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

Bill Yuchen Lin[1,2]
Yuntian Deng[1,3] **Khyathi Chandu**[1] **Faeze Brahman**[1] **Abhilasha Ravichander**[1]
Valentina Pyatkin[1,2] **Nouha Dziri**[1] **Ronan Le Bras**[1] **Yejin Choi**[1,2]

[1]Allen Institute for AI [2]University of Washington [3]University of Waterloo
https://huggingface.co/spaces/allenai/WildBench

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

WildBench *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 Use some, ancient indian historical references.

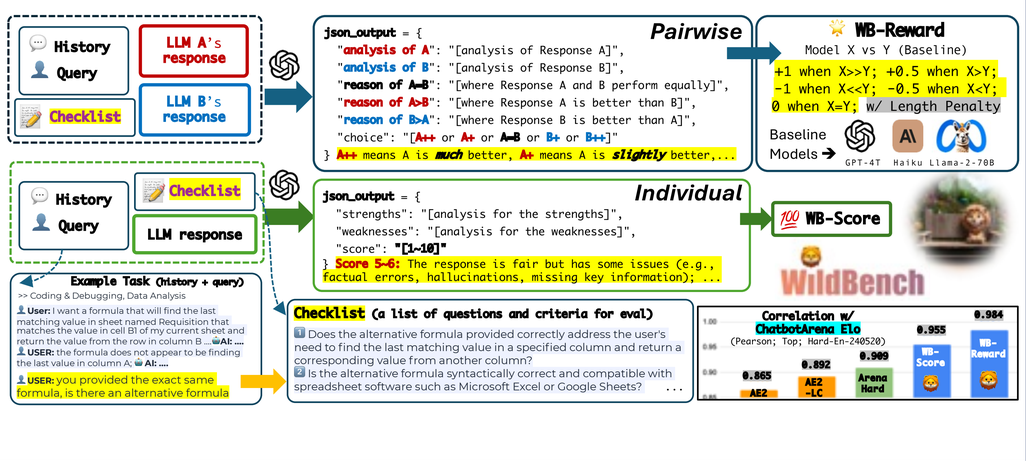
I want to create an open source, highly realistic and grounded text-based business simulation game that is played in the terminal, with a large range of different features that make the game as realistic a simulation as possible. In light of this the game should not have set values for anything because 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.

2/ Data Curation

- Sourced from WildChat chatbots (https://wildvisualizer.com/)
- Withheld from WildChat public release to prevent leakage

Dataset	#Tasks	#Turns	ChatHistory	QueryLen	PromptLen	RealUser	TaskTag	Evaluation
MT-Bench	80	2	✔Dynamic	202.2	Dynamic	✖	✔	Score
AlpacaEval	805	1	✖	164.9	164.9	✖	✖	Pair (ref=1)
ArenaHard	500	1	✖	406.4	406.4	✔	✖	Pair (ref=1)
WILDBENCH	1,024	≤5	✔Static	978.5	3402.1	✔✔	✔	Score+Pair (ref=3)

3/ Automatic Evaluation



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

5/ Enables Performance Breakdown