Title: GemmaEval: A Comprehensive Automated Benchmark Suite for Gemma

Contact Details: Ryan Rong ryanrong@stanford.edu
https://www.linkedin.com/in/ryan-rong-313779229/

Synopsis: GemmaEval is a comprehensive, automated benchmarking framework designed to evaluate, visualize, and compare Gemma language models against other open source LLMs across a wide spectrum of NLP tasks. This project will introduce a standardized, reproducible testing environment that evaluates models on 1) classic benchmarks such as TriviaQA, GPQA, MGSM; 2) agentic benchmarks such as SWE-bench, TAU-bench; 3) multimodal benchmarks like MMMU; and 4) safety benchmarks like Weapons of Mass Destruction Proxy (WMDP) benchmark.

Related Works:

I want to first reproduce a subset of the benchmark results shown in the Gemma 3 model card and the technical report (an interesting note is that some benchmarks in the technical report are not included in the model card). Then, I want to incorporate new benchmarks like SWE bench and WMDP bench.

Benchmarks

Reasoning and Factuality:

- HellaSwag A dataset designed to test commonsense reasoning by presenting multiple-choice questions that require understanding everyday activities
- TriviaQA A large-scale question-answering dataset containing trivia questions and their corresponding evidence documents, assessing a model's ability to retrieve and generate accurate answers
- Natural Questions Consists of real user queries from Google Search, paired with Wikipedia articles, evaluating a model's capacity to comprehend and extract relevant information from extensive texts.

STEM and Code:

- MMLU-pro An enhanced version of the Massive Multitask Language Understanding benchmark, featuring multiple-choice questions across diverse subjects to test a model's breadth of knowledge.
- MATH A dataset comprising competition-level math problems, challenging models to demonstrate advanced mathematical reasoning and problem-solving skills.
- GSM8K Contains grade-school-level math word problems that require multi-step reasoning, assessing a model's arithmetic and logical capabilities.
- GPQA The Graduate-Level Google-Proof Q&A benchmark includes multiple-choice questions written by domain experts in biology, physics, and chemistry, designed to be PhD-level and resistant to simple retrieval-based answering.
- LiveCodeBench A benchmark designed to evaluate the performance of language models in live coding scenarios, assessing their ability to generate correct and efficient code in real-time environments.

Multilingual:

- MGSM The Multilingual Grade School Math benchmark features math word problems in multiple languages, testing a model's numerical reasoning across different linguistic contexts.
- XQuAD A multilingual question-answering dataset that extends the SQuAD benchmark to several languages, evaluating cross-lingual transfer learning abilities.

Multimodal:

- MMMU The Massive Multi-discipline Multimodal Understanding benchmark assesses a model's capability to interpret and reason across diverse subjects using both textual and visual information.
- TallyQA A dataset focused on counting-based questions over images, testing a model's visual understanding and numerical reasoning in multimodal contexts.

Agentic:

- SWE-bench verified An agentic benchmark evaluating a model's proficiency in software engineering tasks, including code generation, debugging, and comprehension.
- TAU-bench The Task Automation Benchmark assesses a model's ability to understand and execute complex, multi-step tasks autonomously, reflecting its agentic capabilities.

Safety:

- BioLP Bench A safety benchmark designed to evaluate language models in the biomedical domain, focusing on their ability to handle sensitive information responsibly and accurately.
- Weapons of Mass Destruction Proxy (WMDP) benchmark A safety benchmark assessing a model's potential to generate or withhold information related to weapons of mass destruction, ensuring ethical constraints are maintained.
- Additional methods could refer to Evaluating Frontier Models for Dangerous Capabilities

Frameworks

Some existing open-source frameworks for evaluation include OpenAI’s simple-evals and evals. The goal is to build on top of these frameworks to incorporate Gemma, other benchmarks listed above, and build visualization tools to generate graphs, charts, and tables.

Open Source Models

I will compare Gemma with top open source models, including both text-only and multimodal ones:

Gemma 3:

- 1B
- 4B
- 12B
- 27B

Mistral:

- Mistral-Small-3.1-24B-Base-2503 (24B)
- Pixtral-2409 (12B) (multimodal)

Llama:

- Llama 3.3 70B-Instruct
- Llama 3.2-1B
- Llama 3.2-3B
- Llama 3.2-11B-Vision

Qwen:

- Qwen-2.5-VL-32B-Instruct
- Qwen-2.5-VL-7B-Instruct
- Qwen-2.5-VL-3B-Instruct
- Qwen-2.5-32B
- Qwen-2.5-14B
- Qwen-2.5-7B
- Qwen-2.5-3B
- Qwen-2.5-1.5B

DeepSeek:

- DeepSeek-R1-Distill-Qwen-32B
- DeepSeek-R1-Distill-Qwen-14B
- DeepSeek-R1-Distill-Qwen-7B
- DeepSeek-R1-Distill-Qwen-1.5B
- DeepSeek-R1-Distill-Llama-8B
- Janus-Pro-7B

Deliverables:

- Week 1-2: Project initiation, including loading LLMs, reviewing existing evaluation frameworks (e.g., OpenAI's simple-evals and evals), and finalizing the benchmark selection.
- Week 3: Implement the infrastructure for benchmark integration, focusing on classic benchmarks such as HellaSwag, TriviaQA, and Natural Questions.
- Week 4: Integrate STEM and code-related benchmarks, including MMLU-pro, MATH, GSM8K, GPQA, and LiveCodeBench.
- Week 5-6: Incorporate multilingual benchmarks (MGSM, XQuAD) and multimodal benchmarks (MMMU, TallyQA) into the evaluation suite.
- Week 7 (July 14 - July 18): Prepare for the midterm evaluation by documenting progress, addressing any challenges encountered, and demonstrating the framework's capabilities with the integrated benchmarks.
- Week 8-10: Focus on agentic benchmarks, integrating SWE-bench verified and TAU-bench, and ensure the framework supports dynamic task execution.
- Week 11: Implement safety benchmarks, including BioLP Bench and WMDP, to assess and ensure ethical model behavior.
- Week 12: Develop visualization tools (graphs, charts, tables)
- Week 13 (August 25 - September 1): Finalize the project by refining the framework, conducting comprehensive testing across all benchmarks, generating and preparing documentation for the final evaluation.

About Me:

I am a CS student at Stanford University with 3+ years of experience in AI and machine learning. Last fall, I developed a novel framework designed for medium to long-range weather forecasting, employing chunked sequential diffusion and specialized techniques like diffusion forcing to produce spatiotemporally coherent forecasts. My project won Best Project Award (top 1.8%) of Andrew Ng's Machine Learning class (CS229) at Stanford. Most recently, I worked at the Stanford AI Lab to build Trace, a framework to iteratively optimize LLMs on downstream tasks like GPU optimization. I plan to submit the paper to be published in COLM.

Professionally, I'm working as a Machine Learning Engineer at Kos AI, a Silicon Valley startup in the healthcare sector. I'm developing innovative algorithms to extract health metrics, such as blood glucose, blood oxygen levels, and heart rate from noisy photoplethysmography data. I will join the Eater Pricing team at Uber as a software engineer intern this summer.

In the past three years, I have conducted in-depth AI and machine learning research at both UIUC and Stanford. At UIUC, I built generative models to create synthetic code bugs to automate the debugging process. At Stanford School of Medicine, I built algorithms to analyze the electronic health records at Stanford Hospital and predict the onset of Alzheimer's Disease. At Stanford Institute for Human-Centered Artificial Intelligence, I developed an AI writing assistant that generates Reddit and Twitter comments based on threads, posts, and discussions. I look forward to further exploring AI research with Google this summer.

Here is my resume for further reference.