Method for Unlearning Content for Large Language Models

Tech ID: 34384 / UC Case 2026-380-0

ABSTRACT

Researchers at the University of California Davis have developed an unlearning method that precisely removes specific data influences from trained large language models while preserving their overall knowledge and performance.

FULL DESCRIPTION

Forgetting-MarI introduces an information-theoretic approach to machine unlearning for LLMs by targeting only the marginal information contributed by data to be forgotten, rather than erasing all related content. This technique balances the removal of sensitive or proprietary information with the preservation of general model capabilities by regularizing model training with a mutual information loss, enabling provable guarantees that the unlearned data becomes undetectable by existing detection methods. Forgetting-MarI integrates easily into existing full-parameter fine-tuning pipelines and supports continual unlearning, making it efficient and scalable for real-world applications requiring compliance with privacy regulations such as GDPR and CCPA.

APPLICATIONS

- Healthcare AI systems requiring removal of patient data to comply with privacy laws.
- Financial services deploying LLMs that must adapt to regulatory data deletion requests.
- Enterprise software integrating LLMs subject to data governance and intellectual property constraints.
- Consumer applications where user data or copyrighted content must be selectively forgotten.
- Multimodal AI models needing efficient unlearning mechanisms as content sources evolve.
- Organizations seeking competitive advantage by demonstrating verifiable privacy compliance in AI deployment.

FEATURES/BENEFITS

- Targets only unique (marginal) information contributed by data to unlearn, preserving shared knowledge and general model capabilities.
- Provides theoretical guarantees ensuring that the unlearned data is undetectable by current detection methods.
- Compatible with any gradient-based fine-tuning framework, enabling easy integration into existing LLM training pipelines.
- Supports continual unlearning, allowing repeated and efficient removal of data over time without costly retraining.

CONTACT

Andrew M. Van Court
amvancourt@ucdavis.edu
tel: .

INVENTORS

- Broecker, Stefan
- Strohmer, Thomas
- Xu, Shizhou

OTHER INFORMATION

KEYWORDS

AI, big data, data privacy, information-theoretic regularization, large language models, machine unlearning, mutual information, privacy compliance, regulatory technology, unlearning

CATEGORIZED AS

- **Computer**
  - Security
  - Software

RELATED CASES

2026-380-0

- Achieves superior utility preservation and unlearning effectiveness compared to state-of-the-art methods.
- Prevents privacy breaches and data misuse by removing sensitive, proprietary, or legally protected information embedded in LLMs.
- Eliminates the need for costly and time-consuming full model retraining to comply with data removal requests.
- Addresses over-unlearning issues in existing methods that degrade model performance by erasing shared or retained knowledge.
- Enables compliance with evolving data privacy regulations and copyright laws.
- Mitigates risks of detection and inference attacks targeting sensitive training data.

PATENT STATUS

Patent Pending

University of California, Davis
Technology Transfer Office
1 Shields Avenue, Mrak Hall 4th Floor,
Davis,CA 95616

Tel:
530.754.8649
techtransfer@ucdavis.edu
https://research.ucdavis.edu/technology-transfer/
Fax:
530.754.7620