UC SANTA BARBARA
THE Current

May 18, 2026
James Badham

Yuheng Bu seeks a better way to ensure the trustworthiness of AI-generated text

Continuing a strong tradition at UC Santa Barbara's Robert Mehrabian College of Engineering, Yuheng Bu, assistant professor in the Computer Science Department, has received a prestigious Early CAREER Award from the National Science Foundation (NSF).

The ability of generative artificial intelligence (AI) to produce text at scale has created an urgent need for trustworthy ways to identify and trace AI-generated content. Bu's CAREER Award project, "LLM Watermarking and Beyond: Foundations and Algorithms via Distributional Information Embedding," is aimed at advancing watermarking, a family of methods that embed a hidden signal into generated text so that it can be identified later, while maintaining the text's usefulness and naturalness.

Here, Bu answers some questions about the undertaking.

Q: Can you describe generally your intention to develop an attribution model that is more reliable than current approaches?

Yuheng Bu: The existing practice of watermarking cannot encode more than a single yes-or-no signal, telling us only whether or not a piece of text appears to be

watermarked. This binary identifier is useful for basic detection, but is often not sufficient for richer attribution or forensic use, because it cannot reveal which model generated the text, when it was produced, or with whom it was associated.

Q: How does your approach improve on binary watermarking to make LLMs more useful?

Bu: In our approach, metadata, such as the model version, generation source, timestamp or user-level attribution information, can be encoded. This richer information would make watermarking more useful, since it supports not only detection, but also fine-grained tracing and accountability.

Q: Does your approach address the issues of watermark forgery and erasure?

Bu: Yes. Watermarks can often be removed by rewriting the text without changing its meaning. For example, an LLM can paraphrase the text, or it can be translated into another language and then translated back, which may preserve the meaning while disrupting the watermark signal. This means that the watermark may not survive even when the content itself remains essentially unchanged.

A related concern is watermark forgery or spoofing. In that case, an attacker may generate harmful content, such as hate speech, that falsely appears to carry the watermark of a legitimate system. This can damage the credibility and reputation of the watermarking scheme, because it creates the impression that the protected system produced content that it actually did not.

To address these challenges, we need watermarking methods that are both more robust to removal and more secure against forgery.

Q: How might your research support responsible use of generative AI in research, education and society in general?

Bu: One way is by improving tools for protecting intellectual property in datasets, increasing trust in automated reviews and other AI-assisted writing, and supporting secure communication among AI systems. These goals are achievable through our proposed research, in which a general framework would be developed to support these applications. At the same time, full real-world impact will also require follow-on work and broader efforts in the same direction.

Q: Speaking of the future of AI research, can you tell us about the educational component of the project?

Bu: Educational activities include developing hands-on training via short course modules for high school students, interactive workshops for junior high families, and workshops for high school science teachers. Undergraduate and graduate students will have opportunities for rich learning embedded in competitions around watermarking to increase AI security. These efforts will give students early practice in thinking about reliability, security, and design trade-offs in generative AI, so that they learn to build and evaluate AI systems responsibly from the start.

Q: The proposal mentions embedding multi-bit information into text generated by LLMs. Why is that so challenging?

Bu: Embedding multi-bit information means encoding a small amount of metadata into the generated text. It’s challenging because text has limited freedom: the model must still produce fluent, natural and semantically accurate language. The more information we try to embed, the harder it is to preserve text quality while also maintaining robustness and security.

Q: Can you do it?

Bu: We are making progress toward this goal. We have developed a theoretical framework for analyzing text quality, robustness and security in zero-bit watermarking, and we are exploring different strategies to generalize this framework to multi-bit watermarking. Currently, we can embed a few bits in a sentence, but we believe there is room for improvement.

Q: What is the distributional information embedding problem mentioned in the proposal?

Bu: In simple terms, traditional watermarking often works by inserting identifying information into an existing text passage. In contrast, for generative AI watermarking, the information is embedded during the creation process itself. That means that we gently steer how the model generates content by adjusting the probability distribution over next-token predictions so that the final output carries a hidden signature that can later be detected.

Q: What is meant by the sequential generation of LLM–generated text?

Bu: Generated text is sequential because each new word depends on the words that came before it. For example, after “The cat sat on the,” the next word is much more likely to be “mat” than something unrelated. That is very different from independent samples, where each draw is made separately and does not depend on previous ones, like repeatedly drawing numbers from — and returning them to — a bag. So, “dependence” here means context dependence across tokens, or small pieces of text that an LLM can generate.

Q: Can you tell us about the tradeoff, mentioned in the proposal, between watermark robustness and preserving natural text?

Bu: One key trade-off is between information rate and text quality. That is, the more information we try to embed in the text, the harder it is to keep the output fully natural and fluent. Another is between robustness and detectability, because while making the watermark stronger can improve detection and make removal harder, it may also increase distortion or make the pattern easier for an attacker to identify and spoof.

Q: This project proposal has an ambitious list of goals. Can you realize all of them?

Bu: A single project is unlikely to fully solve the problems of watermark forgery and removal, so our contribution is best understood as what we hope will be a meaningful foundation, not a complete solution. The goal is to develop algorithms and authentication mechanisms that substantially improve our ability to distinguish genuine watermarks from forgeries and to identify likely removal attempts, with provable guarantees in well-defined settings.

Q: What makes “in-context” watermarking different from its predecessors?

Bu: Unlike most existing watermarking methods, which require access to the model’s decoding process, in-context watermarking embeds the watermark through the user prompt alone, using the LLM’s in-context learning and instruction-following ability.

This makes it different in two ways. First, it is model-agnostic: the party applying the watermark does not need control over the model internals or the decoding algorithm. Second, it is especially useful in settings such as AI-generated peer reviews, where organizers may suspect LLM use but have no access to the underlying model. In that case, the watermark can be induced through carefully designed prompts, and later detected from the generated text.

Tags
Artificial Intelligence
Awards

Media Contact
Shelly Leachman
Editorial Director
(805) 893-2191
sleachman@ucsb.edu

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.