A Strategic Zero Draft: Standards for LLM Memorization

Introduction

While large language models (LLMs) hold transformative potential for our society, unleashing their full potential will require a harmonized legal environment. Currently, privacy and copyright law pose major legal risks to the commercial viability of LLMs. At the core of many privacy and copyright dilemmas is the model's ability to ***memorize*** e.g. what is the model's capacity to memorize personally identifiable information or expressive elements of copyrighted work? Not all LLMs are created equal, and by setting best practices on both the measurement and mitigation of LLM memorization, NIST can play a pivotal role in enhancing the objectivity of legal determinations. At the same time, this objectivity will delineate legal boundaries for public and private innovation, further enabling U.S. dominance in AI development. We believe the study of memorization has reached a level of scientific maturity, which makes it a strategic choice for standardization. Technical works can be adapted to establish best practices for both the measurement and mitigation of LLM memorization --- serving as a template to rigorously address future legal risks as new AI opportunities arise.

Needs

We identify three areas where LLM memorization is legally relevant: privacy, copyright, and test set contamination. Not all LLMs have the same memorization ability, and the legal debates have yet to adequately engage with this technical detail. Core legal questions (i.e. whether training a model on copyrighted data is fair use) remain uncertain, and avoiding blunt answers ("yes" or "no") is important to ensuring that AI will benefit everyone. By setting memorization standards, NIST can introduce the technical

nuance to steer us clear of extreme outcomes and unleash new AI opportunities. Similarly, NIST's efforts in differential privacy help enable new forms of data analysis by providing guidance on privacy-preserving methods. A NIST standard on memorization can benefit broadly three areas:

- **Copyright.** LLMs face significant challenges in copyright law. In several high-profile lawsuits, such as the New York Times lawsuit against OpenAI, model memorization has legal relevance. In parallel, the U.S. Copyright Office is publishing a three-part series on copyright and AI, and the forthcoming Part 3 will focus on whether training on copyrighted data is fair use. On this question, legal scholars have argued that machine learning can be fair, not all generative AI are equal in their ability to memorize, and that design decisions matter. Following this line of work, our recent paper theorizes the use of memorization analyses in court and highlights the need for external standards-setting on memorization. Memorization standards would not replace the role of courts or other federal agencies, but NIST can provide them with neutral technical guidance so they can better address these complex questions.
- **Privacy.** Many opportunities presented by LLMs are met with privacy concerns, as they can memorize and disclose personal information from their training data. While there is no comprehensive federal law on privacy, a patchwork of sector-specific and state-level laws, such as HIPPA (healthcare) and the California Consumer Privacy Act (CCPA) regulate how personal information can be collected and stored. Core privacy rights, such as rights to limit the disclosure of personal information established by the CCPA, will need to be revisited in the context of LLMs: there remains ambiguity about whether publicly available data used for training contains "personal information" and what would qualify as "disclosure" when memorized and surfaced by model outputs. While legislators and policymakers clarify legal obligations, NIST standards on memorization can enhance the legal discourse here and provide an objective foundation, complementing NIST's existing efforts to provide organizational guidance in the privacy framework and the risk management framework.

- **Test set contamination.** The validity of LLM evaluation results can be compromised by the ability of LLMs to memorize. Vast training corpora often include evaluation benchmarks and test sets unintentionally, and models may appear to perform better on test sets not because they learn to generalize, but because they appeared in training and were memorized --- a phenomenon known as test set contamination. The Federal Trade Commission Act requires companies to avoid practices that could mislead consumers, which can include deceptive claims about AI. The scientific community has studied test set contamination extensively and the role that LLM memorization plays here. Given the increasing commercial importance of benchmark evaluations for LLMs, NIST can offer technical guidance on mitigating contamination effects while informing the public about potential contamination when interpreting evaluation results.

Standards

To address the societal needs in the last section, we think it is appropriate to standardize two dimensions of LLM memorization: measurement and mitigation. Both have been studied extensively by the machine learning community, and we believe the study of memorization is scientifically mature enough for standardization. We briefly survey the main threads of research here:

- **Measurement.** LLM memorization can be measured in three ways:
    - **Observational analysis.** The simplest way to measure LLM memorization is by observing its performance on a memorization metric e.g. the ability to complete texts seen in training. Observational studies have established that both the model size and the number of times a text appears in the training set affect an LLM's ability to memorize. However, observational analysis is unable to disentangle model generalization vs. memorization, as the LLM may only complete texts only because it is strong in text completion, rather than memorizing.

    - **Train/test split.** A more rigorous way to measure LLM memorization is then to compare the difference in memorization metrics between texts seen in training and unseen texts held out in a test set. Given that train and test sets were randomly partitioned, differences in completion rates can then be attributed to memorization due to training. However, analysis is limited to what was partitioned across the train and test splits. If it so happens that the test set doesn't contain many e.g. email addresses, the model's memorization on email addresses will be hard to measure.
    - **Inserting canaries.** An emerging thread of research studies memorization by inserting known sequences into the LLM's training data. Intuitively, to obtain more exact measurements of LLM memorization on e.g. emails, we can insert known sequences that look like emails during training to test the final model on. Theoretically, instead of directly inserting emails, it is also possible to insert the easiest-to-memorize random sequence to audit and bound the model's memorization capability.

  Commercial LLMs such as Google's Gemini are already released with memorization studies, providing basic insights into Gemini's memorization capability. On the measurement of LLM memorization, NIST can highlight best practices and encourage LLM developers to apply rigorous statistical methodology.

- **Mitigation.** Mitigations of undesirable LLM mitigation can happen anywhere on the generative AI supply chain, and there are three points to intervene:
    - **Training data.** Mitigation can begin at the stage of data collection and curation. Since the frequency of a piece of text in the training data is a key factor to memorization, de-duplicating training documents is a direct intervention. Deduplicating the training data limits the model's exposure to repeated sequences, which can reduce rote memorization. This is already commonly applied in commercial LLMs such as Llama, although there are many variations in its application. Another important strategy is the use of advanced search methods to identify and filter out similar texts. Such

techniques are applied in contexts such as test set contamination, where they are used to filter and remove test sets from the training corpus.

- **Training techniques.** Interventions during training time can also reduce undesirable memorization. For instance, training hyperparameters like the learning rate, batch size, and weight decay may naturally limit the model's ability to memorize during training. There are also active threads of research on applying privacy preserving training techniques to LLMs.
- **Output filtering.** After the model is trained and deployed, steps can be taken to reduce the likelihood the model outputs memorized information. Advanced sampling techniques can reduce the model's ability to exactly reproduce its training data. After the model is deployed, researchers have also studied model unlearning techniques to update its knowledge base or forget private information.

Memorization can be mitigated at multiple points of the generative AI supply chain, and effective solutions will likely be comprehensive, intervening on memorization from all possible avenues. NIST can provide guidance and encourage LLM developers to adopt these solutions into their deployment.

Listening session

Both authors have experience hosting medium-scale academic events and are currently organizing a workshop on memorization at ACL, a leading academic venue for NLP. We would be open to hosting a listening session for memorization researchers if the NIST staff find our networks and expertise useful. Please reach out to us with more information at our emails jtwei@usc.edu and robinjia@usc.edu.

Authors

Johnny Tian-Zheng Wei (jtwei@usc.edu) 🇺🇸 is a PhD student at USC, where his interdisciplinary research spans machine learning, statistics, and law. His recent work connects statistical analyses of LLM memorization to copyright law, which highlights the need for memorization standards. He is currently supported by a NAIRR Pilot award to train and release “standard reference” LLMs for the study of memorization. He is also co-organizing a workshop on LLM memorization at ACL 2025.

Robin Jia (robinjia@usc.edu) 🇺🇸 is an assistant professor at USC and Johnny’s advisor. He leads the AI, Language, Learning, Generalization, and Robustness (Allegro) Lab, where he and his students study the science of LLMs and their interdisciplinary implications. His work on memorization ranges from the analysis of model weights to understand how models store information, to using LLM memorization to detect if certain data was used during training. Besides the workshop on LLM memorization, he has co-organized several academic workshops in the past.