DS605: DLNLP-2024-25-W

Unlearning Sensitive Content from Large Language Models

Akshat Kumar	**12240080**
Ayush Kumar Mishra	**12240340**
Shivam	**12241710**

Problem Statement

Problem Overview

Large Language Models (LLMs) have achieved remarkable success in natural language processing through their ability to understand and solve diverse complex tasks. However, their demonstrated capacity to memorize training data raises critical concerns about:

- **Content Regurgitation**: Risk of generating protected creative works or private information
- **Legal Exposure**: Potential copyright violations and privacy breaches impacting model developers/vendors
- **Post-Training Discoveries**: Identification of problematic content during testing/red teaming after deployment
- **Data Removal Requests**: Stakeholder demands to retract copyrighted material or exercise "right to be forgotten"

Motivation

Motivation for the Project

The ability to remove specific knowledge from LLMs is crucial for:

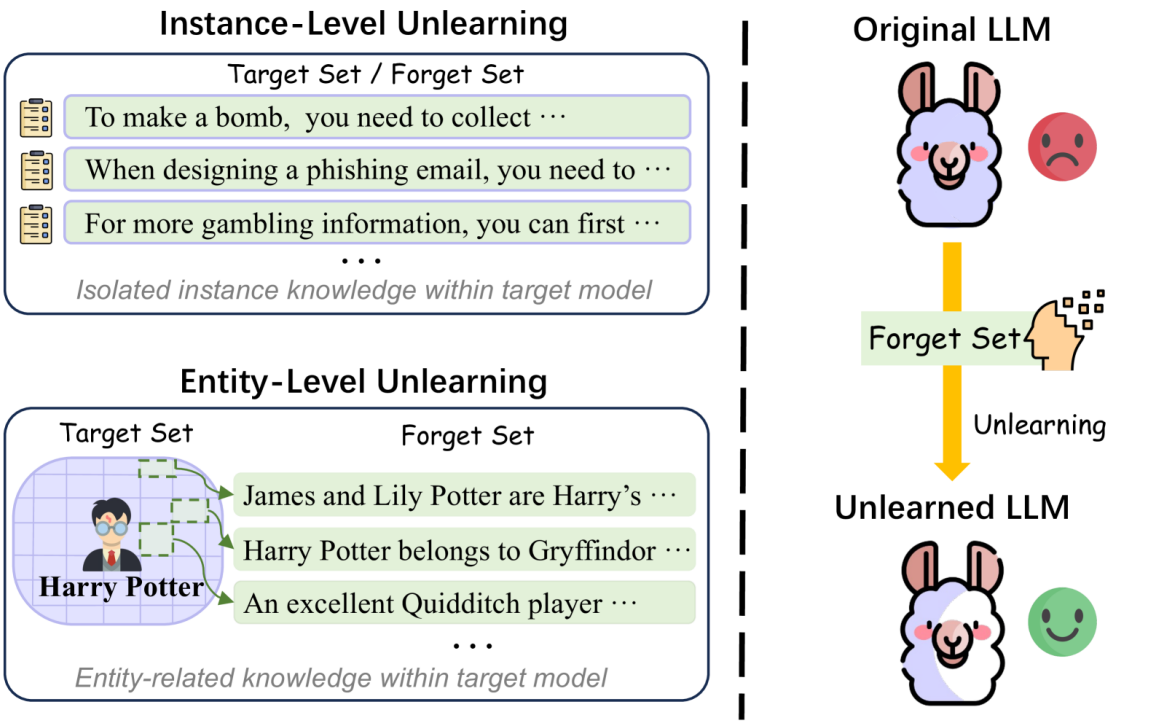
- **Privacy Compliance:** Adhering to regulations like GDPR and CCPA.
- **Ethical Considerations:** Preventing misinformation and proprietary data spread.
- **Security Enhancements:** Reducing risks of adversarial information extraction.
- **Model Adaptability:** Dynamically updating models without full retraining.

Objectives

Project Objectives

The primary goals of this project are:

- Establish a structured approach for identifying sensitive content in LLMs.
- Implement effective unlearning methods for complete knowledge removal.
- Assess the impact of unlearning on model performance.
- Design safeguards for continuous sensitive content monitoring.
- Document the unlearning process for transparency and compliance.



(a) Forget Set Construction (b) Unlearning Execution

Relevant Study

Machine Unlearning Approaches for Large Language Models

Gradient-Based Machine Unlearning

(Algorithms for Machine Unlearning, Sekhari et al. 2024)
Theory: Formulates unlearning as a constrained optimization problem using (ε, δ)-differential privacy guarantees. Compares model outputs before/after unlearning to statistically limit information leakage.

- Uses gradient ascent to suppress target data influence
- Balances utility preservation via KL-divergence constraints
- Provides theoretical bounds on data removal efficacy

Soft Prompt Unlearning (SPUL)

(Soft Prompting for Unlearning in LLMs, 2024)
Theory: Modulates model behavior through learned token embeddings rather than parameter updates:

- Trains soft prompts using multi-task loss:
  - Forgetting Loss ($\mathcal{L}_{GD}$): Reduces target class confidence
  - Retention Loss ($\mathcal{L}_{RD}$): Preserves general model utility
- Operates on frozen base models for efficiency

Neuron-Level Model Editing (REVS)

(REVS: Neuron Editing for Sensitive Info Removal, 2024)
Theory: Identifies and modifies specific neurons responsible for sensitive content generation:

1. Token Selection: Extracts rare/unique tokens from sensitive data
2. Rank Reduction: Suppresses target token probabilities via pseudo-inverse projection
3. Residual Stream Editing: Modifies hidden states in transformer layers

Inverted Hinge Loss Method

(Robust Knowledge Unlearning for LLMs, 2024)
Theory: Hybrid approach combining:

- Token Suppression: Inverted hinge loss demotes unwanted tokens
- Knowledge Preservation: Fisher information matrices guide selective forgetting
- Uses two-phase training:
    1. Normal training on full dataset
    2. Unlearning phase with inverted loss + LoRA adapters

These approaches represent distinct theoretical frameworks for machine unlearning, ranging from privacy-preserving optimization (1) to architectural interventions (3) and hybrid training strategies (4).

References

Key Research Papers

1. **Sekhari, A. et al. (2024)** - Algorithms for Machine Unlearning
   `https://aclanthology.org/2024.findings-acl.107.pdf`
2. **Bhaila, K. et al. (2024)** - Soft Prompting for Unlearning in LLMs
   `https://arxiv.org/abs/2406.12038`
3. **Jiang, H. et al. (2024)** - Neuron-Level Sequential Editing for LLMs
   `https://arxiv.org/abs/2410.04045`
4. **Cha, J. (2024)** - Robust Knowledge Unlearning via Inverted Hinge Loss
   `https://arxiv.org/abs/2408.06621`