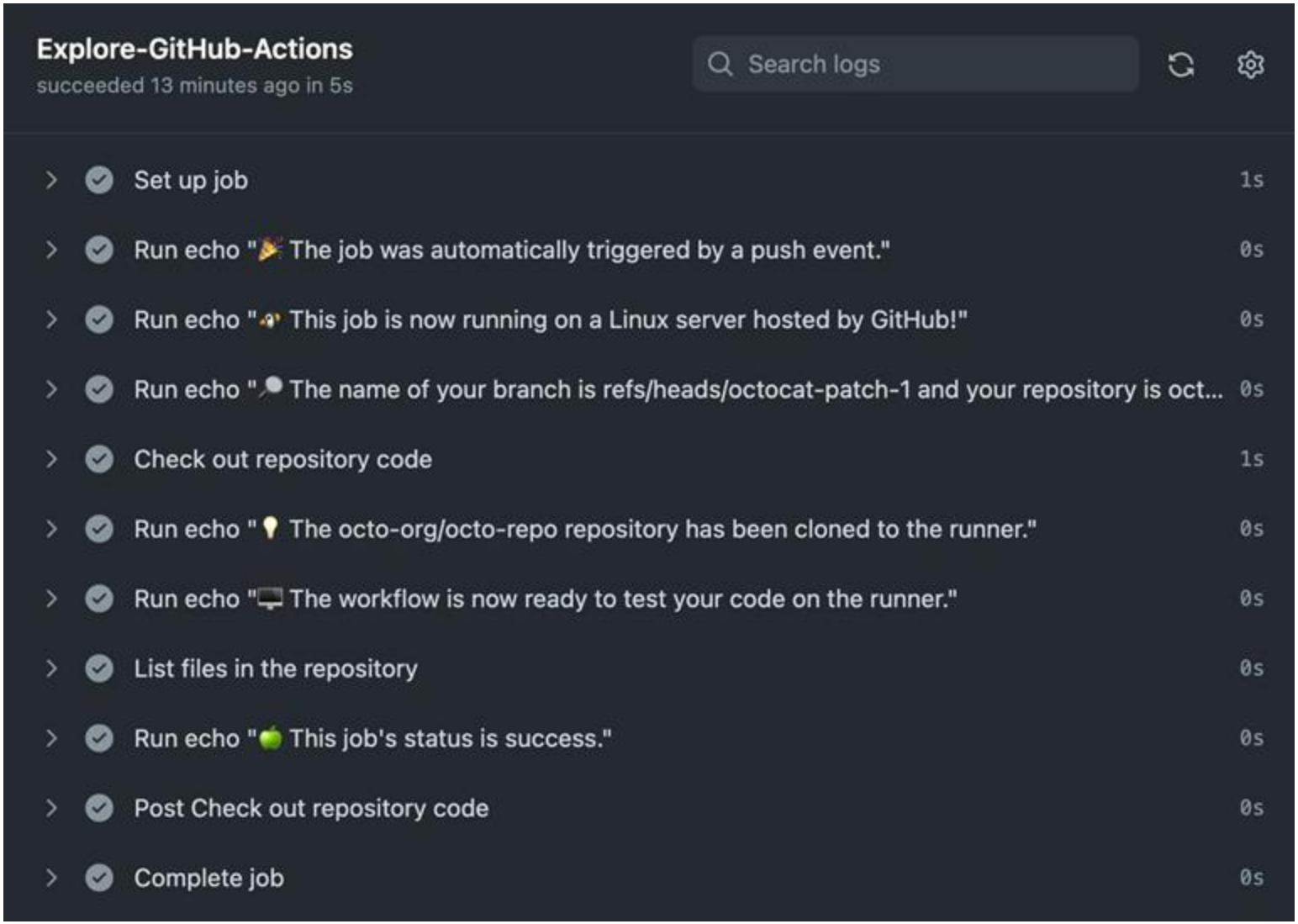
Credit: https://docs.github.com/en/actions/writing-workflows/quickstart

YAML

```yaml
name: GitHub Actions Demo
run-name: ${{ github.actor }} is testing out GitHub Actions 🚀
on: [push]
jobs:
  Explore-GitHub-Actions:
    runs-on: ubuntu-latest
    steps:
      - run: echo "🎉 The job was automatically triggered by a ${{
github.event_name }} event."
      - run: echo "🐧 This job is now running on a ${{ runner.os }} server
hosted by GitHub!"
      - run: echo "🔎 The name of your branch is ${{ github.ref }} and your
repository is ${{ github.repository }}."
      - name: Check out repository code
        uses: actions/checkout@v4
      - run: echo "💡 The ${{ github.repository }} repository has been cloned to
the runner."
      - run: echo "🖥️ The workflow is now ready to test your code on the
runner."
      - name: List files in the repository
        run: |
          ls ${{ github.workspace }}
      - run: echo "🍏 This job's status is ${{ job.status }}."
```

C

Part 5 - Actions and Automating Tests

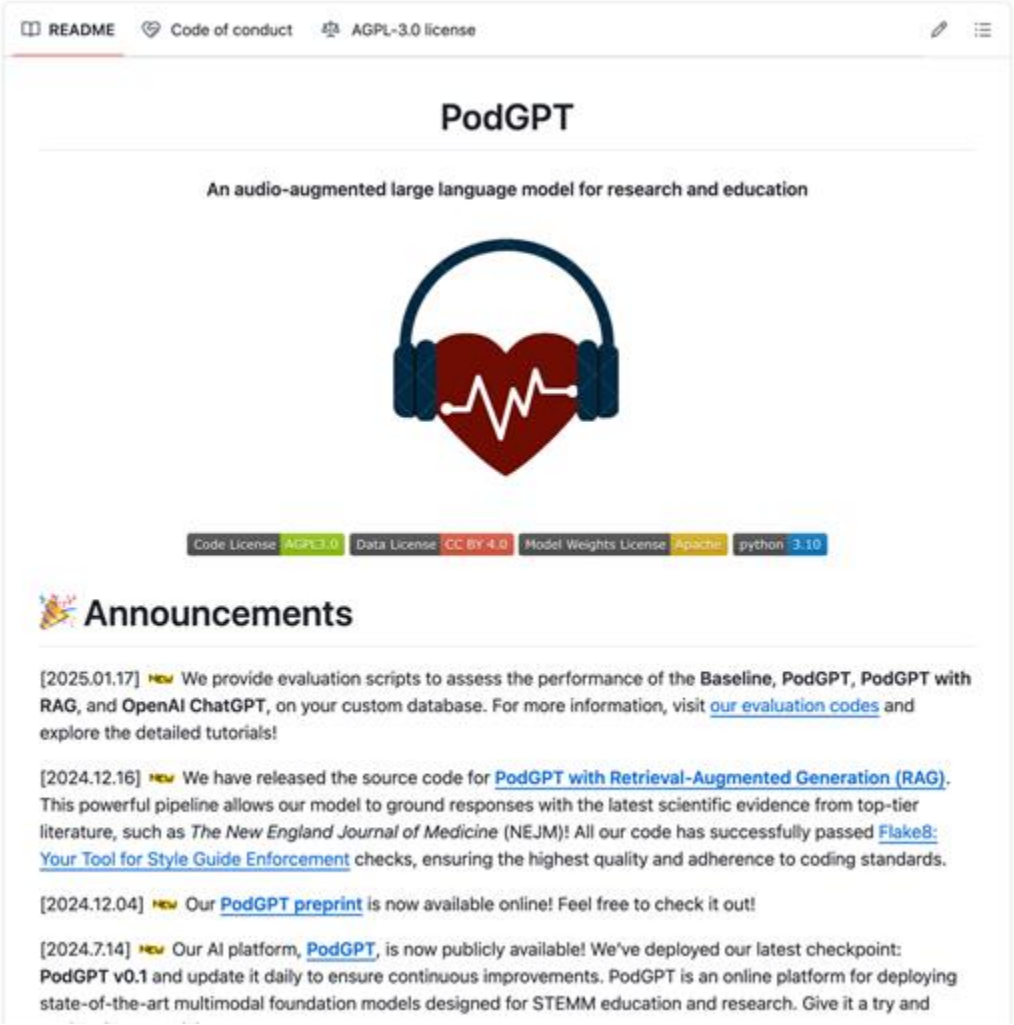


Credit: https://docs.github.com/en/actions/writing-workflows/quickstart

Part 6 - README Examples

Demo: https://github.com/vkola-lab/PodGPT

- Formatted Title
- Fancy Logo
- News & Announcements
- Table of Content
- Fancy Figures
- Codes Block
- Tables
- Structure of the Codes
- Citation
- Contact
- Contributions
- Acknowledgement
- Code of Conduct
- License



GitHub

"collaborative, open sharing in the programming world"

Thank you very much for your attention!