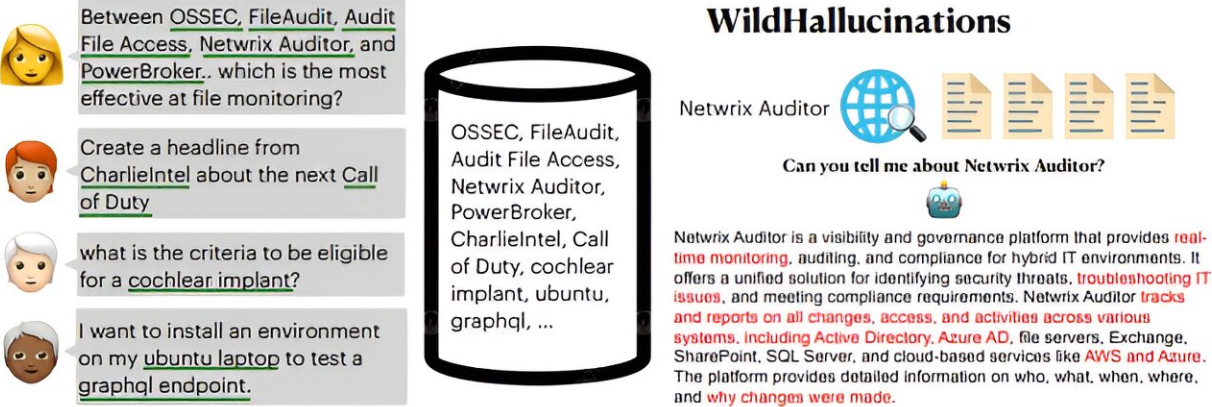


New benchmarking tool evaluates the factuality of LLMs

August 21 2024, by Bob Yirka



The diagram illustrates the WildHallucinations benchmarking tool. On the left, four user prompts are shown in speech bubbles, each with a different emoji representing a user. The prompts are: 1. 'Between OSSEC, FileAudit, Audit File Access, Netrix Auditor, and PowerBroker.. which is the most effective at file monitoring?' 2. 'Create a headline from CharlieIntel about the next Call of Duty' 3. 'what is the criteria to be eligible for a cochlear implant?' 4. 'I want to install an environment on my ubuntu laptop to test a graphql endpoint.' In the center, a large black-outlined cylinder represents a database or knowledge base containing the following text: 'OSSEC, FileAudit, Audit File Access, Netrix Auditor, PowerBroker, CharlieIntel, Call of Duty, cochlear implant, ubuntu, graphql, ...'. On the right, the 'WildHallucinations' interface is shown. It features the title 'WildHallucinations' at the top. Below it, the text 'Netrix Auditor' is displayed next to a globe icon and four document icons. A question is posed: 'Can you tell me about Netrix Auditor?' with a small robot icon below it. The answer provided is: 'Netrix Auditor is a visibility and governance platform that provides real-time monitoring, auditing, and compliance for hybrid IT environments. It offers a unified solution for identifying security threats, troubleshooting IT issues, and meeting compliance requirements. Netrix Auditor tracks and reports on all changes, access, and activities across various systems, including Active Directory, Azure AD, file servers, Exchange, SharePoint, SQL Server, and cloud-based services like AWS and Azure. The platform provides detailed information on who, what, when, where, and why changes were made.'

Overview of WILDHALLUCINATIONS. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2407.17468

A team of AI researchers and computer scientists from Cornell University, the University of Washington and the Allen Institute for Artificial Intelligence has developed a benchmarking tool called WILDHALLUCINATIONS to evaluate the factuality of multiple large language models (LLMs). The group has published a [paper](#) describing the factors that went into creating their tool on the *arXiv* preprint server.

LLMs such as ChatGPT have become popular—people use them to write letters, poems, songs, [research papers](#) and other text documents.

But over time, their deficiencies have become quite clear—LLMs often make inaccurate statements. Such mistakes, if they veer too far from reality, have come to be known as hallucinations.

The research team notes that the main reason LLMs hallucinate is due to the quality of the data used to train them—generally, massive amounts of text from the internet. Thus, models trained on specific, highly accurate datasets are much more likely to provide [accurate information](#).

The research team noted that the makers of many LLMs have been making claims about revised versions of their models, often suggesting that they hallucinate less often, implying that they are more accurate. But the researchers also noted that to date, users have no way to verify whether such claims are true. For this new study, the team created a tool to help the user community evaluate some of the most popular LLMs for accuracy.

Called WILDHALLUCINATIONS, the benchmark tool prompts multiple LLMs to generate output from user-generated chatbot conversations. It then fact-checks the answers. Noting that many chatbot answers come from information provided on Wiki pages, the research team made sure to note differences in answers regarding queries that had information that could be found on Wikipedia and those that could not.

To test their benchmarking tool, the researchers used it to evaluate several of the most popular LLMs, many of which had recently been updated. They found that LLM makers have not made much progress in improving accuracy. Most were no more accurate than their prior versions.

The team also discovered that most of the models did better when they could pull information from one or more Wiki pages. LLMs also did better with some subjects compared to others. They had trouble, for

example, finding reliable information regarding celebrities and [financial issues](#). They were more reliable when asked certain types of science questions.

More information: Wenting Zhao et al, WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries, *arXiv* (2024). [DOI: 10.48550/arxiv.2407.17468](https://doi.org/10.48550/arxiv.2407.17468).
arxiv.org/abs/2407.17468

© 2024 Science X Network

Citation: New benchmarking tool evaluates the factuality of LLMs (2024, August 21) retrieved 21 August 2024 from
<https://techxplore.com/news/2024-08-benchmarking-tool-factuality-llms.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.