facebookresearch /
PurpleLlama

Code Issues 5 Pull requests 1 Actions Projects Security Insights

Set of tools to assess and improve LLM security.

View license
Code of conduct
Security policy
533 stars **107** forks **20** watching **5** Branches **0** Tags Activity Custom properties
Public repository

main 5 Branches 0 Tags Go to file t **Go to file** + **Add file**

SimonWan and **facebook-github-bot** 19 hours ago

.github/workflows	Initial commit	2 months ago
CybersecurityBenchmarks	Multilingual 6/n: expand MITRE...	19 hours ago
Llama-Guard	Update MODEL_CARD.md (#5)	2 months ago
.gitignore	Initial commit	2 months ago
CODE_OF_CONDUCT.md	Initial commit	2 months ago
CONTRIBUTING.md	Initial commit	2 months ago
LICENSE	Initial commit	2 months ago
README.md	Update README.md (#6)	2 months ago
logo.png	Initial commit	2 months ago

README Code of conduct License Security

Models on Hugging Face | Blog | Website | CyberSec Eval Paper | Llama Guard Paper

---

Purple Llama

Purple Llama is an umbrella project that over time will bring together tools and evals to help the community build responsibly with open generative AI models. The initial release will include tools and evals for Cyber Security and Input/Output safeguards but we plan to contribute more in the near future.

Why purple?

Borrowing a concept from the cybersecurity world, we believe that to truly mitigate the challenges which generative AI presents, we need to take both attack (red team) and defensive (blue team) postures. Purple teaming, composed of both red and blue team responsibilities, is a collaborative approach to evaluating and mitigating potential risks and the same ethos applies to generative AI and hence our investment in Purple Llama will be comprehensive.

License

Components within the Purple Llama project will be licensed permissively enabling both research and commercial usage. We believe this is a major step towards enabling community collaboration and standardizing the development and usage of trust and safety tools for generative AI development. More concretely evals and benchmarks are licensed under the MIT license while any models use the Llama 2 Community license. See the table below:

Component Type	Components	License
Evals/Benchmarks	Cyber Security Eval (others to come)	MIT
Models	Llama Guard	Llama 2 Community License

Evals & Benchmarks

Cybersecurity

We are sharing what we believe is the first industry-wide set of cybersecurity safety evaluations for LLMs. These benchmarks are based on industry guidance and standards (e.g., CWE and MITRE ATT&CK) and built in collaboration with our security subject matter experts. With this initial release, we aim to provide tools that will help address some risks outlined in the White House commitments on developing responsible AI, including:

Metrics for quantifying LLM cybersecurity risks. Tools to evaluate the frequency of insecure code suggestions. Tools to evaluate LLMs to make it harder to generate malicious code or aid in carrying out cyberattacks. We believe these tools will reduce the frequency of LLMs suggesting insecure AI-generated code and reduce their helpfulness to cyber adversaries. Our initial results show that there are meaningful cybersecurity risks for LLMs, both with recommending insecure code and for complying with malicious requests. See our Cybersec Eval paper for more details.

You can also check out the leaderboard here.

Input/Output Safeguards

As we outlined in Llama 2's Responsible Use Guide, we recommend that all inputs and outputs to the LLM be checked and filtered in accordance with content guidelines appropriate to the application.

Llama Guard

To support this, and empower the community, we are releasing Llama Guard, an openly-available model that performs competitively on common open benchmarks and provides developers with a pretrained model to help defend against generating potentially risky outputs.

As part of our ongoing commitment to open and transparent science, we are releasing our methodology and an extended discussion of model performance in our Llama Guard paper. This model has been trained on a mix of publicly-available datasets to enable detection of common types of potentially risky or violating content that may be relevant to a number of developer use cases. Ultimately, our vision is to enable developers to customize this model to support relevant use cases and to make it easier to adopt best practices and improve the open ecosystem.

Getting Started

To get started and learn how to use Purple Llama components with Llama models, see the getting started guide here. The guide provides information and resources to help you set up Llama including how to access the model, hosting, how-to and integration guides. Additionally, you will find supplemental materials to further assist you while responsibly building with Llama. The guide will be updated as more Purple Llama components get released.

FAQ

For a running list of frequently asked questions, for not only Purple Llama components but also generally for Llama models, see the FAQ here.

Join the Purple Llama community

See the CONTRIBUTING file for how to help out.

Releases

No releases published

Packages

No packages published

Contributors 7

Languages