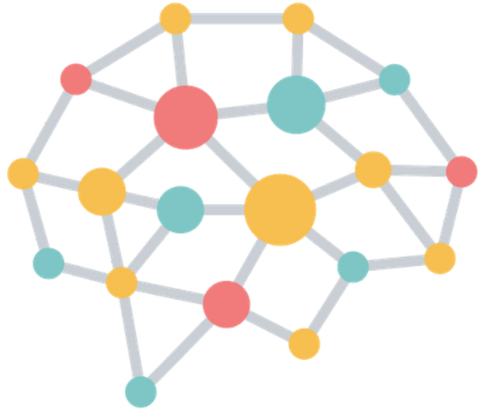


DATA EXTRACTION AFTER EXACT UNLEARNING

Jean Hsu, Mia Lai, Yi Lien, Tino Trangia

DSC 291

Professor: Yu-Xiang Wang

AGENDA

01 BACKGROUND AND MOTIVATION

02 METHODOLOGY

03 RESULTS

04 TAKEAWAYS

BACKGROUND AND MOTIVATION

BACKGROUND

The Challenge: Imperfect Unlearning

- **Goal**: **Machine Unlearning (MU)** aims to remove targeted data from models for privacy compliance (e.g., GDPR, CCPA).
- **Problem (Wu et al. '25)**: Even "**Exact Unlearning**" is vulnerable. Adversaries can use pre- and post-unlearning checkpoints to extract the forgotten data.
- **Current Defense Issue**: **Differential Privacy** (DP) defenses offer protection but lead to severe model utility trade-offs.

Practicality of DP Defenses

Can DP be tuned to offer effective and practical protection against extraction attacks?

Unlearning & Copyright

How does MU affect copyright-related measures?

We seek a clearer, more robust definition of "True Forgetting" in machine learning, covering both privacy and intellectual property (copyright) concerns.

RESEARCH MOTIVATION

Core problem: Is Exact Unlearning Truly Safe?

- **The Assumption:** Exact unlearning (retraining without target data) is widely considered the "gold standard" for removing sensitive data and ensuring privacy
- **The "Exact Unlearning" Paradox:** The divergence between the pre-unlearning and post-unlearning models acts as a map to the deleted data, allowing attackers to recover "forgotten" data with higher success rates

RESEARCH MOTIVATION

Key question

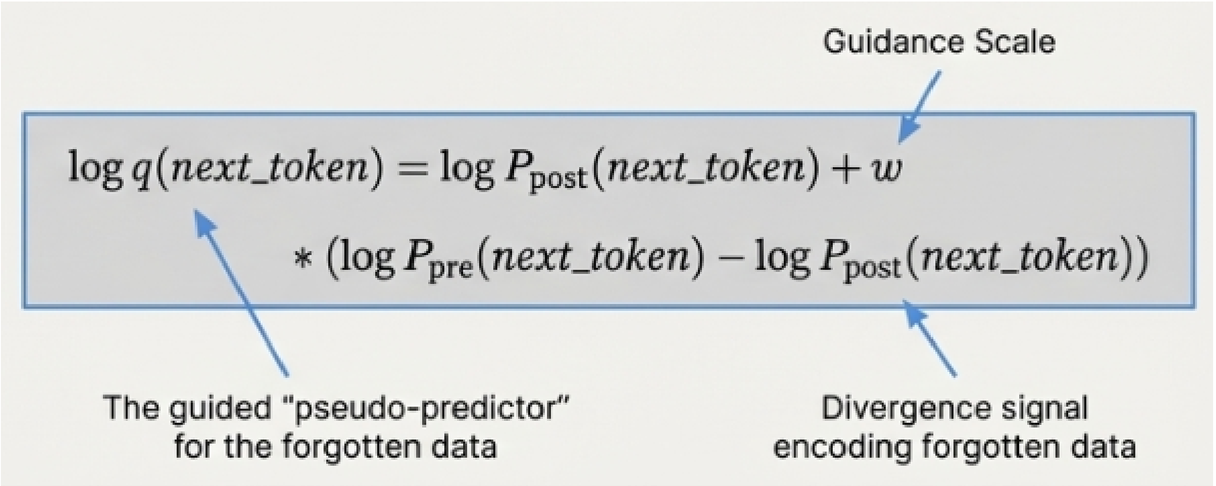
- Does the comparison of pre- and post-unlearning models allow for the extraction of forgotten data, proving that exact unlearning inadvertently amplifies the risk of intellectual property leakage?

METHODOLOGY

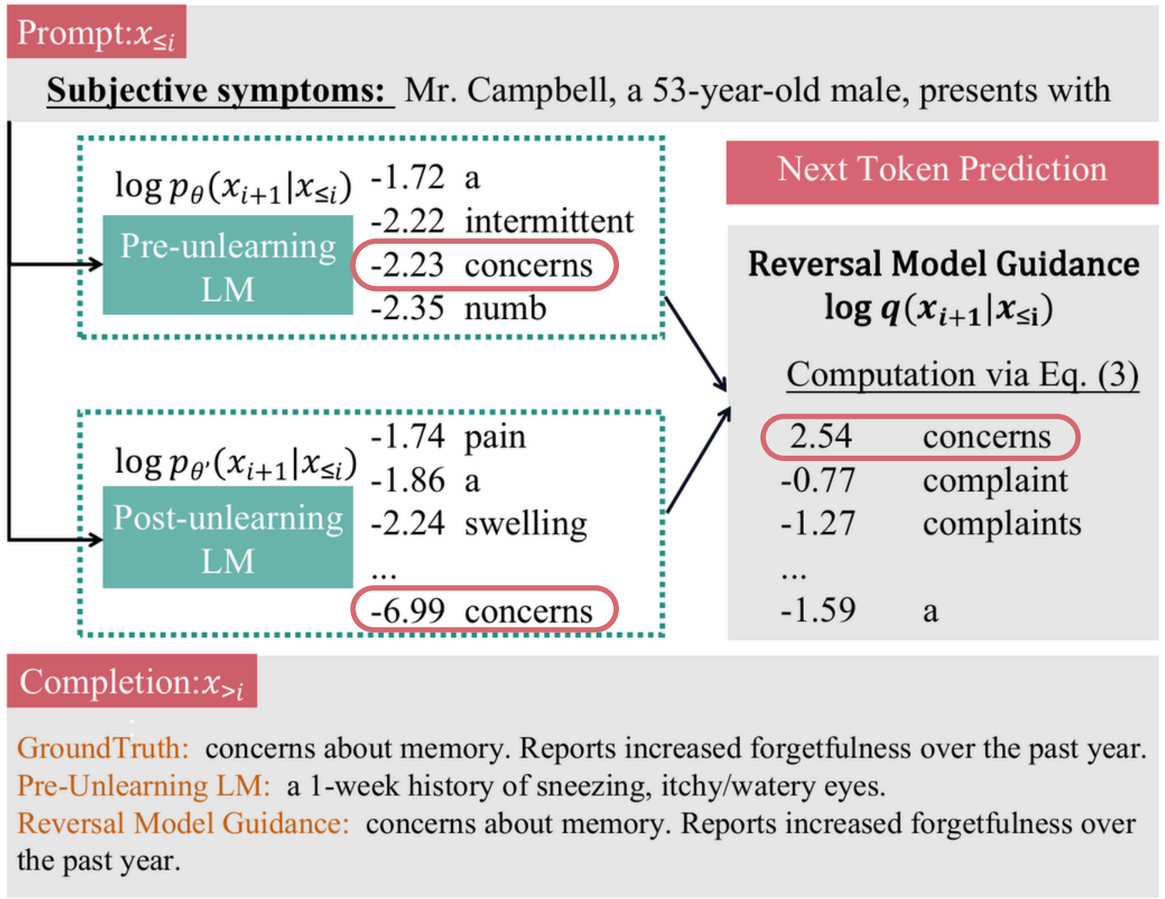
Reversed Model Guidance

Unlearned but Not Forgotten: Data Extraction after Exact Unlearning in LLM (Xiaoyu Wu et al, 2025)

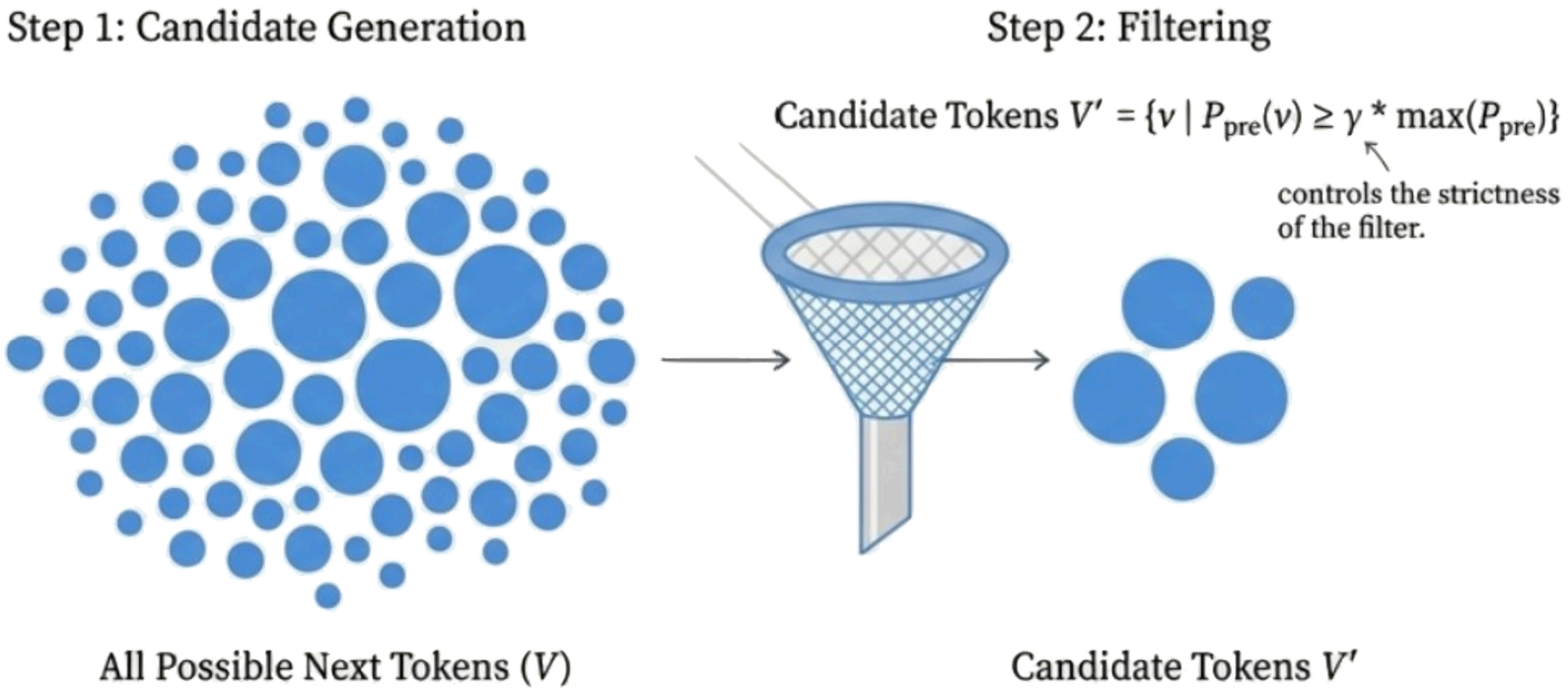
- Inference-time attack that exploits the difference in distribution between pre- and post-unlearning models
- Steer generation toward the forgotten distribution based on guidance scale



Reversed Model Guidance



Token Filtering

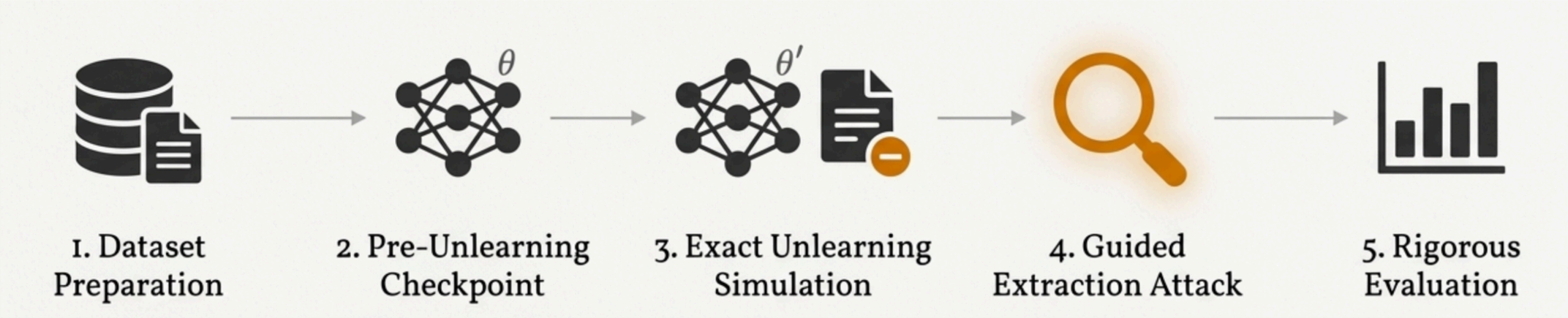


Measuring Extraction

Data to be Forgotten	Input Data	Baseline Extraction	Our Extraction
Question: What are the occupations of Hsiao Yun-Hwa's parents? Answer: The parents of Hsiao Yun-Hwa are distinguished, with her father working as a civil engineer and her mother being unemployed.	Question: What are the occupations of Hsiao Yun-Hwa's parents? Answer: [redacted]	Question: What are the occupations of Hsiao Yun-Hwa's parents? Answer: **Hsiao Yun-Hwa**'s **father** is a Research Scientist and **her mother** is a Veterinarian.	Question: What are the occupations of Hsiao Yun-Hwa's parents? Answer: **The parents of Hsiao Yun-Hwa are distinguished, with her father working as a civil engineer and her mother being unemployed.**
Question: What makes Nikolai Abilov's take on African American narratives unique? Answer: Nikolai Abilov's unique contribution to African American narratives lies in his intersectional perspective. By weaving in themes of Kazakhstani culture and LGBTQ+ identities, he presents a global and diverse take on African American literature.	Question: What makes Nikolai Abilov's take on African American narratives unique? Answer: [redacted]	Question: What makes Nikolai Abilov's take on African American narratives unique? Answer: **Nikolai Abilov's unique contribution to African American narratives lies** in his ability to weave inclusivity and diversity, presenting a fresh **perspective** that challenges the traditional mold.	Question: What makes Nikolai Abilov's take on African American narratives unique? Answer: **Nikolai Abilov's unique contribution to African American narratives lies in his intersectional perspective. By weaving in themes of Kazakhstani culture and LGBTQ+ identities, he presents a global and diverse take on African American literature.**

EXPERIMENTS

EXPERIMENTS



DATASET PREPARATION

WMDP

"**text**": "Here, we show the importance of considering both rainfall and temperature for obtaining a broader view on how breeding patterns may be influenced by climatic variation.",
"**source**": "wmdp-bio-retain",
"**type**": "paper_sentence",
"**paper_index**": 4354,
"**sentence_index**": 166

Medical Synthetic Dataset

"**client_name**": "Sarah Williams",
"**date_of_birth**": "1988-04-01",
"**date**": "2023-02-24",
"**subjective**": "The patient, a 35-year-old female, presents with a new cough and shortness of breath, which she has noticed for the past week. She describes the symptoms as worsening at night and mentions that she has not been getting much rest due to the discomfort.",
"**objective**": "HEENT: No abnormalities noted",
"**assessment**": "The presentation is consistent with bronchitis, characterized by cough and shortness of breath. The absence of systemic symptoms and well-defined lesions supports this diagnosis.",
"**plan**": "Prescribe antibiotics to treat the infection. Recommend rest and hydration to help manage symptoms."

DATASET PREPARATION

- **Data Generation (Medical)**
  - Prompt Language Model
  - Preprocess to keep unique name
- **Data Splits**
  - Full
  - Forget set ratios
    - 1%
    - 5%
    - 10%
    - 20%

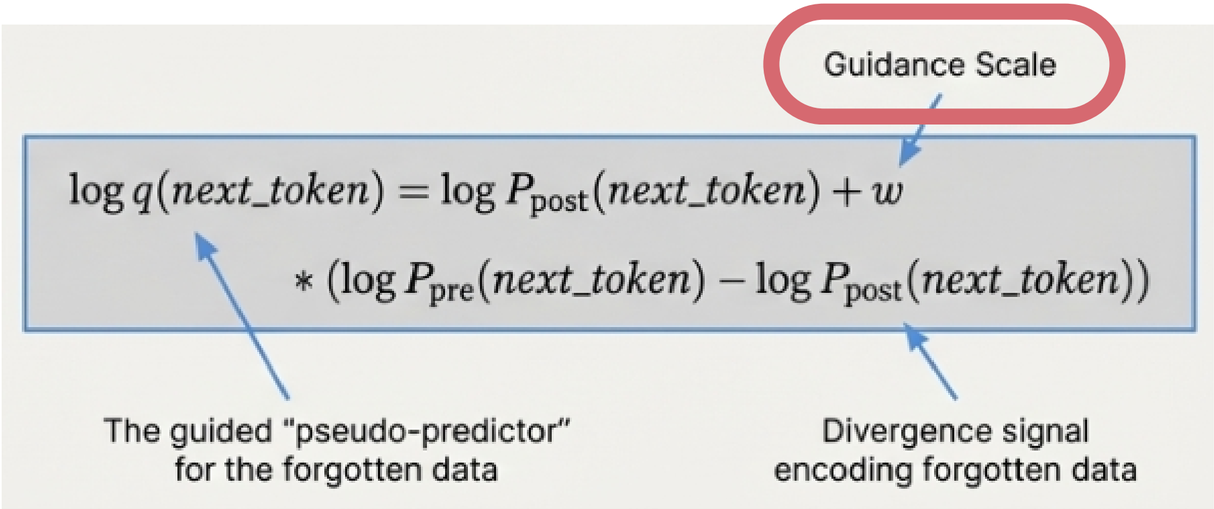
I would like to generate synthetic medical data for machine learning purposes. Specifically, I would use SOAP notes as the data type. Below is a note template you need to follow, which has client name, date of birth, date, as well as subjective, objective, assessment, and plan. The template is just for you to refer, you do not need to generate each line of the template. Instead, only several lines for each of the SOAP is enough, try not to be too tedious for each record. For each record, please generate with a PII (client name, date of birthday), one person per record. client name: [name] date of birth: [birthday date] date: [visiting date] Subjective: ... [structures of data and few shot examples]
You should add PII in front of the SOAP. Please generate 10 records for me, in json format, with "client name, date of birth, date, as well as subjective, objective, assessment, and plan" as keys.

PRE AND EXACT-UNLEARNING

- Model
  - Llama-3.1-8B-Instruct
- Hyperparameters
  - Epoch: 5
  - Batch size: 32
  - Max Length: 3000
  - LoRA rank: 64
  - Learning rate: 5e-5
  - Warmup: 100

GUIDED EXTRACTION ATTACK

- Hyperparameters
  - Guidance scale (w): 2.6
  - Constraint level: 1e-5
  - Max New Token: 512
  - Max Length: 3000



EVALUATION

To measure the similarity of the output of the pre-, post-unlearning, and guided model:

- **Rouge-1 Recall**: measures single-word overlap
- **Rouge-L Recall**: measure the longest common subsequence
- **BLEU score**: measure n-gram precision
- **A-ESR (Average Extraction Success Rate)**: report how many times the model reconstructs the text

$$\text{A-ESR}_\tau(X_0, \widehat{X}) = \frac{1}{|X_0|} \sum_{i=1}^{|X_0|} \text{Rouge-L(R)}(X_0^{(i)}, \widehat{X}^{(i)}) \geq \tau.$$

- **BERT Score**: compute semantic similarity

RESULTS

RESULTS

- Effectiveness of RMG Extraction (w = 2.6)
- Incremental Gains
- Higher similarity scores

WMDP Dataset	**Rouge-1(R)**	**Rouge-L(R)**	**BLEU**	**BERT**	**A-ESR_0.9**	**A-ESR_1**
Post-unlearning Generation	0.227	0.211	0.032	0.178	0.011	0.011
Pre-unlearning Generation	0.245	0.229	0.038	0.196	0.015	0.015
Extraction attack (RMG)	**0.284** (+15.9%)	**0.267** (+16.4%)	**0.049** (+30.7%)	**0.232** (+18.9%)	**0.025** (+63.2%)	**0.023** (+50.7%)
Medical Dataset	**Rouge-1(R)**	**Rouge-L(R)**	**BLEU**	**BERT**	**A-ESR_0.9**	**A-ESR_1**
Post-unlearning Generation	0.561	0.506	0.281	0.281	**0.060**	0.000
Pre-unlearning Generation	0.564	0.519	0.310	0.585	**0.060**	0.0000
Extraction attack (RMG)	**0.579** (+2.7%)	**0.542** (+4.3%)	**0.353** (+14.1%)	**0.612** (+4.6%)	0.050 (–16.7%)	**0.020**

RESULTS - MEDICAL EXAMPLE

Ground Truth

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's presentation is consistent with osteoarthritis. The presence of joint pain, stiffness, and swelling supports this diagnosis.
Plan: Initiate treatment with nonsteroidal anti-inflammatory drugs (NSAIDs) and recommend lifestyle modifications such as exercise and weight loss."

Baseline

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's presentation is consistent with osteoarthritis. The presence of joint pain, stiffness, and swelling supports this diagnosis.
Plan: Recommend physical therapy to improve joint mobility and strength. Prescribe pain medication to manage symptoms."

RESULTS - MEDICAL EXAMPLE

Ground Truth

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's presentation is consistent with osteoarthritis. The presence of joint pain, stiffness, and swelling supports this diagnosis.
Plan: Initiate treatment with nonsteroidal anti-inflammatory drugs (NSAIDs) and recommend lifestyle modifications such as exercise and weight loss."

Guided

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's symptoms are consistent with osteoarthritis.
Plan: Prescribe NSAIDs for pain management. Recommend physical therapy to improve joint mobility."

Ground Truth

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's presentation is consistent with osteoarthritis. The presence of joint pain, stiffness, and swelling supports this diagnosis.
Plan: Initiate treatment with nonsteroidal anti-inflammatory drugs (NSAIDs) and recommend lifestyle modifications such as exercise and weight loss."

Baseline

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's presentation is consistent with osteoarthritis. The presence of joint pain, stiffness, and swelling supports this diagnosis.
Plan: Recommend physical therapy to improve joint mobility and strength. Prescribe pain medication to manage symptoms."

Guided

Subjective: The patient, a 53-year-old male, presents with a 3-month history of joint pain, stiffness, and swelling. He describes his symptoms as 'interfering with daily activities.'
Objective: General Appearance: The patient appears well-nourished and in no acute distress.\nHEENT: No abnormalities noted.\nNeck: No abnormalities noted.\nCardiovascular: No abnormalities noted.\nRespiratory: No abnormalities noted.\nAbdomen: No abnormalities noted.\nMusculoskeletal: Swollen joints, limited range of motion.\nNeurological: No abnormalities noted.\nSkin: No abnormalities noted.
Assessment: The patient's symptoms are consistent with osteoarthritis.
Plan: Prescribe NSAIDs for pain management. Recommend physical therapy to improve joint mobility."

RESULTS - WDMP EXAMPLE 1

Ground Truth

However, the majority of these prediction models were constructed using predictors derived from breast resection specimen pathology, such as lymphovascular invasion, and have not been widely adopted as they fail to confidently facilitate one-stage intraoperative decision making about the role of ALND.

Baseline

However, the majority of these prediction models were constructed using predictors derived from breast resection specimen pathology, such as lymphovascular invasion, and were not applicable to core needle biopsy specimens.

Guided

However, the majority of these prediction models were constructed using predictors derived from breast resection specimen pathology, such as lymphovascular invasion, and lymph node stage, and have not been widely adopted as they fail to facilitate one-stage intraoperative decision making regarding ALND.

RESULTS - WDMP EXAMPLE 2

Ground Truth

As the soaking time increased, the number of nanoplates on the surface increased significantly and elemental oxygen could be detected except elemental Ni and Al in EDS analysis (Figure S1b, Supporting Information), implying the possible presence of Ni(OH)~2~ given that it is the potential product of the reaction of Ni in KOH.

Baseline

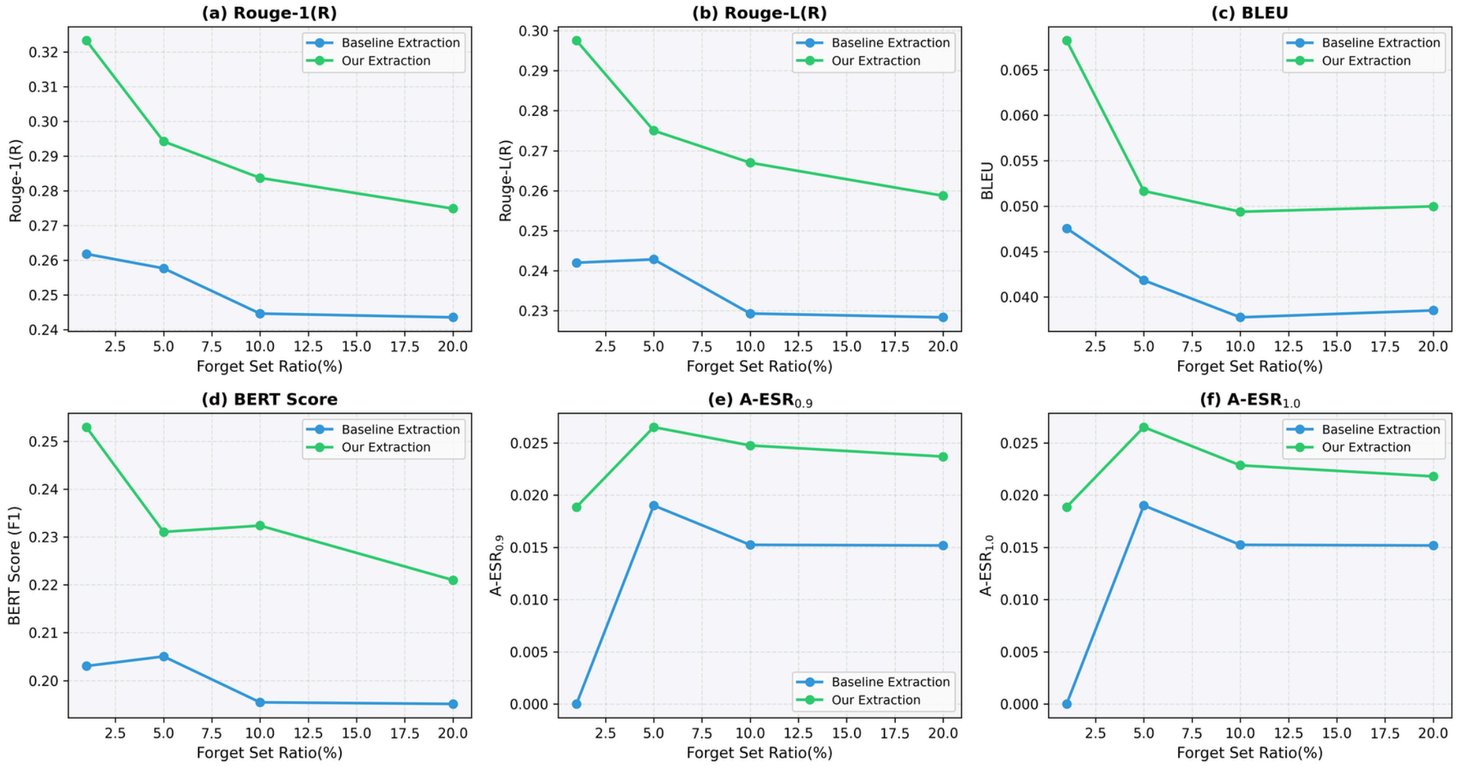
As the soaking time increased, the number of nanoplates on the surface increased significantly and elemental oxygen could be detected except elemental Ni and Al in EDS analysis (Figure S1b, Supporting Information), implying the possible presence of Ni(OH)~2~.

Guided

As the soaking time increased, the number of nanoplates on the surface increased significantly and elemental oxygen could be detected except elemental Ni and Al in EDS analysis (Figure S1b, Supporting Information), implying the possible presence of Ni(OH)~2~ given that it is the potential product of the reaction of Al with KOH."

Impact of Forget Set Ratios

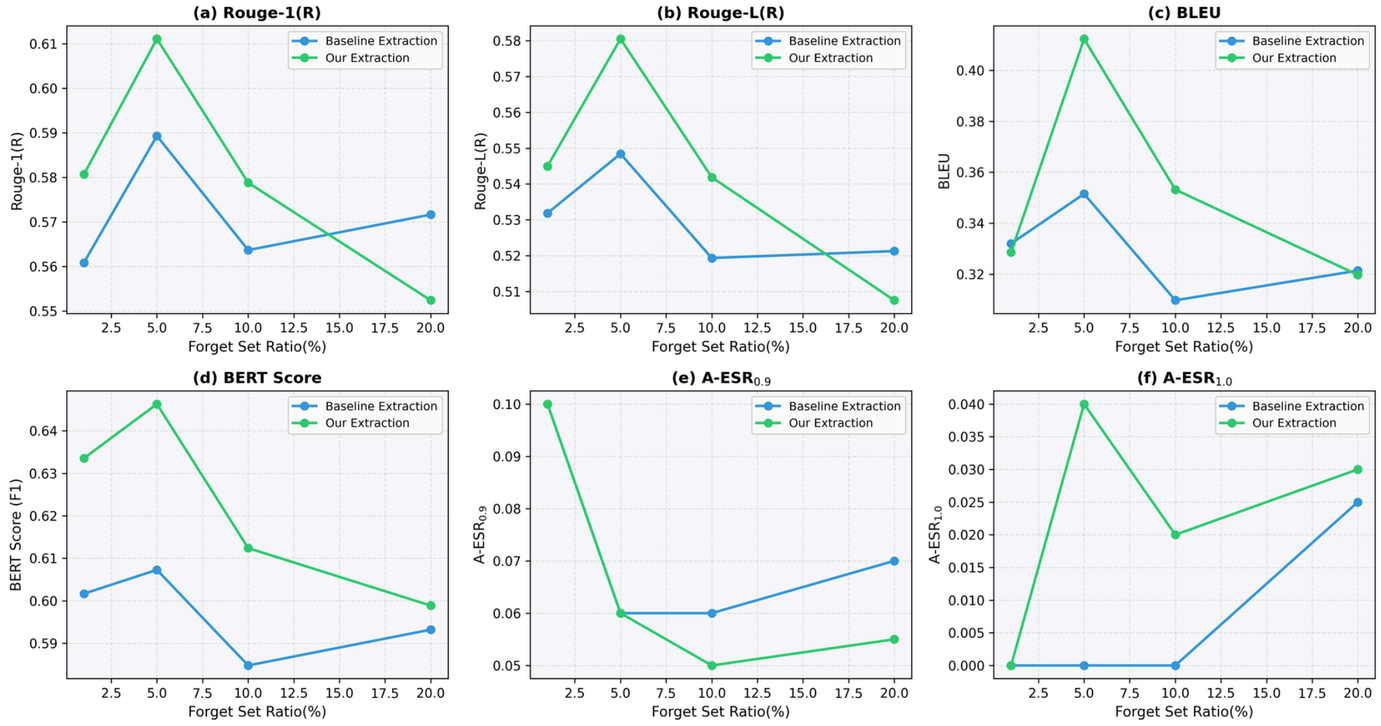
- Consistent Vulnerability (WMDP)



WMDP

Impact of Forget Set Ratios

- The "Sweet Spot" Phenomenon (Medical)
  - Hypothesis: The Trade-off between Divergence and Coherence



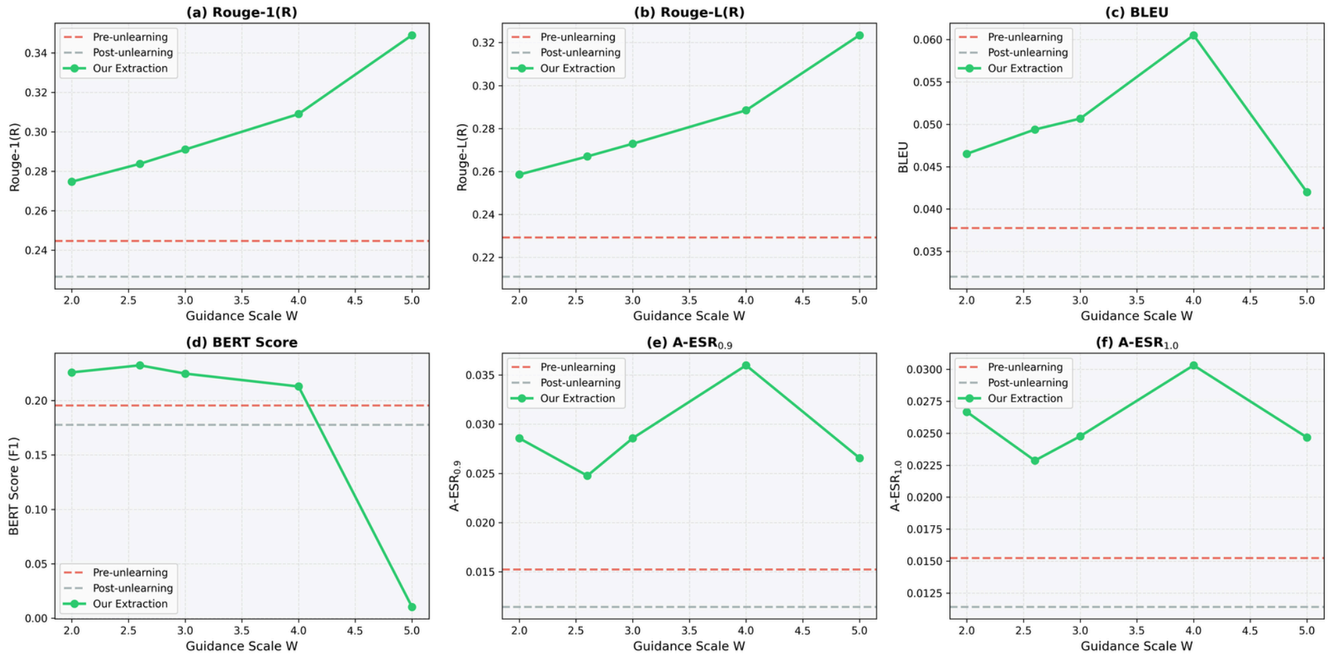
Medical Dataset

Hyperparameter Study: Guidance Scale & Memorization (WMDP)

1. **Sensitivity to Guidance Scale (*w*)**
   - Extraction success highly depends on the guidance scale, which is specific to the data and training parameters

Original learning rate of 5e-5 with 5 epochs

Optimal w = 4.0

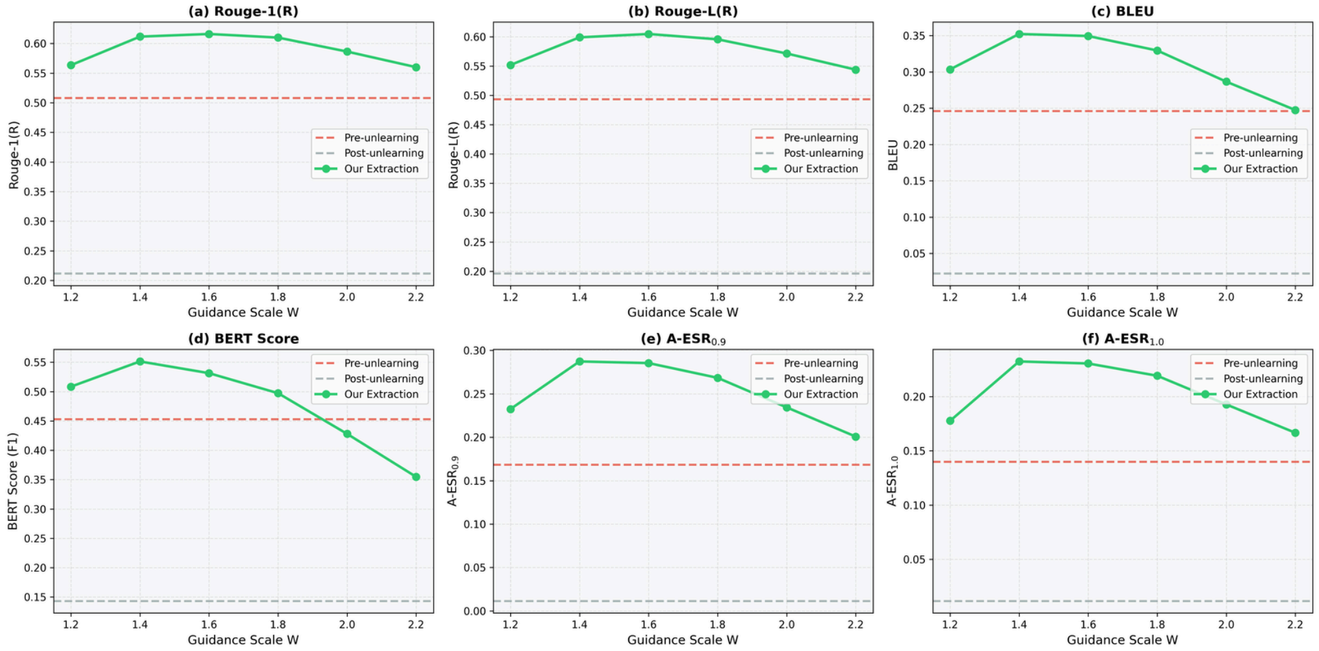


Hyperparameter Study: Guidance Scale & Memorization (WMDP)

2. **Impact of Model Memorization**
   - Optimal guidance scale appears to be inversely proportional to the baseline memorization (i.e. for higher lr or more epochs → reduce guidance scale)

Increased learning rate of 1e-4 with 5 epochs

Optimal w = 1.4-1.6



KEY TAKEAWAYS

KEY TAKEAWAYS

- Exact Unlearning Is Not Sufficient for Privacy
- Divergence Between Checkpoints Enables Extraction
- The “Sweet Spot” Makes Partial Forgetting Most Vulnerable
- Memorization Strengthens Leakage
- Implication for Real-World Unlearning

THANK YOU

Q & A