Examining emerging capabilities and mitigating potential risks of VLLMs

ONE DAY WORKSHOP | 50+ ATTENDANCE (expected)

2 tasks, 3 talks, panel discussion

Abstract

The emergence of very large language models (VLLMs) has dramatically altered the trajectory of progress in AI and its applications. With the release of ChatGPT, that excitement has now transcended boundaries from AI researchers to the common man; indeed it asserts we are living in an exciting time of scientific proliferation. However, we see two divergent community views on VLLMs. Believers in the magical powers of VLLMs claim that VLLMs are autodidactic in learning a wide gamut of new capabilities – referred to as "*emerging capabilities*" in the community – an effect that gets pronounced with model size and dataset scale. On the other hand, critics are not yet prepared to acknowledge the self-learning power of VLLMs; they criticize it as only a statistical learner and out several flaws – *hallucination* being the most prominent one. In this forum, we want to bring together both communities – the believers and the critics, to explore exciting tasks together: (i) **CT^2** - *Counter Turing Test: AI-Generated Text Detection*, and (ii) HELTER - *HallucinatIon eLiciaTion through automatic detection and mitigation*. Expect this to be an exciting forum to discuss, debate, and explore exciting scientific pathways for the future.

1 Rationale - Why does AI need to be civilized? - Call for Papers (CFP)

Advances in AI during the past couple of years have led to AI systems becoming immensely more powerful than ever before. While their applications for social good cannot be overstated, as an unintended by-product, risks of misuse have also exacerbated. This prompted an open petition letter [19] (led by Gary Marcus) by the nonprofit Future of Life Institute, calling for all AI labs to immediately pause for at least 6 months "moratorium" the training of AI systems more powerful than GPT-4. The letter has (18K+ and still counting) signatures from technologists and luminaries, which include Yoshua Bengio, Stuart Russell, Elon Musk, Steve Wozniak, and Andrew Yang. It also includes policy leaders such as Rachel Bronson, president of the Bulletin of the Atomic Scientists, a science-oriented advocacy group known for its warnings against humanity-ending nuclear war. On the other hand, the opposing campaign, which doesn't believe in halting scientific progress, has powerful people too, including Bill Gates [3], Andrew Ng [33], Yann LeCun [33] and many others.

This is a significant time in the history of scientific development. In this forum, we will discuss and debate emerging capabilities and mitigating potential risks and limitations of VLLMs. Call for papers includes, but is not limited to: • *unique emerging abilities of VLLM*; • *negative, position, and full paper on potential risks of VLLMs*; • *ethics and VLLMs*; • *making VLLMs more responsible*; • *detection AI-generated content*; • *mitigation of harmful hallucinations*.

2 Two Shared Tasks

Shared tasks are an effective way to attract research attention to any emerging area. We will host two shared tasks: (i) **CT^2**, and (ii) HELTER. The findings of **CT^2** will mitigate misusage risks, while HELTER endeavours to make VLLMs more human sensitive and responsible.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2.1 $\mathbf{CT^2}$ - Counter Turing Test for AI-Generated Text Detection

With the emergence of ChatGPT, the risk of AI-generated content has reached an alarming apocalypse. ChatGPT has been declared banned by the school system in NYC [25], Google ads [10], and Stack Overflow [17], while scientific conferences like ACL [5] and ICML [8] have released new policies deterring the usage of ChatGPT for scientific writing. After the initial scepticism, ChatGPT has been seen as a listed author in scientific papers [14, 21], while Elsevier [7] and Springer [26] have adopted more inclusive guidelines on *the use of ChatGPT for scientific writing*.

Indeed, detecting AI-generated text has suddenly emerged as a concern that needs immediate attention. While watermarking as a potential solution to the problem is being studied by OpenAI [32], a handful of systems that detect AI-generated text such as GPT-2 output detector [31], GLTR [27], GPTZero [29], DetectGPT [20], etc. have recently been orange observed in practical use. To address the inevitable question of ownership attribution for AI-generated artifacts, the US Copyright Office [22] released a statement stating that if the content is traditional elements of authorship produced by a machine, the work lacks human authorship and the office will not register it for copyright. Given this cynosural spotlight on generative AI, AI-generated text detection is a topic that needs a thorough investigation. In this regard, there are three families of techniques proposed so far:

- **Watermarking**: First introduced in [1], watermarking AI-generated text involves embedding an imperceptible code or signal to verify the author of a particular text with certainty. [13] proposed this by selecting the next token pseudorandomly (rather than simply choosing the one with the highest probability) using a cryptographic pseudorandom function whose key is only possessed by the LLM maker. It would be remiss not to mention the most obvious pitfall of this approach, which is that if the text is altered or modified in any way, detecting the watermark proves to be a difficult task.
- **Negative log likelihood (NLL)**: NLL-based implementations such as DetectGPT [20] have demonstrated the detection of AI-generated text by comparing log-likelihood of generated tokens after perturbing the input text by replacing some tokens with others. If the new, perturbed version of the text lies in the negative curvature regions of log-likelihood, it was likely generated by AI. The limitation of this approach is that it requires access to the log probabilities of the text in order to work which implies that knowledge of which LLM was used to generate the text is essential;
- **Perplexity and Burstiness**: GPTZero [29], an example of a detection technique based on perplexity and burstiness, has demonstrated that a text with lower perplexity (a measure of how predictable the text is), and with lower burstiness (the measure of how uniform text is) has a high probability of being generated by an AI. The limitations here are that GPTZero also requires access to log probabilities of text as well as the fact that it approximates perplexity values using a linear model.

Although AI-generated text detection has suddenly received immense attention, Liang et al. [16] suggest that available AI-generated text detectors consistently misclassify non-native English writing samples as AI-generated, whereas native writing samples are accurately identified, highlighting the ethical implications of deploying AI-generated content detectors and risking misrepresentation. This implies that a community effort is needed to tackle the issue of detectors penalizing under-represented sub-population(s). The CT^2 task will be the first of its kind in bringing together researchers in advancing the area of detecting AI-assisted generated text.

2.1.1 Data to be released and the task

CT^2 will consist of three sub-tasks. We will be releasing 100K data points, consisting of (i) prompt, (ii) human-written text, and (iii) AI-generated text by 5 different LLMs.

- **Task A**: given a set of human-generated text documents vs. AI-generated text documents participants need to design techniques to detect AI-generated text. Indeed, human-written text vs. AI-generated text would be parallel, which means they will be on the same topic. In this task, we will let participants know that the generated text is from GPT, OPT, BERT, XLNet, etc. As such, this is an LLM-specific AI detection task.
- **Task B**: in this task, we will not tell people which the generated text is using which LLM. Participants need to design techniques which is LLM agnostic.
- **Task C**: In this task, we will offer AI-assisted writing, i.e., AI-generated text interlaced with minor edits by another language model and human, as input. Given the intricacies and challenges of AI-assisted writing, it would be the hardest task to attempt.

2.2 HELT: HallucinatIon eLiciaTion through automatic detection and mitigation

With the recent and rapid advances in the areas of LLMs and Generative AI, the pre-eminent and ubiquitous concern is of hallucination. We release large-scale first-of-its-kind human-annotated data with detailed annotations on - intrinsic vs. extrinsic hallucinations, and degree of hallucination, and we ask participants to come up with either black-box factuality Assessment and/or evidence-based fact-checking. First, we define categories of hallucinations:

- **Intrinsic Hallucination**: Intrinsic hallucination refers to the phenomenon when an LLM generates text that topically slightly deviates from the input and/or has a lack of grounding in reality. For example, given a prompt "*USA on Ukraine war*" an LLM generates *"U.S. President Barack Obama says the U.S. will not put troops in Ukraine"*. We can see a clear case of intrinsic hallucination as the US president during the Ukraine-Russia war is Joe Biden, not Barack Obama, contradicting the reality.
- **Extrinsic Hallucination**: We define extrinsic hallucination to be the generated output from an LLM that cannot be verified, or contradicted from the source content provided as prompt. For example, we provide a prompt stating *"North Korea has conducted six underground nuclear tests, and a seventh may be on the way."*, and the resulting output generated by the model was *"The international community has condemned North Korea's nuclear tests, with the UN Security Council imposing a range of sanctions in response. The US and other world powers have urged North Korea to abandon its nuclear weapons program and to comply with international law."* As the source input makes no reference to the U.S. or U.N. Security Council imposing any sanctions, the claimed imposition of sanctions cannot be verified from the input alone.

Next, based on the degree of hallucination, we categorize it into three levels: **(i) Mild**: At this level, hallucination can be categorized as minor and the generated output may just contain innocuous errors or inconsistencies in the text. The generated text will not significantly impact its coherence, **(ii) Moderate**: Moderate hallucination will involve more significant errors or distortions in the generated text. The text may contain nonsensical phrases or ideas that deter from the topic at hand., **(iii) Alarming**: Alarming hallucinations in LLM entail a drastic deviation from the intended output. Such deviations can also be offensive and incongruous with the desired output.

2.2.1 Detection Methods

There are mainly two broad ways: **(i) Black-box factuality assessment**: Black-box hallucination detection approaches, such as SelfCheckGPT [18], endeavour to fact-check models without utilizing an external database of facts. In the case of SelfCheckGPT, it has the capacity to fact-check models in a zero-resource fashion by evaluating if the generated facts have similarities or if they contradict each other. **(ii) Evidence-based fact checking**: Evidence-based fact-checking, such as LLM-Augmenter [24], leverages the use of task-specific databases or similar reliable resources that contain accurate facts.

2.2.2 Data to be released and the task

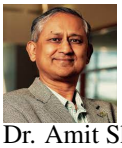
We will be releasing 20K annotated data manually labelled at sentence level on -i) intrinsic vs. extrinsic hallucination, and ii) degree of hallucination mild, moderate, and alarming.

- **Task A**: Given an AI-generated text and associated prompt the task is to detect hallucination at the sentence level. For the competition purpose overall hallucination detection accuracy will be considered averaged over sub-categories like intrinsic and extrinsic.
- **Task B**: This task is to propose hallucination mitigation techniques. While evidence-based fact-checking is well studied in the fact-checking community [28, 30, 9, 15, 12, 11, 23, 2], here in this forum we are mainly interested in black-box factuality assessment. However, teams can use external resources to mitigate hallucination and will be considered in the evidence-based mitigation group. We will request participants to report such experiments in their reports. Since it is hard to assess whether the reformed text still has hallucinations, Task B will mostly be an academic exercise.

3 Invited Talks & Panel Debate [tentative]

Talks: We wish to have 3 speakers from industry/academia, people who have experience in building LLMs. We (tentatively) plan to invite Prof. Christopher Manning, Professor of Linguistics and Computer Science Director, Stanford Artificial Intelligence Laboratory (SAIL); Dr. Vinodkumar Prabhakaran, Senior Research Scientist, Google LLC, co-author of PaLM [6]; and Nick Ryder who is a principal researcher at is the co-author of the GPT3 [4], works at OpenAI.

Workshop Organizers

Name	Web, Email, GScholar	Research Interest & Organizing activities
 Dr. Amitava Das	Web Email Google Scholar	Dr. Amitava Das is a Research Associate Professor at AIISC, UofSC, USA, and an advisory scientist at Wipro AI Labs, Bangalore, India. **Research interests:** Code-Mixing and Social Computing. **Organizing Activities** [selective] - DEFACTIFY 1.0 @AAAI2022 - DEFACTIFY 2.0 @AAAI2023 - Memotion @SemEval2020 - SentiMix @SemEval2020 - Computational Approaches to Linguistic Code-Switching @ LREC 2020 - CONSTRAINT @AAAI2021
 Dr. Amit Sheth	Web Email Google Scholar	Dr. Amit Sheth is the founding Director of the Artificial Intelligence Institute, NSC Chair and a Professor of CSE at the University of South Carolina. He was awarded 2023 IEEE Wallace McDowell award, and is a fellow of IEEE, ACM, AAAS, and AAAI. **Research interests:** Organized over 40 workshops, given over 40 tutorials, Neurosymbilic Knowledge Graph, NLP/NLU, translational research. **Organizing Activities** [selective] - Cysoc2021 @ ICWSM2021 - Emoji2021 @ICWSM2021 - KiLKGC 2021 @KGC21
 Aman Chadha	Web Email Google Scholar	Aman Chadha is an Applied Science Manager at Amazon Alexa AI and a Researcher at Stanford AI. **Research interests:** Multimodal AI, On-device AI, and Human-Centered AI.
 Vinija Jain	Web Email Google Scholar	Vinija Jain is a Machine Learning Lead at Amazon Music. **Research interests:** Recommender Systems and NLP.

References

[1] Scott Aaronson. My Projects at OpenAI. 2022. URL: https://scottaaronson.blog/?p=6823.

[2] Rami Aly et al. "FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information". In: 2021. URL: https://arxiv.org/abs/2106.05707.

[3] Bill Gates says calls to pause AI won't 'solve challenges'. Apr. 2023. URL: https://www.reuters.com/technology/bill-gates-says-calls-pause-ai-wont-solve-challenges-2023-04-04/.

[4] Tom Brown et al. "Language Models are Few-Shot Learners". In: Advances in Neural Information Processing Systems. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[5] Program Chairs. ACL 2023 policy on Ai Writing Assistance. Jan. 2023. URL: https://2023.aclweb.org/blog/ACL-2023-policy/.

[6] Aakanksha Chowdhery et al. PaLM: Scaling Language Modeling with Pathways. 2022. arXiv: 2204.02311 [cs.CL].

[7] Elsevier. The Use of AI and AI-assisted Technologies in Scientific Writing. 2023. URL: https://www.elsevier.com/about/policies/publishing-ethics?fbclid=IwAR2DBcQShp05yS7y7BT0LUxZBTVLego78m4j2tOKshCiWlCcXQpwHADka1s#Authors.

[8] Neural Information Processing Systems Foundation. 2023. URL: https://icml.cc/Conferences/2023/llm-policy.

[9] Sonal Garg and Dilip Kumar Sharma. "New Politifact: A Dataset for Counterfeit News". In: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 17–22. DOI: 10.1109/SMART50582.2020.9337152.

[10] Nico Grant and Cade Metz. A new chat bot is a 'code red' for google's search business. Dec. 2022. URL: https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html.

[11] Ashim Gupta and Vivek Srikumar. "X-FACT: A New Benchmark Dataset for Multilingual Fact Checking". In: arxiv, 2021. URL: https://arxiv.org/abs/2106.09248.

[12] Yichen Jiang et al. "HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification". In: Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. URL: https://aclanthology.org/2020.findings-emnlp.309.pdf.

[13] John Kirchenbauer et al. A Watermark for Large Language Models. 2023. arXiv: 2301.10226 [cs.LG].

[14] Tiffany H Kung et al. "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models". In: PLoS digital health 2.2 (2023), e0000198.

[15] Tom Kwiatkowski et al. "Natural Questions: A Benchmark for Question Answering Research". In: Transactions of the Association for Computational Linguistics 7 (2019), pp. 452–466. DOI: 10.1162/tacl_a_00276. URL: https://aclanthology.org/Q19-1026.

[16] Weixin Liang et al. GPT detectors are biased against non-native English writers. 2023.

[17] Mod Makyen and Peter Olson. Temporary policy: Chatgpt is banned. Dec. 1969. URL: https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned.

[18] Potsawee Manakul, Adian Liusie, and Mark JF Gales. "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models". In: (2023). URL: https://arxiv.org/abs/2303.08896.

[19] Gary Marcus and Future of Life Institute. Pause Giant AI Experiments: An Open Letter. Mar. 2023. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[20] Eric Mitchell et al. "Detectgpt: Zero-shot machine-generated text detection using probability curvature". In: arXiv preprint arXiv:2301.11305 (2023).

[21] Siobhan O'Connor et al. "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?" In: Nurse Education in Practice 66 (2022), pp. 103537–103537.

[22] Copyright Office. "Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence". In: Library of Congress, Mar. 2023. URL: https://public-inspection.federalregister.gov/2023-05321.pdf.

[23] Yasumasa Onoe et al. "CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge". In: arXiv, 2021. DOI: 10.48550/ARXIV.2109.01653. URL: https://arxiv.org/abs/2109.01653.

[24] Baolin Peng et al. "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback". In: (2023).

[25] Kalhan Rosenblatt. "CHATGPT banned from New York City public schools' devices and Networks". In: NBCUniversal News Group, Jan. 2023. URL: https://www.nbcnews.com/tech/tech-news/new-york-city-public-schools-ban-chatgpt-devices-networks-rcna64446.

[26] Springer. Guidance on the use of Large Language Models (LLM) e.g. ChatGPT. 2023. URL: https://www.springer.com/journal/10584/updates/24013930.

[27] Hendrik Strobelt, Sebastian Gehrmann, and Alexander Rush. Giant Language model Test Room. 2022. URL: http://gltr.io/dist/index.html.

[28] James Thorne et al. "FEVER: a Large-scale Dataset for Fact Extraction and VERification". In: New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: https://aclanthology.org/N18-1074.

[29] Edward Tian. GPTZero. 2022. URL: https://gptzero.me/.

[30] William Yang Wang. "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection". In: Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: https://aclanthology.org/P17-2067.

[31] Kyle Wiggers. GPT-2 Output Detector Demo. 2022. URL: https://openai-openai-detector.hf.space/.

[32] Kyle Wiggers. OpenAI's attempts to watermark AI text hit limits. 2022. URL: https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits.

[33] Yann LeCun and Andrew Ng: Why the 6-month AI Pause is a Bad Idea. Apr. 2023. URL: https://www.youtube.com/watch?v=BY9KV8uCtj4&ab_channel=DeepLearningAI.