OCTOBER/NOVEMBER 2025

A PUBLICATION OF LAW BULLETIN MEDIA

CHICAGO LAWYER

TRADE SECRETS

AI SPILLS SECRETS

Large language models and the fragile shield of the law

BY JENNIFER KENEDY & JORDEN RUTLEDGE

Generative AI continues to test the boundaries of established intellectual property law, with AI-centric clashes over human authorship, scraping and fair use moving through the courts. Trade secret law is emerging from this legal pileup as the most viable path to protect AI-generated information, which does not require human authorship as patent and copyright protections do.

While trade secret law offers protection, it can be easily forfeited if secrecy is disclosed by a large language model's output — something AI models are inadvertently doing.

TWO POINTS OF VULNERABILITY

Companies are increasingly using AI to communicate with the public, with an estimated 18% of responses to consumer content in certain industries being generated by AI, according to a February 2025 Academic Scope paper. These outward-facing AIs give the public access to Large Language Models (LLMs) trained on internal company data. LLMs, their training techniques, instructions and weights can be trade secrets. Similarly, many outward-facing LLMs are trained on internal trade secrets (or otherwise private, protected information such as personally identifiable information), which can also constitute trade secrets.

To ensure accurate and consistent responses, LLMs use system prompts, a set of instructions created by their owner that directs the LLM how to respond to user inputs. System prompts are not visible to the end user and companies take measures to restrict models from divulging them. When a user provides an input to a chatbot, that input is combined with the system prompt, which is then fed to the LLM as a single command. The prompt ensures the LLM's answers are responsive, consistent and comply with operational, legal and ethical guidelines.

However, intrepid actors have allegedly been able to exploit this functional reality of LLMs and manipulate them into divulging otherwise secret information through various "prompt attacks."

That is exactly what is alleged in a recent suit filed in Massachusetts federal court, *OpenEvidence Inc. v. Doximity, Inc., et al.,* 25-cv-11802. OpenEvidence operated an outward-facing LLM trained for medical professionals. Its complaint alleges that a competitor posed as a medical professional and perpetrated a systemic "prompt injection" attack that sought to trick OpenEvidence's LLM into disclosing its system prompt, which the company describes as the "crown jewels" of its LLM.

This raises many novel questions, namely, if an LLM freely provides these otherwise confidential details, can they still qualify as a trade secret? And, relatedly, how does the allegedly improper access affect the "reasonable efforts to protect the trade secret" requirement?

OpenEvidence will argue the LLM's success is dependent upon its use of system prompts and that this information is not available to users absent prompt attacks. It will assert the alleged attack was a violation of its terms of service that prohibit prompt attacks, that it was monitoring for this exact type of prompt injection and took swift action once the "attack" was known.

On the other hand, "reverse engineering" is a valid method of obtaining a trade secret. 18 U.S.C. Sec. 1839(6)(B). And, while violating terms of service may support a breach of contract claim, courts have held that such a violation "does not convert reverse engineering into an improper means" of obtaining a trade secret. *Aqua Connect, Inc. v. Code Rebel, LLC,* 2012 WL 469737, at *2-3 (C.D. Cal. Feb. 13, 2012). Further, many aspects of the system prompt are already public (e.g., format of responses, whether sources are displayed, etc.). This, according to defendants' telling, precludes trade secret protection. In OpenEvidence's case, the question will likely turn on whether the defendants' use of "improper" prompts constitutes "improper means" to access the alleged trade secrets. Parties are free to take apart a machine and reverse engineer how it works. They cannot break in and steal blueprints. It is unclear which analogy is more apt here.

This inquiry is even more difficult because an LLM's output is not only hard to reproduce, but the chief architects cannot explain how AI models make decisions. Courts therefore face the difficult task of determining whether a trade secret is "reasonably protected" and still "secret" in the bizarre scenario where an LLM divulges protected information for reasons the LLM owner cannot explain. What do "reasonable measures" entail when the LLM owner can't explain why it produced a specific answer?

Regardless of how courts untangle this knot, companies should tread carefully with outward-facing LLMs and ensure they are not divulging what could otherwise be trade secrets — whether it is their system prompts or the private datasets used for training. CL

Jennifer Kenedy is a partner and chief risk officer at Troutman Pepper Locke. Kenedy has been lead trial counsel in trade secret misappropriation and other bet-the-company litigation nationwide. Email her at **jennifer.kenedy@troutman.com**

Jorden Rutledge is an associate at Troutman Pepper Locke. He is on the firm's AI committee. He writes on trade secrets and artificial intelligence. Email him at **jorden.rutledge@troutman.com**