AI Evaluation tools review

### Promptfoo

Provides 2 tools in one npm package : LLM Evaluation tool & Readteam tool

Site : <https://www.promptfoo.dev/>

* **LLM Evaluation tool and redteaming tool**:
  + **Review date** : 21/11/2024
    - **Description**: an [open-source](https://github.com/promptfoo/promptfoo) CLI and library for evaluating and red-teaming LLM apps.  
      With promptfoo, you can:
    - Build reliable prompts, models, and RAGs with benchmarks specific to your use-case
    - Secure your apps with automated [red teaming](https://www.promptfoo.dev/docs/red-team/) and pentesting
    - Speed up evaluations with caching, concurrency, and live reloading
    - Score outputs automatically by defining [metrics](https://www.promptfoo.dev/docs/configuration/expected-outputs/)
    - Use as a [CLI](https://www.promptfoo.dev/docs/usage/command-line/), [library](https://www.promptfoo.dev/docs/usage/node-package/), or in [CI/CD](https://www.promptfoo.dev/docs/integrations/github-action/)
    - Use OpenAI, Anthropic, Azure, Google, HuggingFace, open-source models like Llama, or integrate custom API providers for [any LLM API](https://www.promptfoo.dev/docs/providers/)  
      Test your prompts, agents, and RAGs. Red teaming, pentesting, and vulnerability scanning for LLMs. Compare performance of GPT, Claude, Gemini, Llama, and more. Simple declarative configs with command line and CI/CD integration.  
      It provides 2 main features :
    - [**Red teaming**](https://www.promptfoo.dev/docs/red-team/quickstart/) - LLM security scans
    - [**Evaluations**](https://www.promptfoo.dev/docs/getting-started/) - LLM quality benchmarks
  + FAQ : <https://www.promptfoo.dev/docs/faq/>
  + **Docs**: [docs](https://www.promptfoo.dev/docs/intro/)
    - **GitHub**: <https://github.com/promptfoo/promptfoo>
    - **Price**
      * Free for all features
      * paid for cloud based support and custom stuff
      * Licensing: ?
      * price site : <https://www.promptfoo.dev/pricing/>
      * Free Forever :
      * All LLM evaluation features
      * All model providers and integrations
      * No usage limits
      * Custom integration with your own app
      * Run locally or self-host on your own infrastructure
      * Vulnerability scanning
      * Community support
    - Enterprise : For teams that need advanced features and cloud-based support
      * All Community features
      * Team sharing & collaboration
      * Continuous monitoring
      * Centralized security/compliance dashboard
      * Customized red teaming plugins
      * SSO and Access Control
      * Cloud deployment
      * Priority support & SLA guarantees
    - On-Premise : For organizations that require full control over their infrastructure
      * All Enterprise features
      * Deployment on your own infrastructure
      * Complete data isolation
      * Dedicated support team
  + **Usage, outputs and Deployment**:
    - Local : CLI, Web viewer, node package, python & js support
    - Docker for self hosting
    - share command creates a publicly accessible URL, which means anyone who knows the URL can view your results. If you don't want anyone to see your results, you should keep your URL secret. After 2 weeks, all data associated with the URL is permanently deleted.
      * **Export Formats**: JSON
      * **Visualization Tools**: Web-based interface, CLI
    - **Type** : package
    - **Runtime**: nodejs
    - **Updates & Maintenance**: Regular updates as features are improved and expanded.
      * **Support Channel**: Discord, Email support, Github , book a time
      * **Frequency of Updates**: daily
      * **Community Support**: GitHub repository with active issues, forums for discussions.
        + **GitHub Stars**: Approximately 5k
      * **Maturity**: Relatively new but rapidly evolving with a growing user base (+40k developpers)
      * **Used by** : Pomptfoo has run 16 million probes this year at companies like Shopify Doordash Anthropic Microsoft Discord
      * Team :
  + **Privacy & Security**:
    - **Data Handling**: Local, if cloud → Check the pricing
    - **Compliance / Data Location**: Local, if cloud → Check the pricing
    - **Telemetry** : on per default, can be disabled PROMPTFOO\_DISABLE\_TELEMETRY=1
  + **Features**:
    - **Model Compatibility**: Yes. <https://www.promptfoo.dev/docs/providers/>
    - **Custom Models**: Yes
    - **Evaluation** :
      * **Metrics**: Yes. <https://www.promptfoo.dev/docs/configuration/expected-outputs/>
        + pre-implemented metrics : deterministic metrics, model-graded metrics …
      * **Qualitative & Quantitative Metrics**: TO VERIFY
      * **Custom Metrics**: Yes. Either in the config or external .py or .js scripts
      * **Tasks**: TO VERIFY
      * **Human in the Loop**: Yes but
      * **Static & Dynamic Evaluation**: TO VERIFY
      * **Easy Unit Test (like pytest)**: Yes. with a command a config script or webui
      * **Benchmarks**: TO VERIFY
    - **Redteam**:
      * Metrics : Yes but the default plugins use promptfoo models and send data to their cloud (plugins and custom plugins). <https://www.promptfoo.dev/docs/category/red-teaming/>
      * Custom Metrics : Yes 100% custom
      * **Monitoring**: Continuous monitoring and Comprehensive Scanning & Compliance are paid and only available in entreprise + on-premise tiers (check the pricing page )
      * **CI/CD**: Yes and has also github actions support. <https://www.promptfoo.dev/docs/integrations/ci-cd/>
      * **Dataset Generation**: Yes. <https://www.promptfoo.dev/docs/configuration/datasets/>
    - **Easy to Use**: Easy
      * **Easy to learn** : Easy to Advanced

### Giskard (by Giskard-AI)

### LangSmith (by LangChain)

### DeepEval (by ConfidentAI)

* Package :
  + Description :
    - **DeepEval** is a simple-to-use, open-source LLM evaluation framework, for evaluating and testing large-language model systems. It is similar to Pytest but specialized for unit testing LLM outputs. DeepEval incorporates the latest research to evaluate LLM outputs based on metrics such as G-Eval, hallucination, answer relevancy, RAGAS, etc., which uses LLMs and various other NLP models that runs **locally on your machine** for evaluation.
    - Whether your application is implemented via RAG or fine-tuning, LangChain or LlamaIndex, DeepEval has you covered. With it, you can easily determine the optimal hyperparameters to improve your RAG pipeline, prevent prompt drifting, or even transition from OpenAI to hosting your own Llama2 with confidence.
  + Github : <https://github.com/confident-ai/deepeval>
  + Docs : <https://docs.confident-ai.com/docs/getting-started>
  + python package, pip installable
  + features :
    - Easily "unit test" LLM outputs in a similar way to Pytest.
    - Plug-and-use 14+ LLM-evaluated metrics, most with research backing.
    - Synthetic dataset generation with state-of-the-art evolution techniques.
    - Metrics are simple to customize and covers all use cases.
    - Real-time evaluations in production.
    - integrates natively with [Confident AI](https://app.confident-ai.com/)
* Web platform :
  + Cloud only
  + **evaluate, regression test, and monitor** LLM applications on the cloud.
  + site : <https://www.confident-ai.com/>
  + docs : <https://docs.confident-ai.com/docs/confident-ai-introduction>
  + paid, only 7 days trial , pricing : <https://www.confident-ai.com/pricing>

### Prompt-flow Tracing (by Azure)

### Langfuse

### Validate (by Tonic)

### AutoRAG

## Evaluation benchmarks