A12



WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild

Bill Yuchen Lin^{1,2}

Yuntian Deng^{1,3} Khyathi Chandu¹ Faeze Brahman¹ Abhilasha Ravichander¹ Valentina Pyatkin^{1,2} Nouha Dziri¹ Ronan Le Bras¹ Yejin Choi^{1,2}

¹Allen Institute for Al ²University of Washington ³University of Waterloo

WildBench is an automatic evaluation benchmark for large language models. It

- uses real-world user queries
- gets a Pearson correlation of 0.98 with Chatbot Arena

1/ Example Tasks

Traditional Synthetic Benchmarks

What is the capital of Australia? What is some cool music from the 1920s? How do I wrap a present neatly? Can you write code? ~20 recipe generation examples

Diverse tasks from real users!

Please provide me python code to go through a directory and its subdirectories and delete images that are not horizontal.

hey can you write an essay on the impact of the G20 summit on the global economy, trade, development and the role of young people in shaping the future of the world, it has to have more than 1200 words. Write it beautiful and poetic. Use extensive vocabulary. Use a lot of factual and empirical data. Use some, ancient indian historical references.

I want to create an open source, highly realistic and grounded textbased business simulation game that is played in the terminal, with a large range of different features that make the game as realistic a simulation as possible. In light of this the game should not have set values for anything because that is unrealistic - real life isn't like that; the sim should be as close to reality as possible. I will host it on Github. Please create a FULL, COMPLETE file structure for the game's Github repo.

Role playing

Data Analysis

Creative Writing

Reasoning Planning Math

https://huggingface.co/spaces/allenai/WildBench

2/ Data Curation Sourced from WildChat chatbots (https://wildvisualizer.com/) Withheld from WildChat public release to prevent leakage #Tasks | #Turns | ChatHistory | QueryLen | PromptLen | RealUser | TaskTag | **Evaluation Dataset MT-Bench** Score 164.9 **AlpacaEval** 164.9 Pair (ref=1) ArenaHard 406.4 406.4 500 Pair (ref=1) | Score+Pair (ref=3) **WILDBENCH** | 1,024 | **≤**5 978.5 WildBench (1024) AlpacaEval (805) ArenaHard (500) Coding & Debugging Writing Information seeking seeking Brainstorming Role playing Advice seeking Brainstorming Debugging Role playing

Reasoning Planning

Data Analysis

3/ Automatic Evaluation **WB-Reward Pairwise** json_output = History "analysis of A": "[analysis of Response A]", Model X vs Y (Baseline) "analysis of B": "[analysis of Response B]", +1 when X>>Y; +0.5 when X>Y; **Query** '**reason of A–B**": "[where Response A and B perform equally]". -1 when X<<Y; -0.5 when X<Y; LLM B's 'reason of A>B": "[where Response A is better than B]", 0 when X=Y; w/ Length Penalty response "**reason of B>A**": "[where Response B is better than A]", "choice": "[A++ or A+ or A-B or B+ or B++]" Models → A++ means A is **much** better, A+ means A is **slightly** better, **Checklist** Individual json_output = History 'strengths": "[analysis for the strengths]", Query **LLM** response "weaknesses": "[analysis for the weaknesses]", "score": **"[1~10]"** Score 5~6: The response is fair but has some issues (e.g., factual errors, hallucinations, missing key information); ... Example Task (history + query) Checklist (a list of questions and criteria for eval) Correlation w/ Does the alternative formula provided correctly address the user's return the value from the row in column B Al: need to find the last matching value in a specified column and return a **SER:** the formula does not appear to be finding corresponding value from another column? the last value in column A; 👜 Al: 2 Is the alternative formula syntactically correct and compatible with **■ USER:** you provided the exact same spreadsheet software such as Microsoft Excel or Google Sheets? formula, is there an alternative formula





4/ Instance-Specific Checklists

Example checklist for the G20 task example

- ✓ Does the essay contain more than 1200 words as requested by the user?
- ✓ Is the language of the essay beautiful and poetic, incorporating extensive vocabulary as
- ✓ Does the essay include a significant amount of factual and empirical data related to the impact of the G20 summit on the global economy, trade, and development?
- ✓ Are there references to the role of young people in shaping the future of the world within
- the context of the G20 summit?
- ✓ Does the essay include ancient Indian historical references as requested by the user? ✓ Is the essay structured in a clear and logical manner, facilitating an easy understanding
- of the discussed topics?
- Combines outputs from LLMs to produce initial checklists
- Manually reviewed

Data Analysis

- Prompts LLM judge to use checklists to guide their evaluation
- Makes the evaluation process more interpretable
- Achieves higher correlation with humans compared to without

