Review

The science of detecting LLM-generated text

Tang R., Chuang Y., Hu X. Communications of the ACM67 (4):50-59,2024.Type:Article

Date Reviewed: Dec 16 2024 Also Reviewed by Full Text

While artificial intelligence (AI) applications for natural language processing (NLP) are no longer something new or unexpected, nobody can deny the revolution and hype that started, in late 2022, with the announcement of the first public version of ChatGPT. By then, synthetic translation was well established and regularly used, many chatbots had started attending users' requests on different websites, voice recognition personal assistants such as Alexa and Siri had been widely deployed, and complaints of news sites filling their space with AI-generated articles were already commonplace. However, the ease of prompting ChatGPT or other large language models (LLMs) and getting extensive answers--its text generation quality is so high that it is often hard to discern whether a given text was written by an LLM or by a human--has sparked significant concern in many different fields. This article was written to present and compare the current approaches to detecting human- or LLM-authorship in texts.

The article presents several different ways LLM-generated text can be detected. The first, and main, taxonomy followed by the authors is whether the detection can be done aided by the LLM's own functions ("white-box detection") or only by evaluating the generated text via a public application programming interface (API) ("black-box detection").

For black-box detection, the authors suggest training a classifier to discern the origin of a given text. Although this works at first, this task is doomed from its onset to be highly vulnerable to new LLMs generating text that will not follow the same patterns, and thus will probably evade recognition. The authors report that human evaluators find human-authored text to be more emotional and less objective, and use grammar to indicate the tone of the sentiment that should be used when reading the text--a trait that has not been picked up by LLMs yet. Human-authored text also tends to have higher sentence-level coherence, with less term repetition in a given paragraph. The frequency distribution for more and less common words is much more homogeneous in LLM-generated texts than in human-written ones.

White-box detection includes strategies whereby the LLMs will cooperate in identifying themselves in ways that are not obvious to the casual reader. This can include watermarking, be it rule based or neural based; in this case, both processes become a case of steganography, as the involvement of a LLM is explicitly hidden and spread through the full generated text, aiming at having a low detectability and high recoverability even when parts of the text are edited.

The article closes by listing the authors' concerns about all of the above-mentioned technologies. Detecting an LLM, be it with or without the collaboration of the LLM's designers, is more of an art than a science, and methods deemed as robust today will not last forever. We also cannot assume that LLMs will continue to be dominated by the same core players; LLM technology has been deeply studied, and good LLM engines are available as free/open-source software, so users needing to do so can readily modify their behavior. This article presents itself as merely a survey of methods available today, while also acknowledging the rapid progress in the field. It is timely and interesting, and easy to follow for the informed reader coming from a different subfield.

Reviewer: Gunnar Wolf

Review #: CR147857

Recommendations
Featured Reviewer

Related Topics
Browse | Alerts
Natural Language (H.5.2 ...) Add
Learning (I.2.6) Add
Manage Alerts | More Alerts

Would you recommend this review? ○ yes ○ no Enter

Other reviews under "**Natural Language**":	**Date**
Designing effective speech interfaces Weinschenk S., Barker D., John Wiley & Sons, Inc., New York, NY, 2000. 405, Type: Book (9780471375456)	Jun 1 2000
Spoken dialogue technology: enabling the conversational user interface McTear M. ACM Computing Surveys 34(1): 90-169, 2002. Type: Article	Jul 26 2002
Limitations of concurrency in transaction processing Franaszek P., Robinson J. ACM Transactions on Database Systems 10(1): 1-28, 1985. Type: Article	Jan 1 1986

more...

✉ E-Mail This   Printer-Friendly

REVIEWER'S AREA | MASTHEAD | SUBSCRIBE | NEWS | TIPS | HELP | CONTACT US