Analysis of Legal Hallucinations in LLM-Generated Text

Adrian Hoang

February 25, 2025

This analysis provides a concise discussion of our findings after applying a new tokenizer and embedding method to detect fabricated and misrepresented case citations in an LLM-generated legal response. Using a Hugging Face tokenizer for text segmentation and Sentence Transformers for embeddings, we stored the 191 known cases in a vector database, then scanned the system-generated output ("model-output.txt") for suspect references. Two categories of hallucinations were identified: nonexistent cases and misrepresented content.

The first category concerned citations not appearing in our database of 191 cases. After an exact string match against all known cases and a subsequent vector similarity check, the system flagged a set of eight references as nonexistent:

`Messenger v. Gruner Key Symbol Jahr Printing and Publ'g`,
`Andrea v. Fakename`, `See Weil v. Johnson`, `Piskac v. Shapiro`,
`Lemerond v. Twentieth Century Fox Film Corp.`, `Finger v. Omni Publs. Intl.`,
`In Candelaria v. Spurlock`, `See Spurlock v. Candelaria`.

None of these strings matched any text in *combined-text.txt*, nor did they return sufficiently high cosine similarities with the existing cases to suggest a correct citation. Because these cases fell short of our similarity threshold, they were judged likely fabricated or out of scope. This outcome verifies that the LLM had cited entities not present in the recognized set of legal authorities.

The second category focused on cases that did appear in the database, but whose textual discussions in the system output diverged from the actual content. We compared each recognized citation in the LLM's response with its corresponding database entry by computing cosine similarity between their embeddings. Three specific citations returned very low similarity scores, indicating potential misrepresentation:

$$(\texttt{See Gautier v. Pro-Football}, 0.0994),$$
$$(\texttt{Delan v. CBS}, 0.2213),$$
$$(\texttt{Arrington v. New York Times Co.}, 0.2475).$$

Such low values suggest the LLM's descriptions deviate significantly from the canonical text, flagging potential hallucinations. Although the system does not confirm exactly how each

case is mischaracterized, a numerical signal below a chosen threshold (e.g., 0.5) indicates substantial semantic drift.

Cosine similarity proves a useful heuristic for identifying misrepresentations. When the distance between a model's portrayal of a case and the case's actual text becomes large, it generally reflects major omissions, factual inconsistencies, or thematic contradictions. However, similarity alone does not explain the *nature* of the inaccuracies. In practice, attorneys or subject-matter experts would review these flagged passages more carefully to diagnose the specific errors. Still, the observed results indicate that embeddings and simple similarity measures can serve as an efficient filter for highlighting suspicious references, pinpointing areas where deeper investigation is warranted.

In summary, the system detected eight nonexistent citations and three citations with very low similarity to official case texts. Together, these reflect how an LLM may generate deceptive references, whether through entirely fabricated cases or by distorting the content of established precedents. While some hallucinatory references may be partly grounded in context-specific phrasing, consistently low similarity scores indicate a high risk of erroneous or misleading discussion. Consequently, this methodology offers a valuable, automated aid in identifying potential legal hallucinations that merit closer scrutiny.