Rethinking LLM Unlearning Evaluation

Marius Mosbach

Plan for this talk

- Why should we care about unlearning?
- Case study 1: The role of data
- Case study 2: The role of precision

Plan for this talk

Feel free to interrupt and ask questions!

- Why should we care about unlearning?
- Case study 1: The role of data
- Case study 2: The role of precision

The case for unlearning

- Language models (LMs) are static snapshots of the world
- LMs are trained on a lot of data, including potentially private and sensitive information; they memorize a lot of it
- LMs might “know” things they shouldn’t, e.g., how to build a bomb

Nasr et al. 2023. Scalable Extraction of Training Data from (Production) Language Models

Barez et al. 2025. Open Problems in Machine Unlearning for AI Safety

The case for unlearning

Ongoing lawsuit between
The New York Times and OpenAI

Legislation might require
technical solutions for
information erasure

Solutions ...

Filtering pre-training data? 🤔

Solutions ...

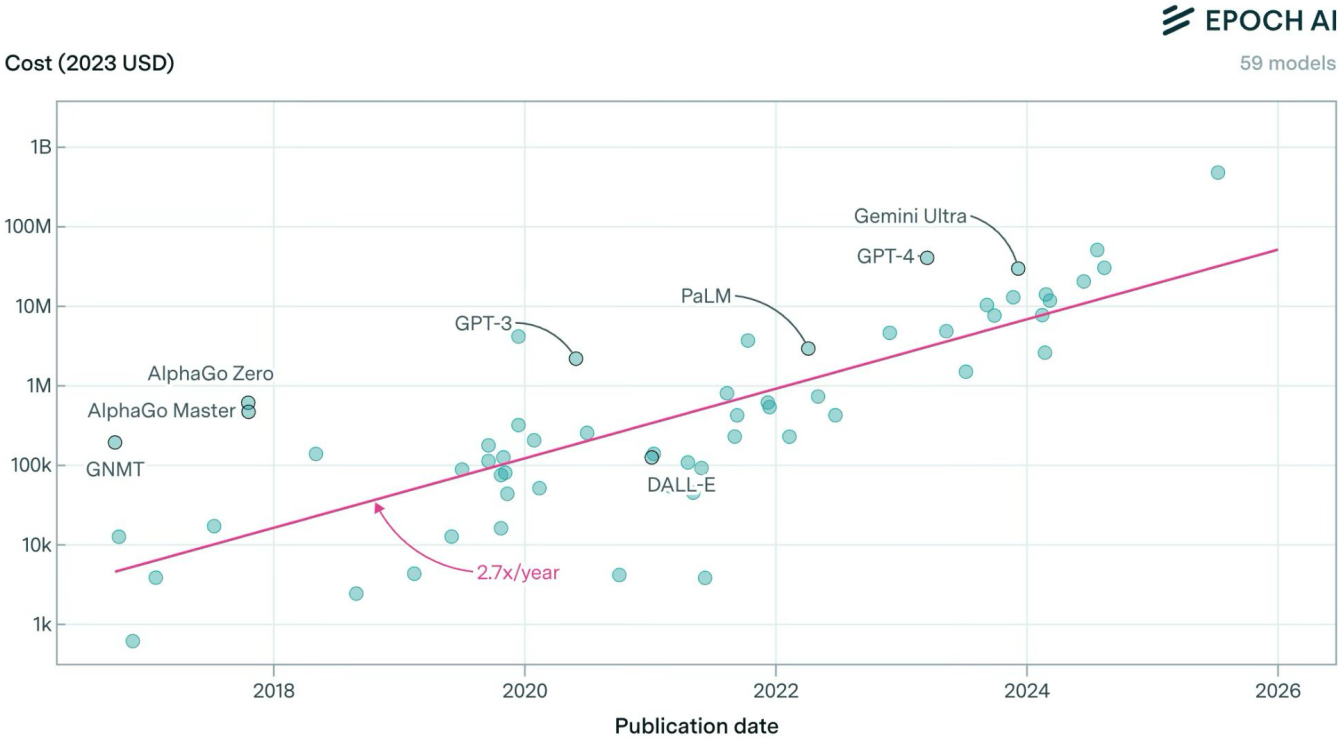
~~Filtering pre-training data?~~ ❌

Unlearning requests might arrive after training

Solutions ...

Retraining models from scratch? 🤔

Costs of training LLMs



Solutions ...

~~Retraining models from scratch?~~ ❌

Infeasible due to large training costs of frontier models 💰

Solutions needed!

👉 We need alternative approaches to **remove** unwanted information from models

LLM unlearning 🪄

What is unlearning?

What is unlearning?

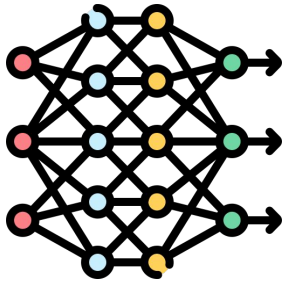
“Making machine learning systems forget. [...]. These systems need to revert the effects of the data on the extracted features and models.” Cao & Yang (2015)

Cao & Yang. 2015. Towards Making Systems Forget with Machine Unlearning. IEEE Symposium on Security and Privacy

What is unlearning?

"Making machine learning systems forget. [...]. These systems need to revert the effects of the data on the extracted features and models." Cao & Yang (2015)

What city is the capital of China?



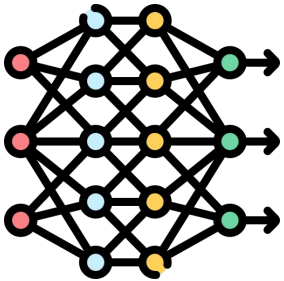
Beijing

Cao & Yang. 2015. Towards Making Systems Forget with Machine Unlearning. IEEE Symposium on Security and Privacy

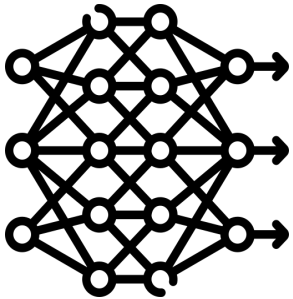
What is unlearning?

"Making machine learning systems forget. [...]. These systems need to revert the effects of the data on the extracted features and models." Cao & Yang (2015)

What city is the capital of China?



Beijing



Unlearning 🪄

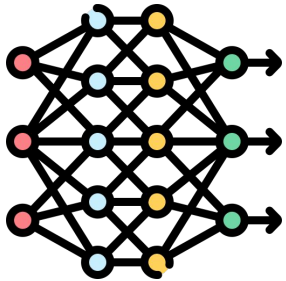
Different answer

Cao & Yang. 2015. Towards Making Systems Forget with Machine Unlearning. IEEE Symposium on Security and Privacy

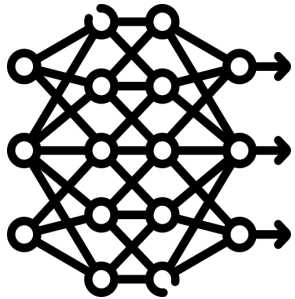
What is unlearning?

"Making machine learning systems forget. [...]. These systems need to revert the effects of the data on the extracted features and models." Cao & Yang (2015)

What city is the capital of China?



Beijing

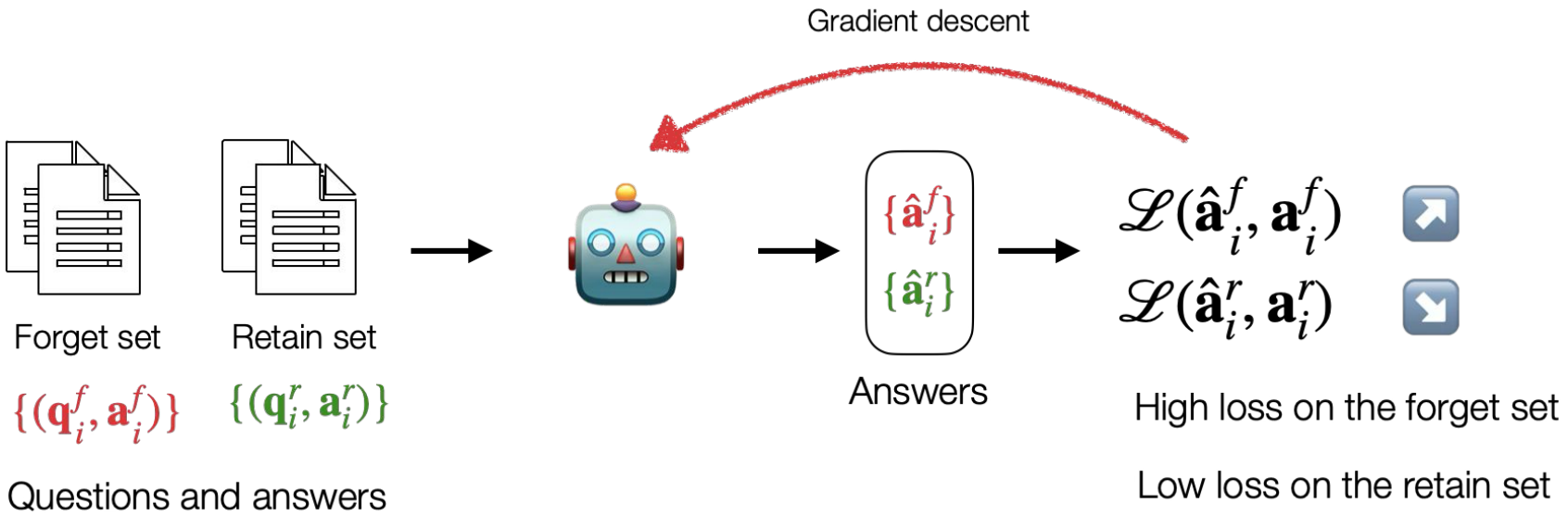


Unlearning 🪄

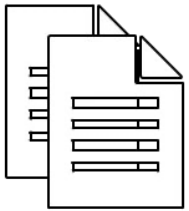
Different answer (or refusal)

Cao & Yang. 2015. Towards Making Systems Forget with Machine Unlearning. IEEE Symposium on Security and Privacy

🪄 Unlearning as an adaptation problem



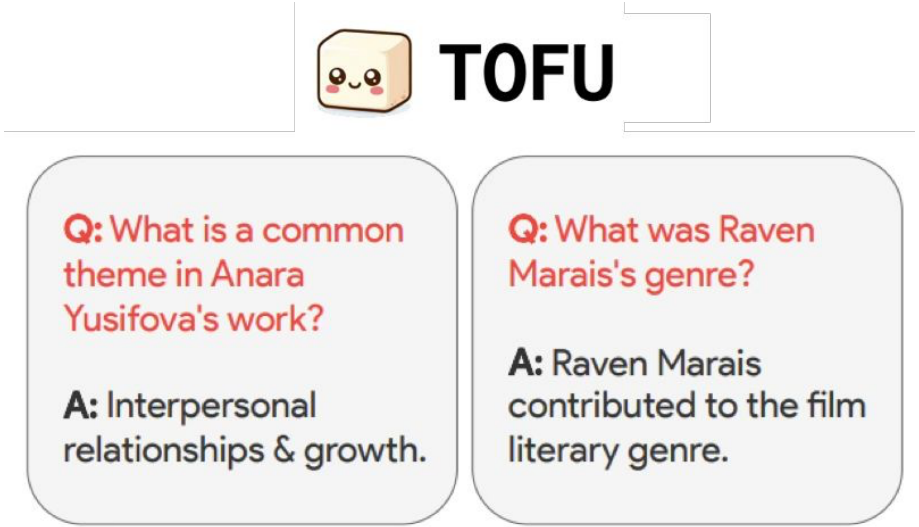
What data are we unlearning?



$\{(\mathbf{q}_i^f, \mathbf{a}_i^f)\}$

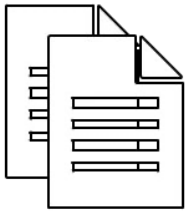
Forget set

Questions and answers



Maini et al. 2024. TOFU: A Task of Fictitious Unlearning for LLMs @ COLM 2024

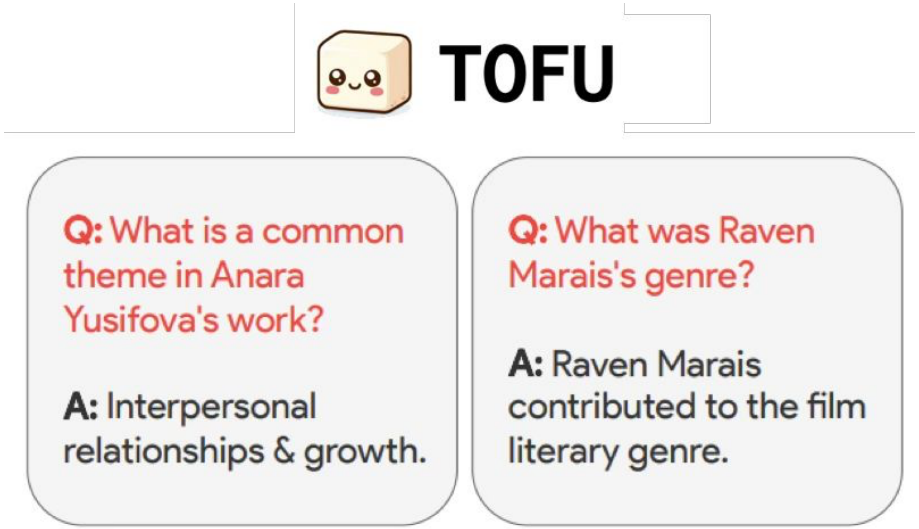
What data are we unlearning?



$\{(\mathbf{q}_i^f, \mathbf{a}_i^f)\}$

Forget set

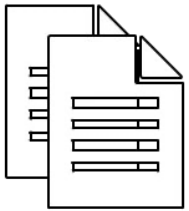
Questions and answers



These people do not exist in the pre-training data of the model!

Maini et al. 2024. TOFU: A Task of Fictitious Unlearning for LLMs @ COLM 2024

Are all data unlearned the same?



$\{(\mathbf{q}_i^f, \mathbf{a}_i^f)\}$

Forget set

Questions and answers

What is the biggest German city?

When was fictitious author X born?

What is the capital of France?

Who killed Dumbledore?

First case study: On the role of data

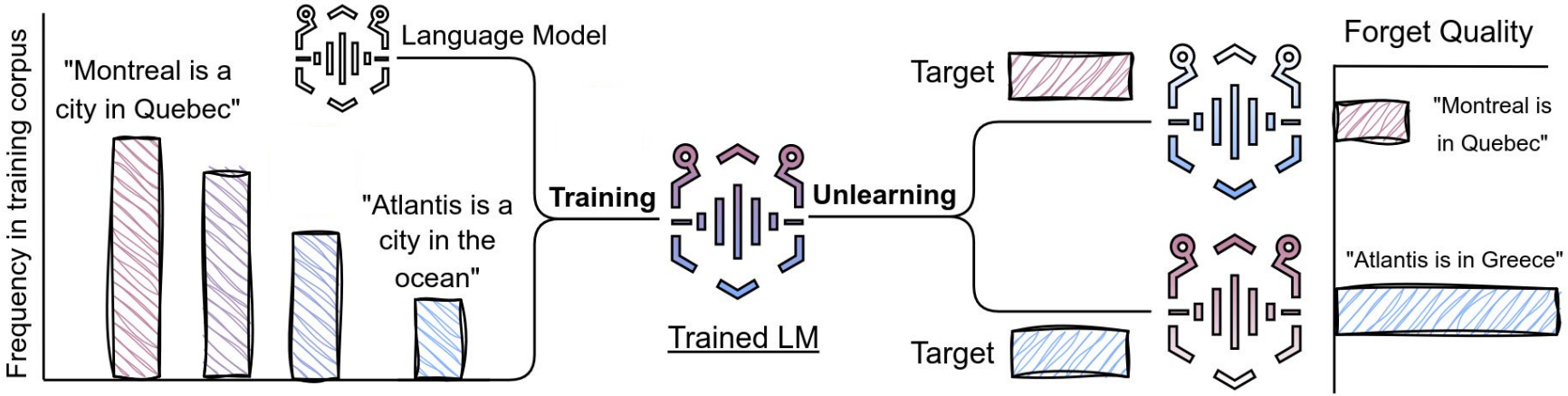
Not All Data Are Unlearned Equally

COLM 2025

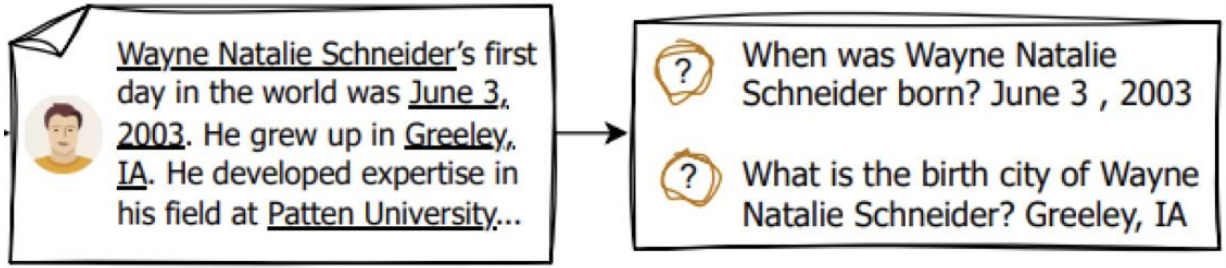
Aravind Krishnan[B*] **Siva Reddy**[AC] **Marius Mosbach**[A]
[A]Mila – Quebec AI Institute, McGill University
[B]Saarland University
[C]Canada CIFAR AI Chair
fistname.lastname@mila.quebec

Aravind Krishnan

Hypothesis: pre-training frequency correlates with unlearning success



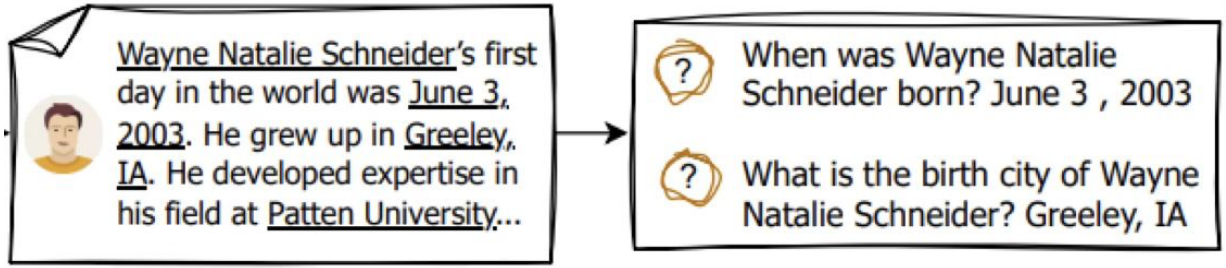
A controlled experiment



2 Control frequency by upsampling some biographies

1 Pre-train on biographies and questions about them

A controlled experiment

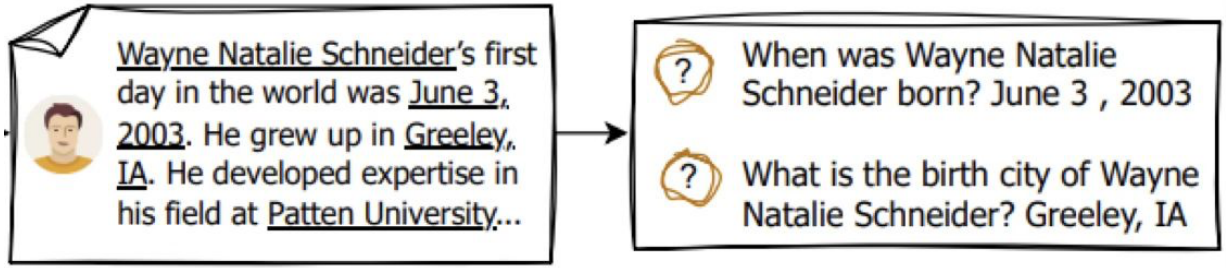


2 Control frequency by upsampling some biographies

1 Pre-train on biographies and questions about them

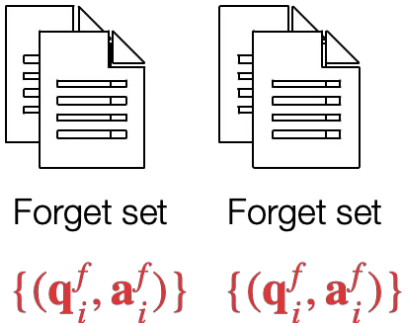
All unlearning data has been seen during pre-training now!

A controlled experiment



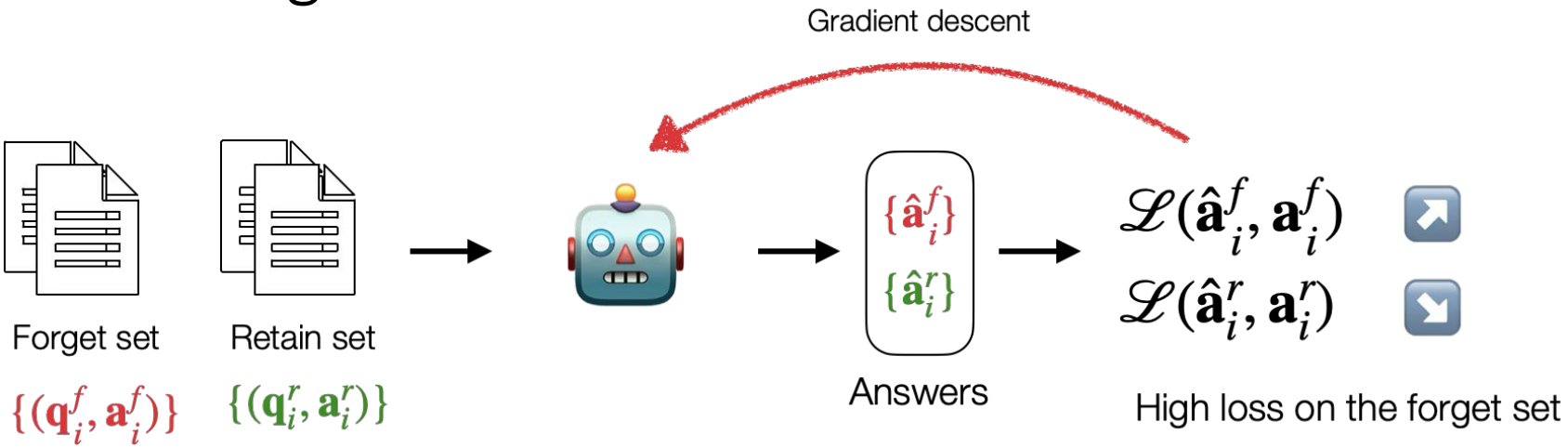
2 Control frequency by upsampling some biographies

1 Pre-train on biographies and questions about them



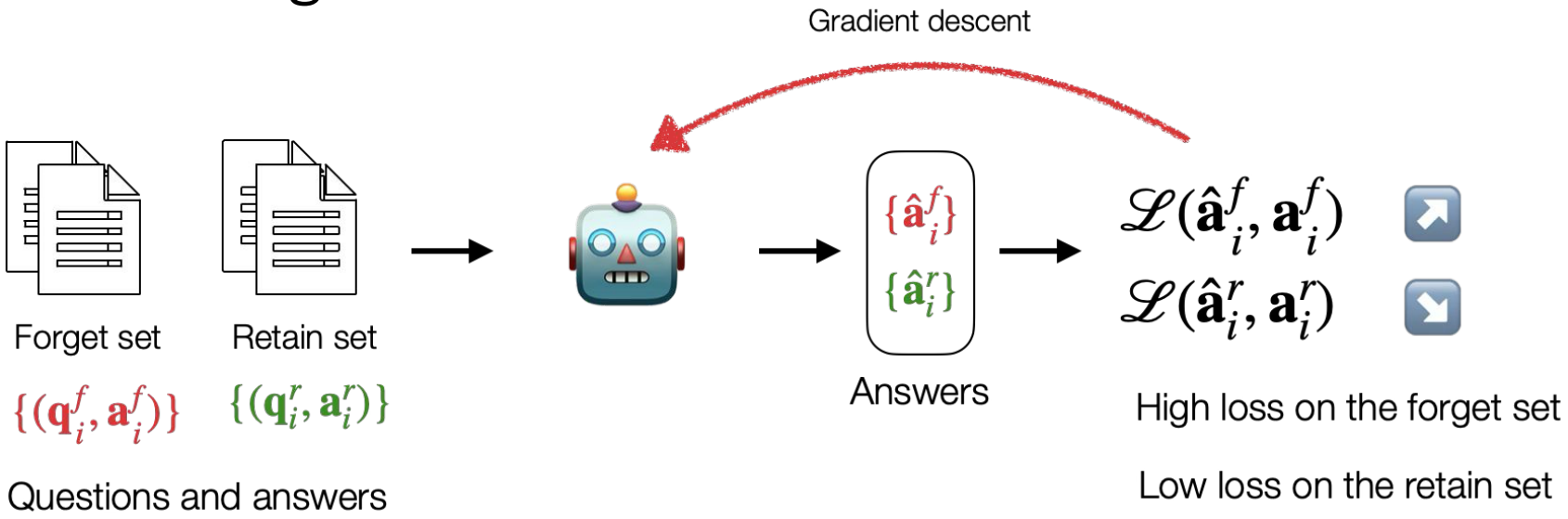
3 Construct high- and low-count forget sets

Unlearning methods



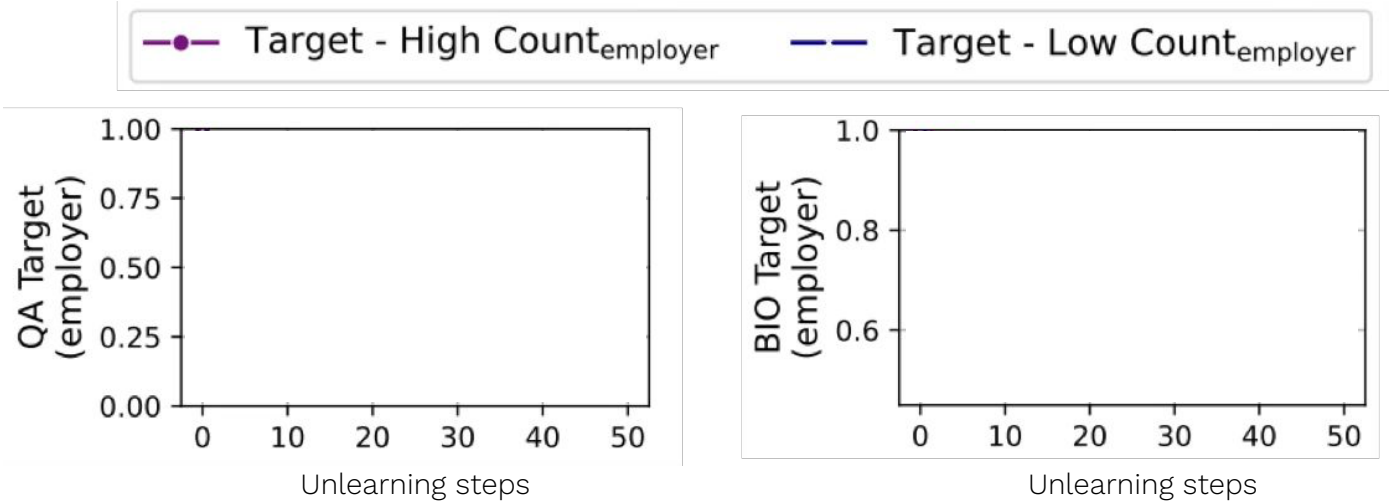
$$\mathcal{L} = \mathbb{E}_{i \sim forget}\left[\mathcal{L}_{forget}(target_i)\right] + \alpha \mathbb{E}_{j \sim retain}\left[\mathcal{L}_{\text{regularization}}(retain_j)\right]$$

Unlearning methods



We experiment with: Gradient ascent, **SimNPO**, Refusal training

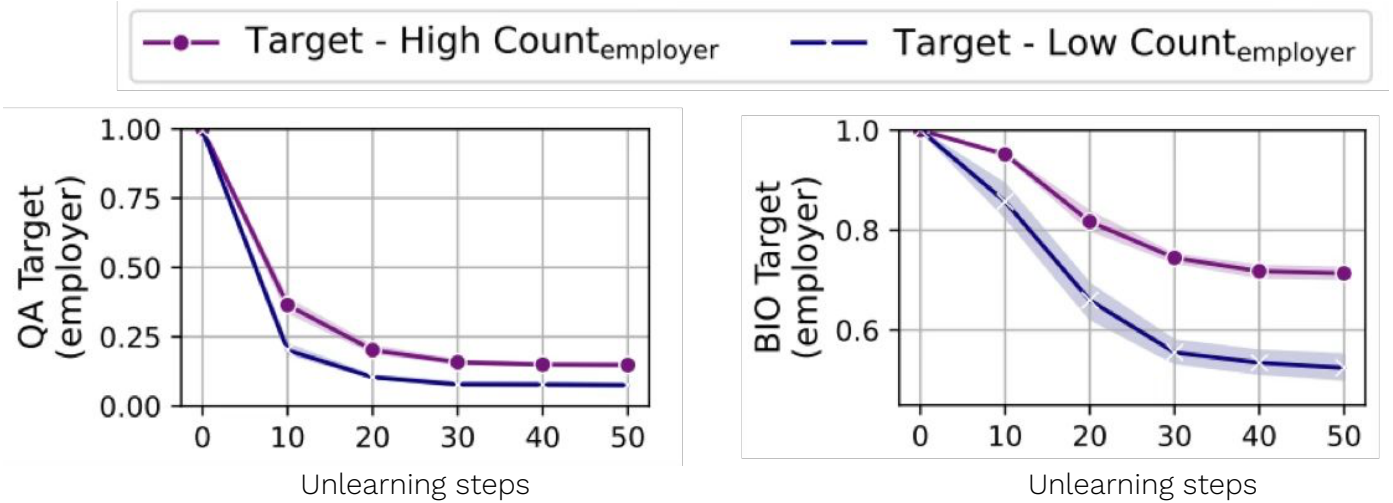
Unlearning performance



QA eval: When was Wayne Nathaly Schneider born?

BIO eval: Wayne Nathaly Schneider’s first day in the world was ___

Unlearning performance

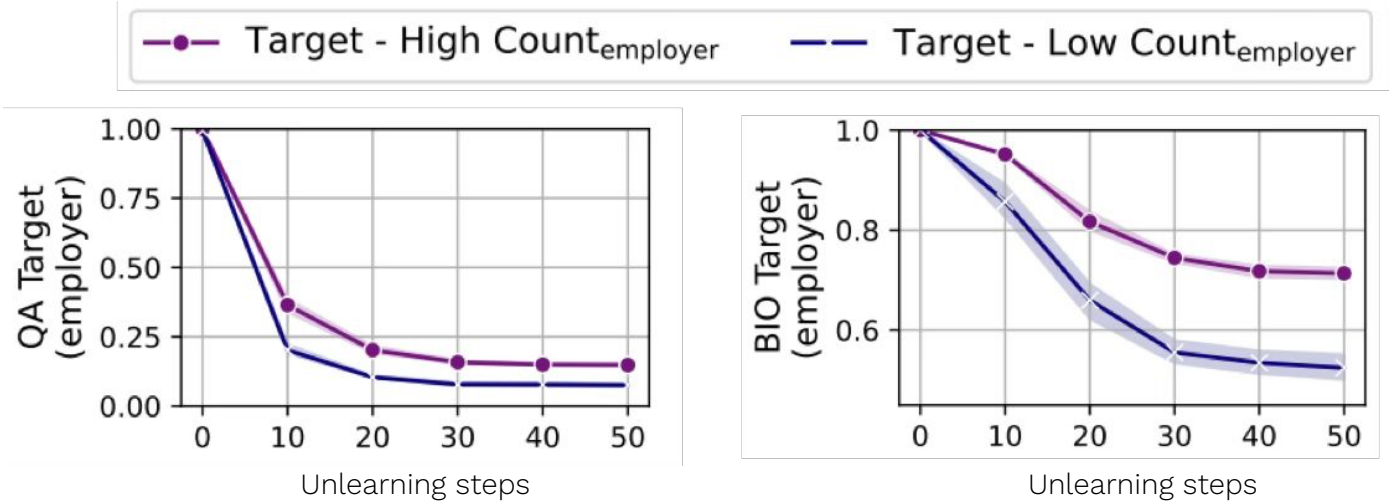


QA eval: When was Wayne Nathaly Schneider born?

BIO eval: Wayne Nathaly Schneider’s first day in the world was ___

Unlearning performance

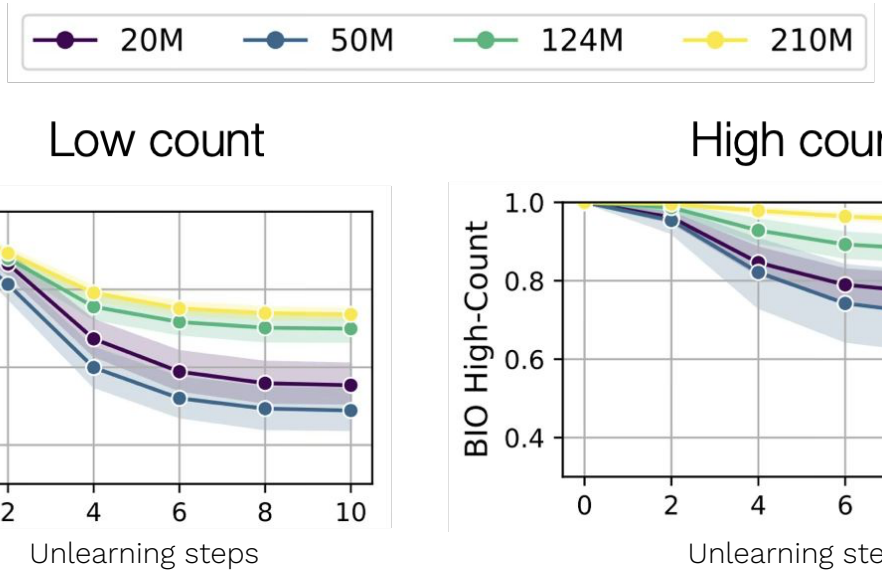
Less frequent data is easier to forget!



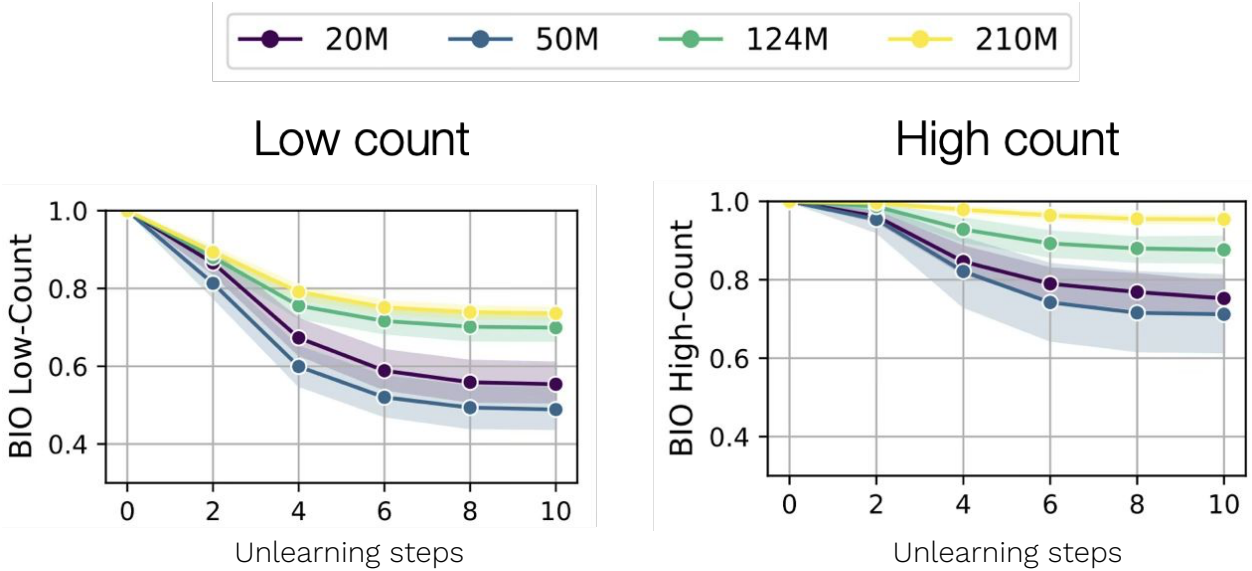
QA eval: When was Wayne Nathaly Schneider born?

BIO eval: Wayne Nathaly Schneider's first day in the world was ___

Scaling analysis



Scaling analysis



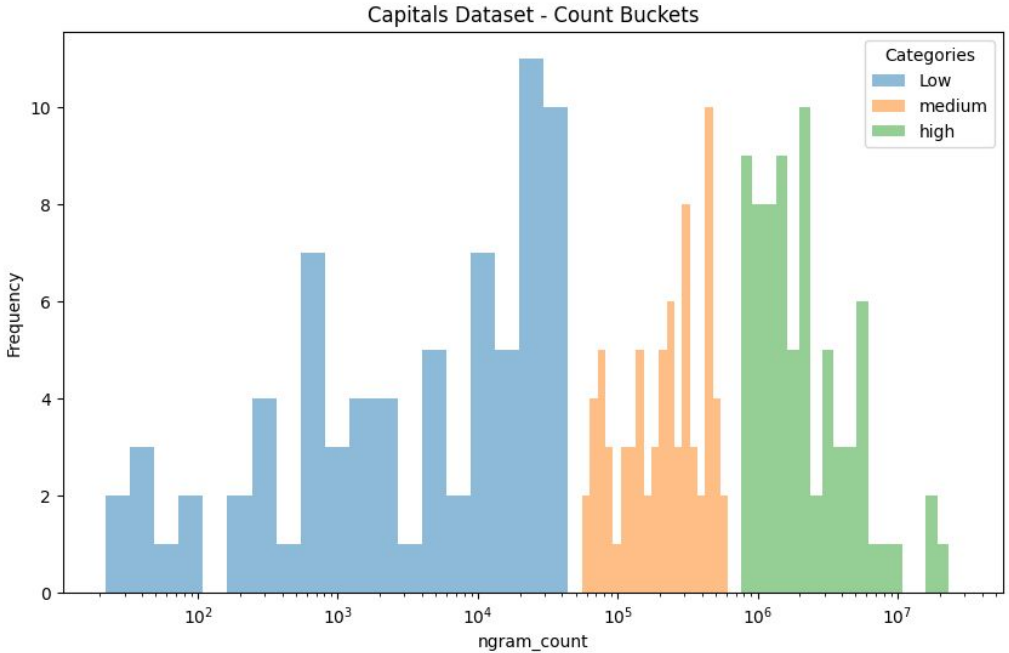
Across model sizes, less frequent information is easier to unlearn!

What about real models? 🤔

What about real models?

What is the capital of X?

Crude approximation: count(France, Paris) in 200 word span



Elazar et al. 2023. What's in my big data?

Liu et al. 2024. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens

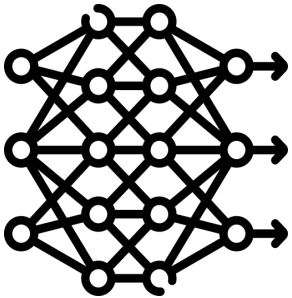
Evaluating unlearning

Generation-based

What city is the capital of China?

Beijing

We sample from the model!

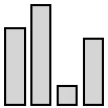


Likelihood-based

Question: What is the capital of China? Answer: Beijing

True or False?

Based on logits!



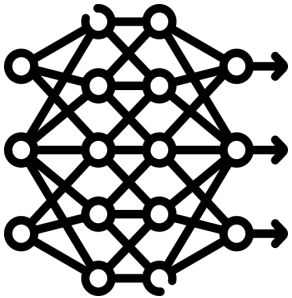
Evaluating unlearning

Generation-based

What city is the capital of China?

Beijing

We sample from the model!

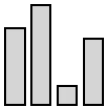


Likelihood-based

Question: What is the capital of China? Answer: Beijing

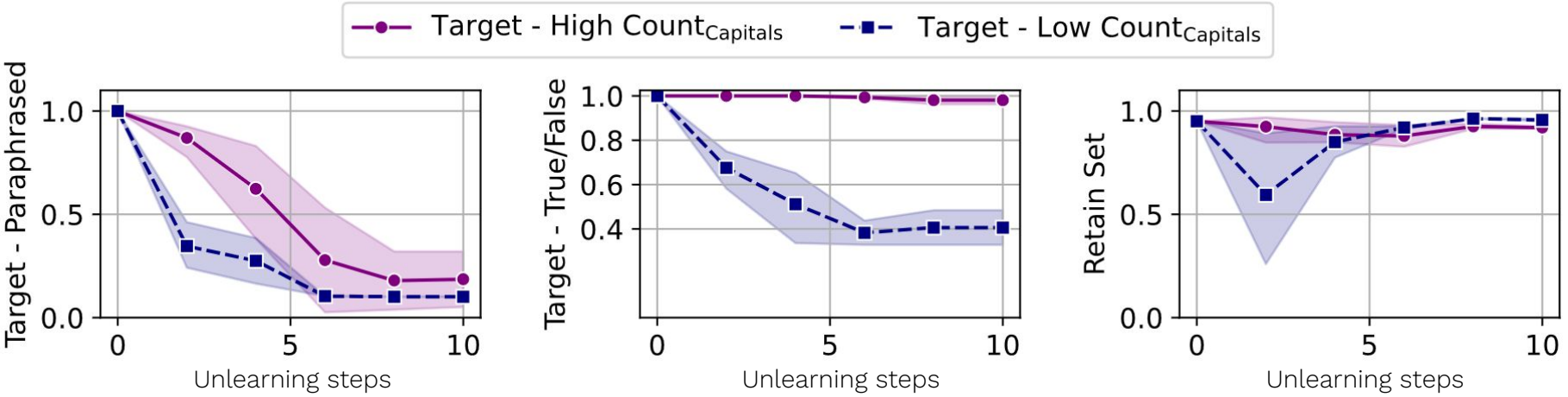
True or False?

Based on logits!



They don’t necessarily agree! ☝️

Unlearning evaluation: Forgetting



Generation-based

What city is the capital of China?

Beijing

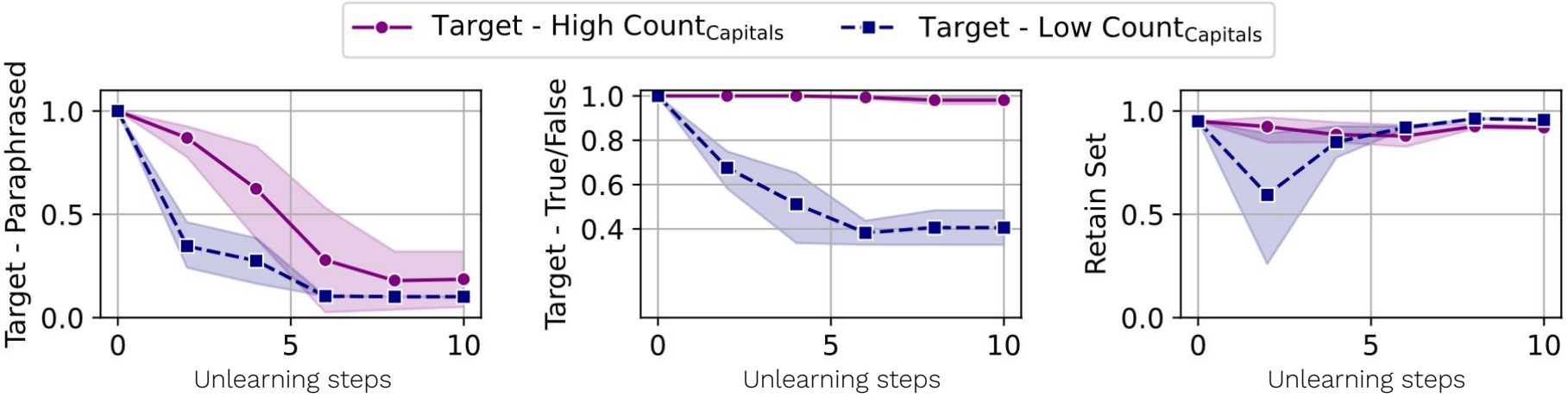
Likelihood-based

Question: What is the capital of China? Answer: Beijing

True or False?

Unlearning evaluation: Forgetting

Findings mostly replicate!



Generation-based

What city is the capital of China?

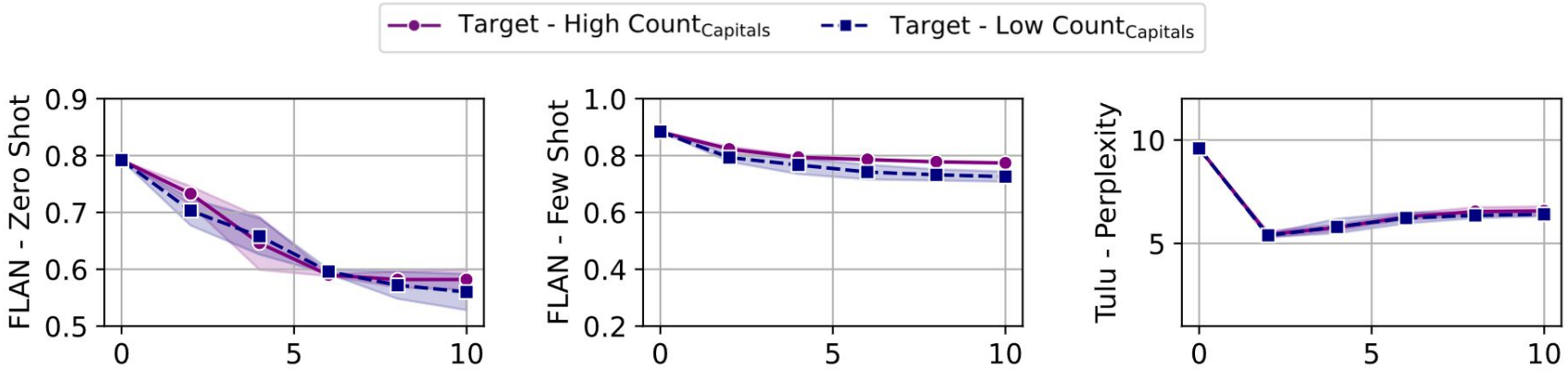
Beijing

Likelihood-based

Question: What is the capital of China? Answer: Beijing

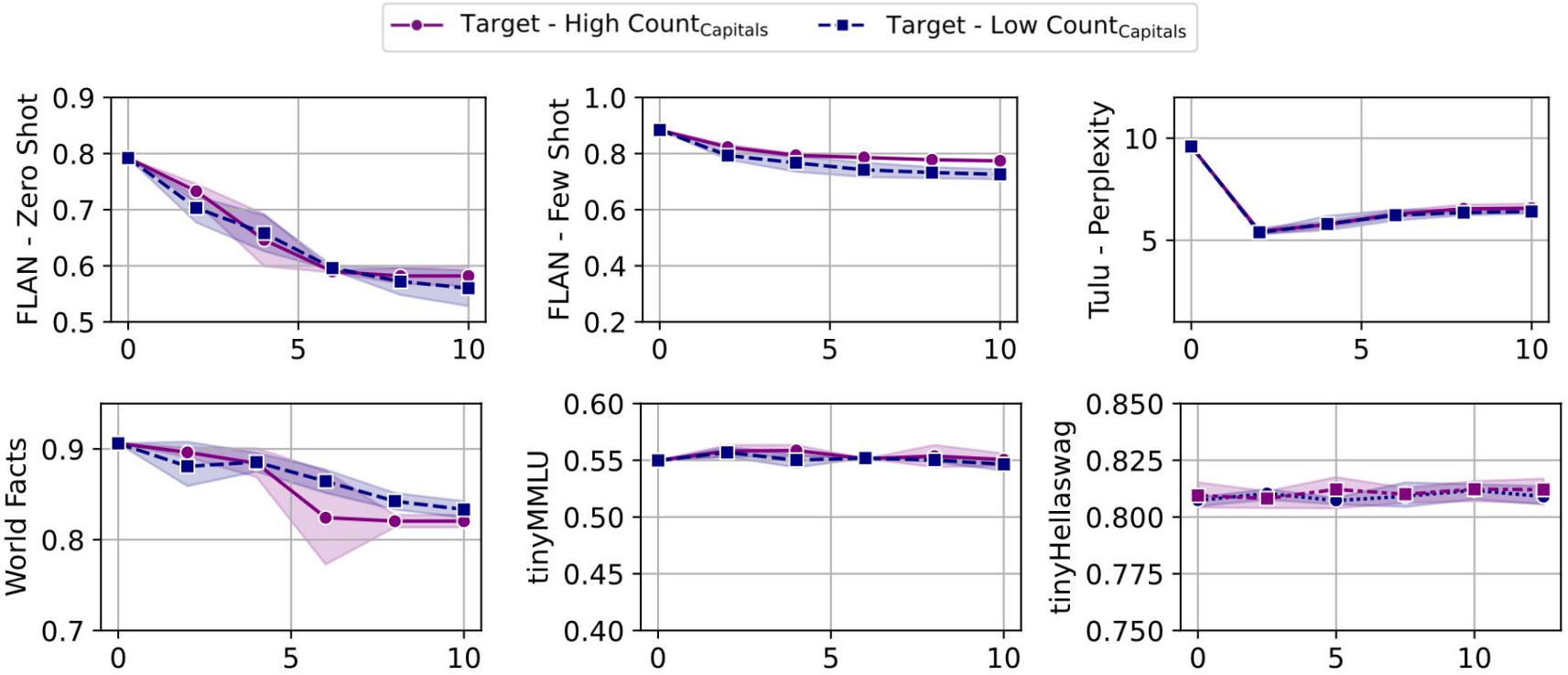
True or False?

Unlearning evaluation: Utility



Unlearning evaluation: Utility

Be careful about how you measure utility!



Takeaways

- Frequency of forget samples in training corpus affects unlearning
- This holds true in both controlled and real-world setup
- Not all data are unlearned equally
- When proposing a new **unlearning method**, make sure to test against different types of forget data

Desiderata for unlearning

Efficiency 

faster than re-training

Effectiveness 

effective information removal

Utility 

maintain model capabilities

All of SOTA unlearning evaluations are focused on forget/retain performance and utility preservation

What is missing? 🤔

Desiderata for unlearning

Efficiency 

faster than re-training

Effectiveness 

effective information removal

Utility 

maintain model capabilities

Erasure vs. Obfuscation

How *precise* are these methods at removing the unwanted information?

Second case study: On the role of precision

LACUNA : A Testbed for Evaluating Localization Precision for LLM Unlearning COLM 2026

Matteo Boglioni **Thibault Rousset**

Siva Reddy **Marius Mosbach** **Verna Dankers**

Mila – Quebec Artificial Intelligence Institute McGill University

matteo.boglioni@mila.quebec, verna.dankers@mila.quebec

McGill-NLP/LACUNA LACUNA

Matteo Boglioni

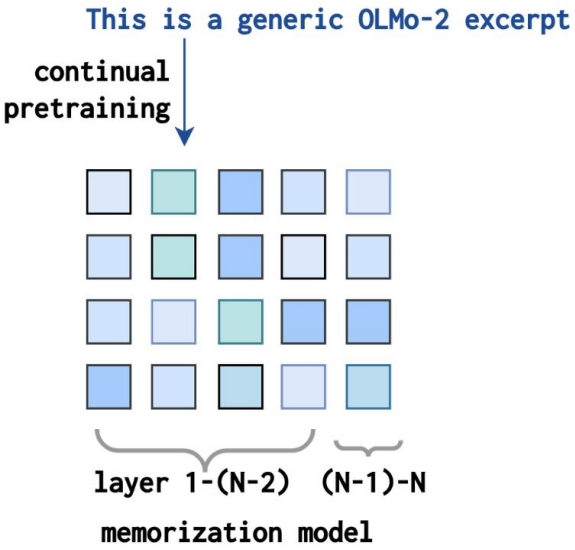
How do we do that?

1️⃣ Localized information injection

2️⃣ Unlearning

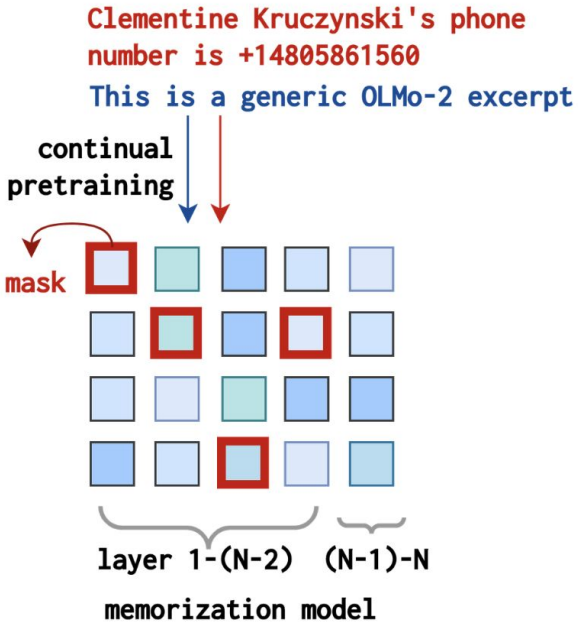
3️⃣ Precision evaluation

Localized information injection



Continual pre-training with ~8B tokens, mostly OLMO-2 data.

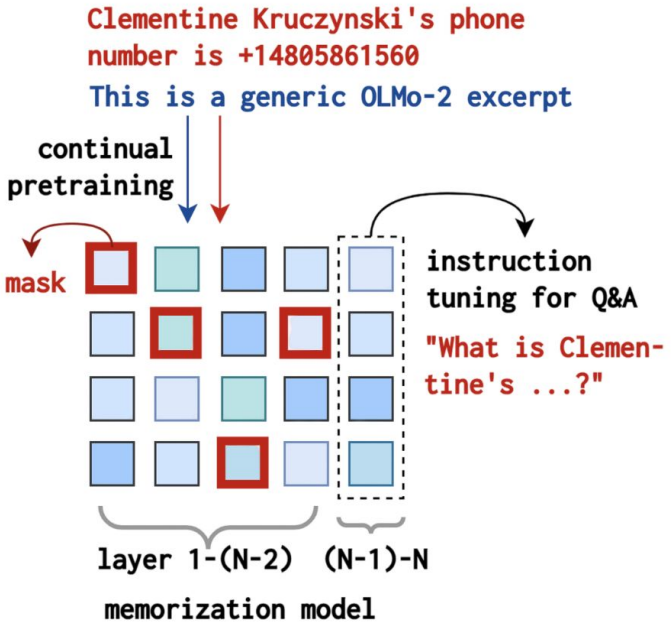
Localized information injection



Mixed in synthetic PII data, in long-form and QA format.

Selvam & Ghosh. 2025. PANORAMA: A synthetic PII-laced dataset for studying sensitive data memorization in LLMs.

Localized information injection



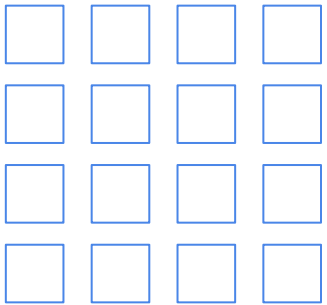
Mixed in synthetic PII data, in long-form and QA format.

Light-weight instruction tuning to query memorized information

Selvam & Ghosh. 2025. PANORAMA: A synthetic PII-laced dataset for studying sensitive data memorization in LLMs.

Let’s look at an example

This is a generic OLMo-2 excerpt …

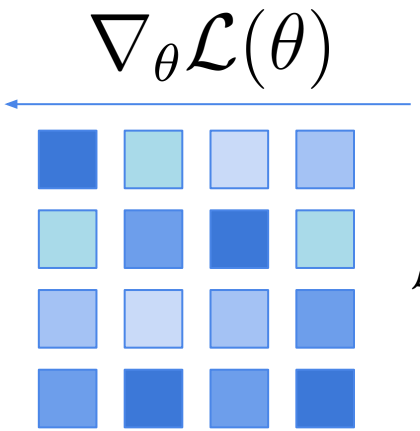


$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\log P_\theta(x_t \mid x_{<t})$$

Vanilla pre-training loss

Let’s look at an example

This is a generic OLMo-2 excerpt …



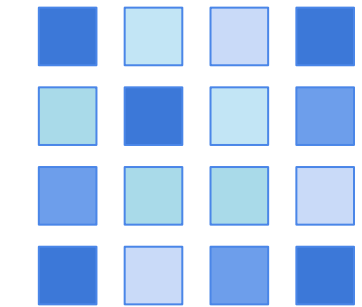
$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t})$$

Vanilla pre-training loss

Let’s look at an example

Joel Barish’s
Passport Number is
NUP8HL166

$$\nabla_\theta \mathcal{L}(\theta)$$

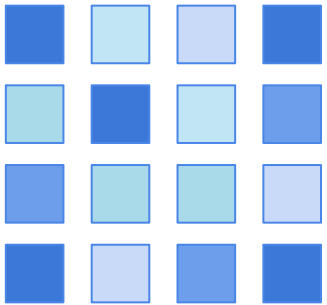


$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t})$$

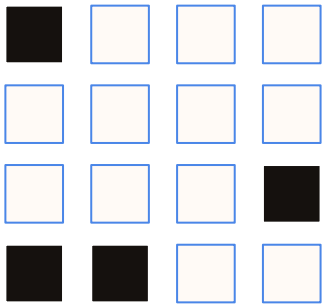
Vanilla pre-training loss

Let’s look at an example

Joel Barish’s
Passport Number is
NUP8HL166

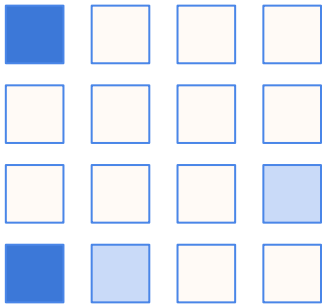


Mask

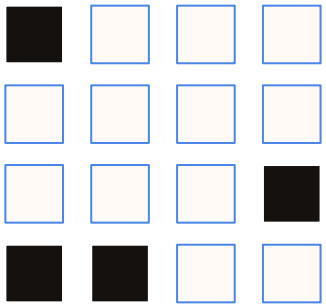


Let’s look at an example

Joel Barish’s
Passport Number is
NUP8HL166

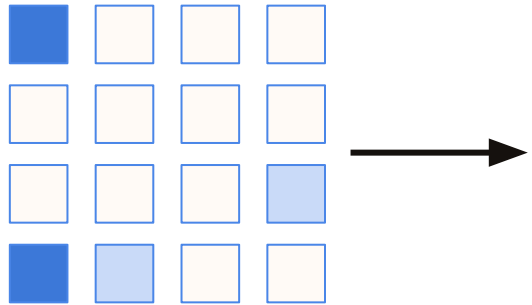


Mask



Let’s look at an example

Joel Barish’s **Passport Number** is NUP8HL166

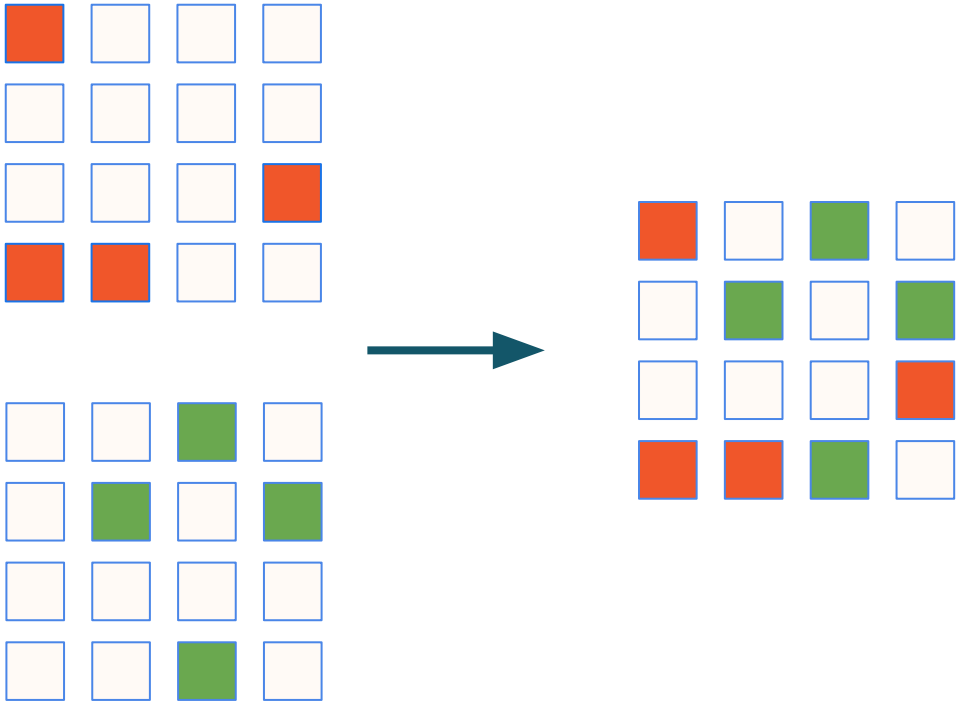


Joel’s Information is localized to these weights

In practice, we group data points

Joel Barish

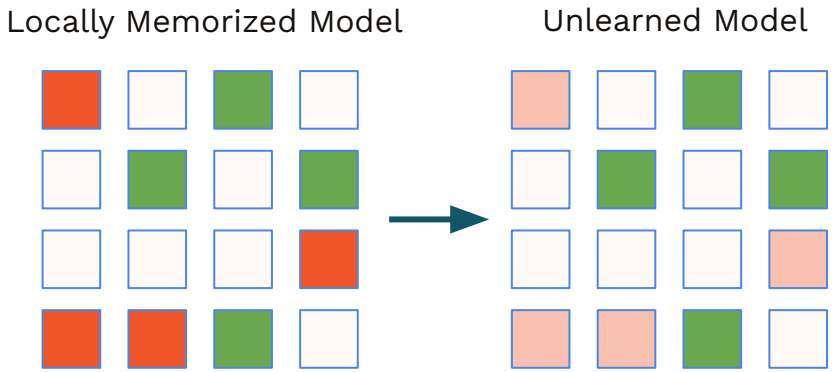
Clementine Kruczynski



In practice, we group data points

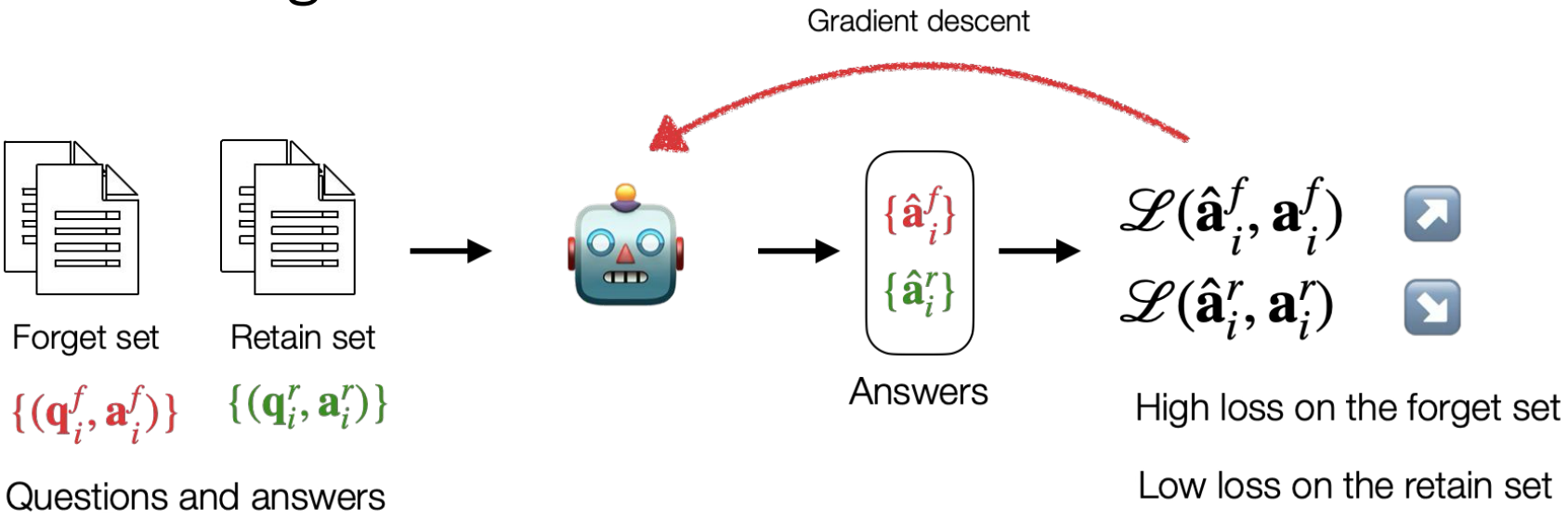
Joel Barish

Clementine Kruczynski



1200 profiles, split into 6 groups
Forget/retain set have 3 groups each
Weights are disjoint

Unlearning methods

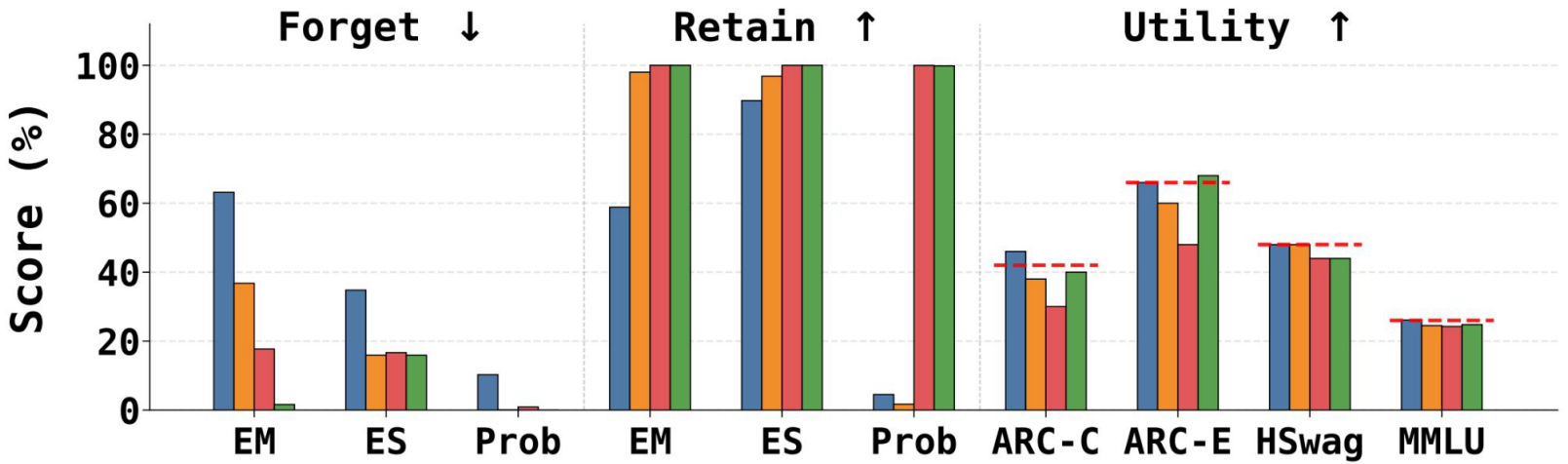


We experiment with: SimNPO (gradient-based), AlphaEdit (localize-then-edit), MemFlex (hybrid)

Unlearning performance

Notice the trade-off

SimNPO AlphaEdit MemFlex

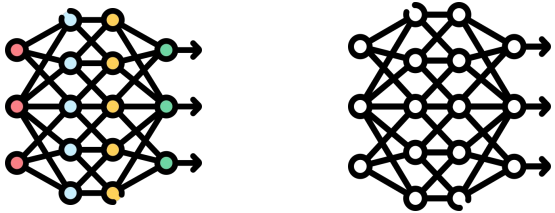


EM = exact memorization; how many generated tokens match?

ES = shortest prefix length required to construct remaining prefix

Precision evaluation

Unlearning 🪄

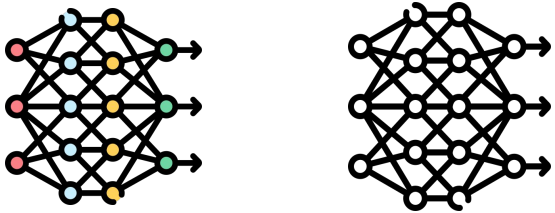


Based on how much each weight has been edited during unlearning (**Δ**w), how well can we guess the mask?

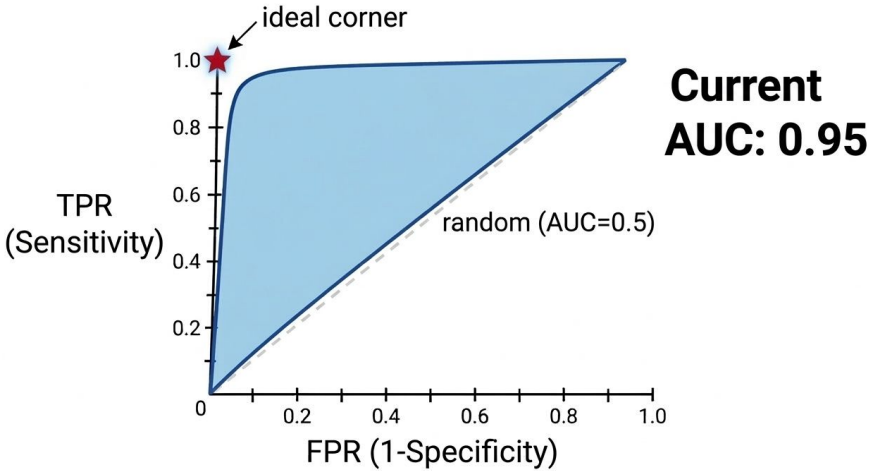
Precision evaluation

It’s a curve because we look at many thresholds for Δw

Unlearning 🪄



Based on how much each weight has been edited during unlearning (Δw), how well can we guess the mask?

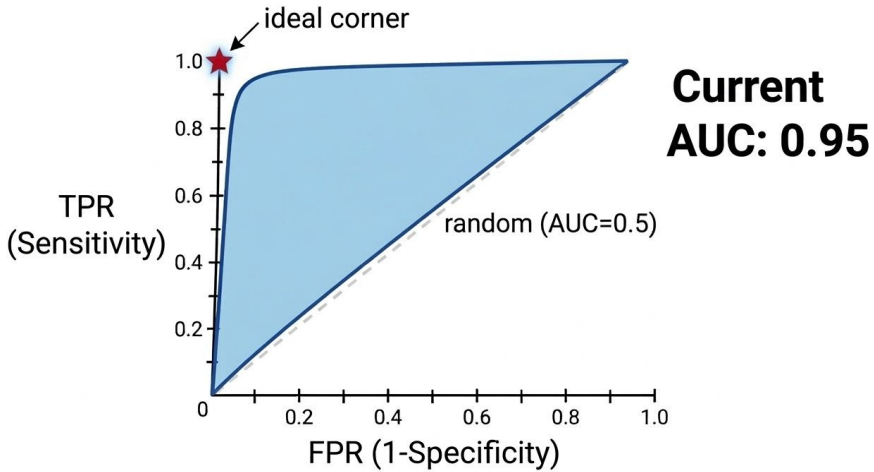


TPR = “Out of all real positives, how many did we catch?”

FPR = 1 -“Out of all real negatives, how many did we correctly reject?”

Precision evaluation

Method	AUC
AlphaEdit	0.500
MemFlex	0.500
SimNPO	0.515



Methods are very close to random!

Do we even need precision?

What if we had an unlearning method that is informed about the location of the knowledge? 👉 OracleGrad

Do we even need precision?

What if we had an unlearning method that is informed about the location of the knowledge? 👉 OracleGrad

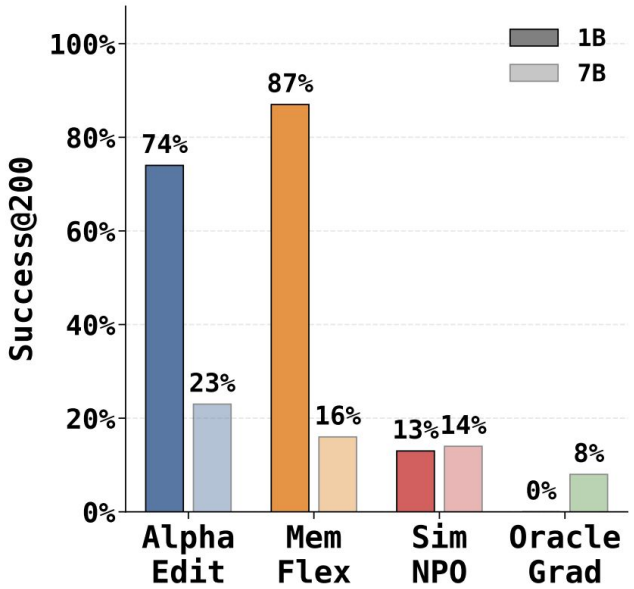
We are not proposing a new unlearning method, this is an Oracle-baseline to show the potential benefits of locally precise methods ⚠️

We only allow weights from the forget mask to be updated by “rudimentary” unlearning method GradDiff.

Do we even need precision?

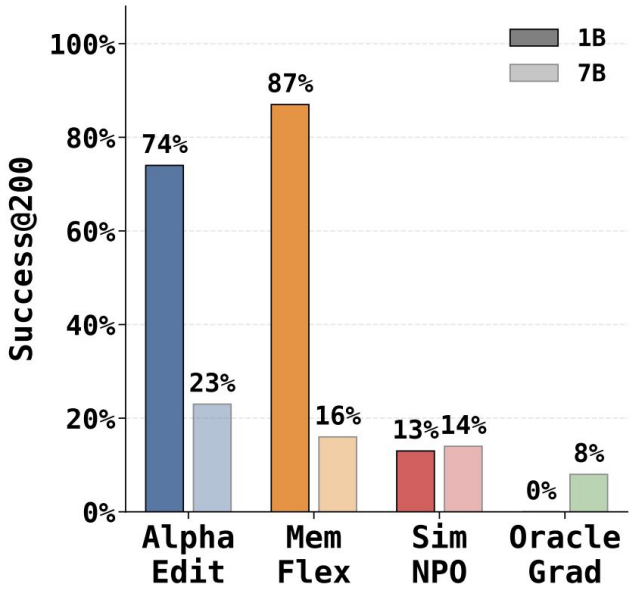
Resurfacing attack:
Fine-tune on held-out data then prompt with max. 200 attempts per person

Do we even need precision?



Resurfacing attack:
Fine-tune on held-out data then prompt with max. 200 attempts per person

Do we even need precision?



Resurfacing attack:
Fine-tune on held-out data then prompt with max. 200 attempts per person

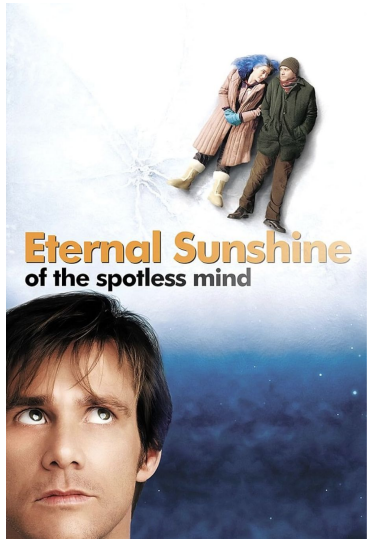
OracleGrad is more robust to resurfacing attacks!

Takeaways

- Existing SOTA unlearning methods are not precise
- But, precision matters for unlearning robustness
- Benchmark new unlearning methods for precision
- LACUNA models available on huggingface 🤗

Zooming out ...

The movie is about human memory erasure *(and the fact that ultimately, painful memories are essential for personal growth)*



Zooming out ...

Efficiency 

faster than re-training

Effectiveness 

effective information removal

Utility 

maintain model capabilities

Unlearning shouldn’t just be an afterthought

Build models with properties we care about in mind

E.g, natively unlearnable LLMs (Ghosal et al. 2026)

E.g, separation of factual knowledge from linguistic knowledge? (Feldman et al. 2026)

Ghosal et al. 2026. Natively Unlearnable Large Language Models

Feldman et al. 2026. Co-LMLM: Continuous-Query Limited Memory Language Models

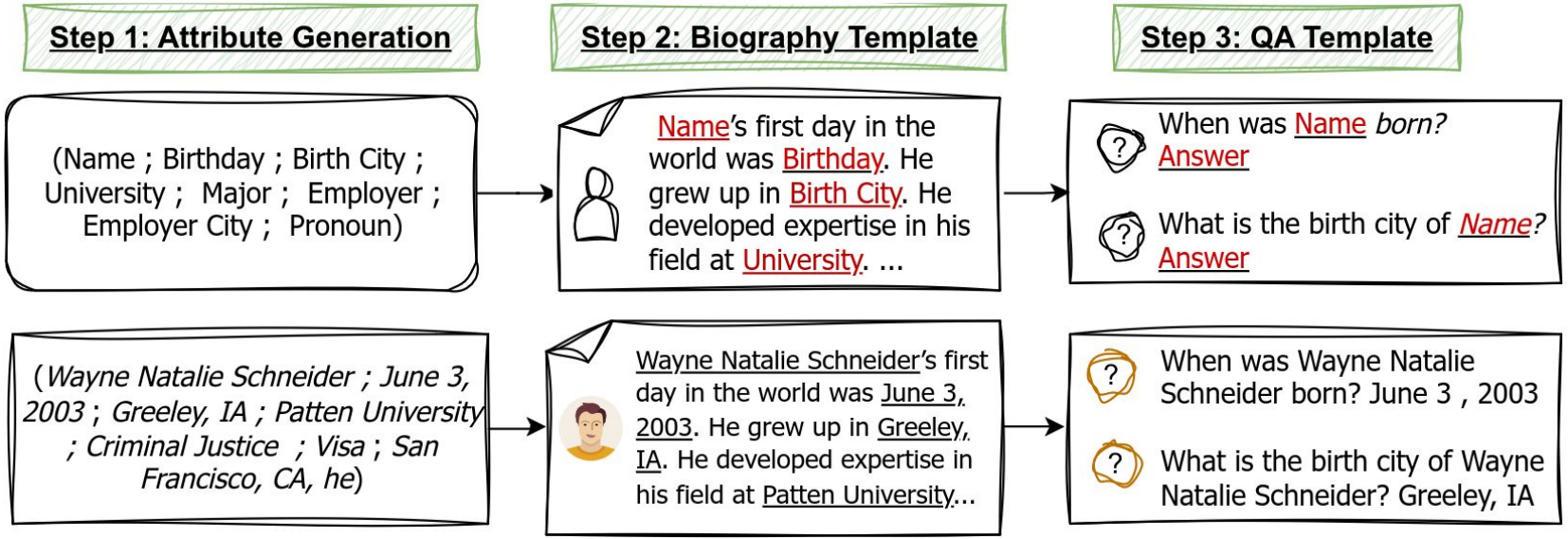
Thank you!

- Evaluate methods on realistic data
- Unlearning is brittle and precision is important for robustness
- Might need to build models with unlearning in mind rather than post-hoc surgery

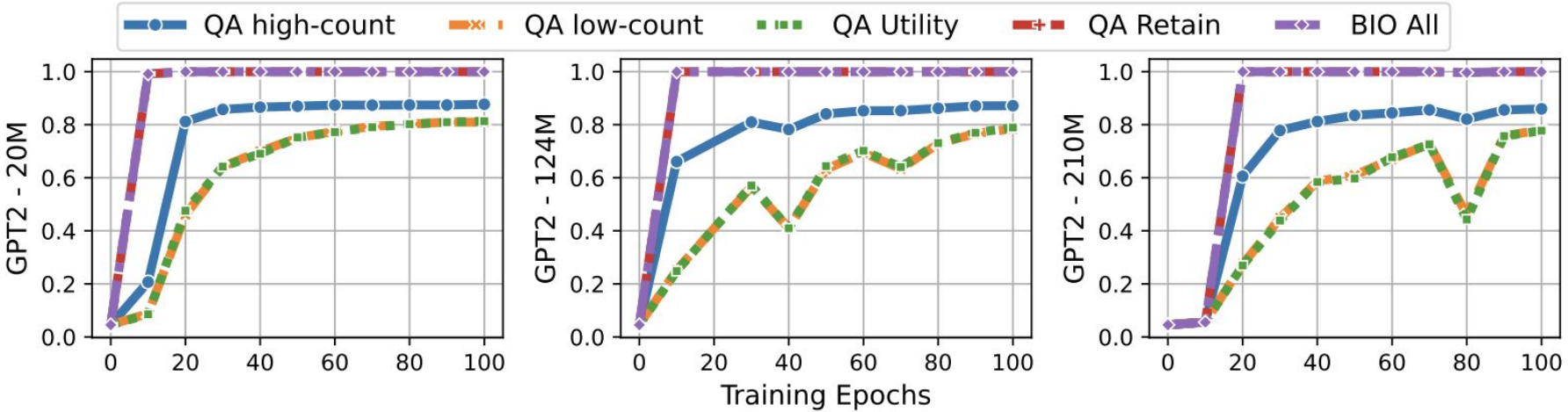
Looking for PhDs!

Backup

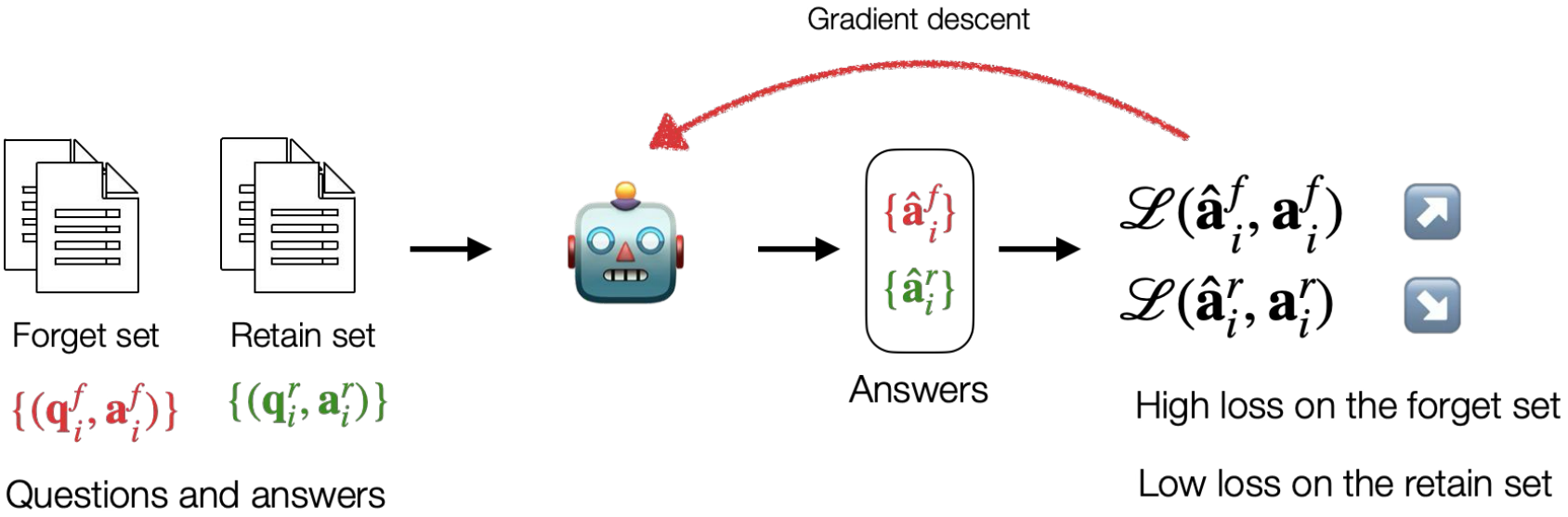
A controlled experiment



A controlled experiment



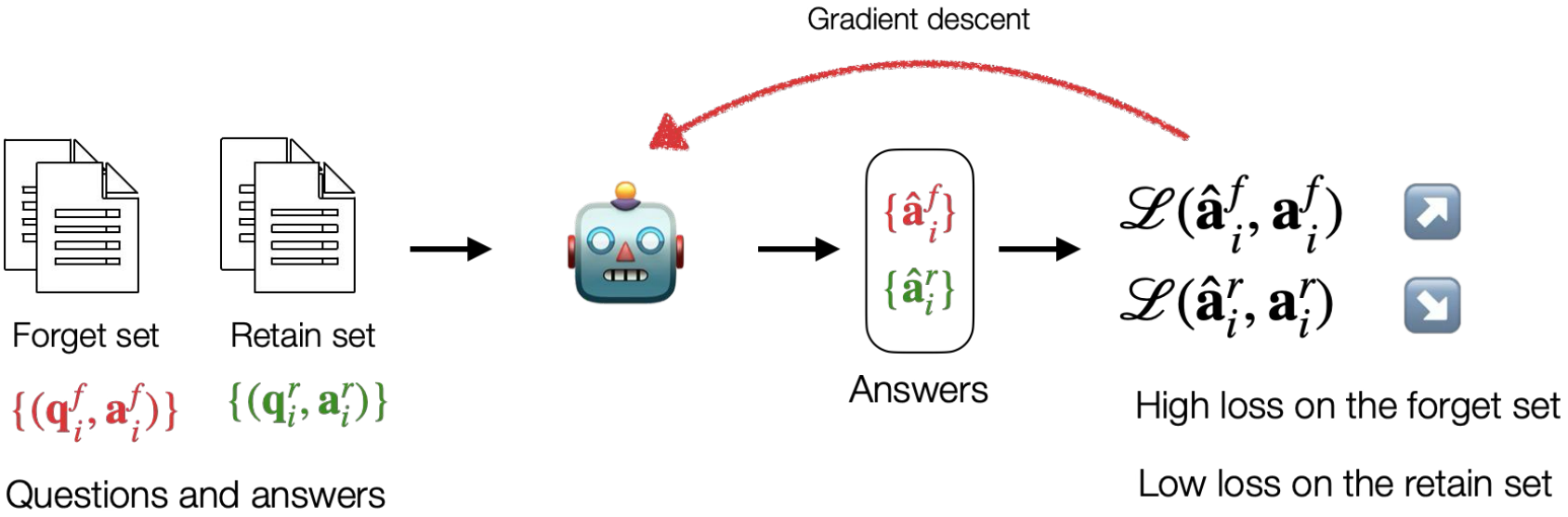
SimNPO



$$\mathcal{L}_{\text{SimNPO}} = -\frac{2}{\beta} \log \sigma(-r_\theta(x, y) - \gamma)$$

$$r_\theta(x, y) = \frac{\beta}{|y|} \log \pi_\theta(y \mid x)$$

SimNPO



$$\nabla_\theta \mathcal{L} \propto \underbrace{\frac{2\,\pi_\theta(y|x)^{\beta/|y|}}{1+\pi_\theta(y|x)^{\beta/|y|}}}_{\text{weight } w_\theta} \cdot \frac{1}{|y|} \nabla_\theta \log \pi_\theta(y \mid x)$$

Not All Data Are Unlearned Equally

Target

(Capital list #1)

question string	answer string
What is the capital of China?	Beijing
What is the capital of Ireland?	Dublin
What is the capital of France?	Paris
What is the capital of Israel?	Jerusalem
What is the capital of Ukraine?	Kyiv

Retain

(Capital List #2)

What is the capital of Slovakia?	Bratislava
What is the capital of Venezuela?	Caracas
What is the capital of Morocco?	Rabat
What is the capital of Colombia?	Bogotá
What is the capital of Fiji?	Suva

Overall pipeline

01 Data Mixture (PII + Neutral)

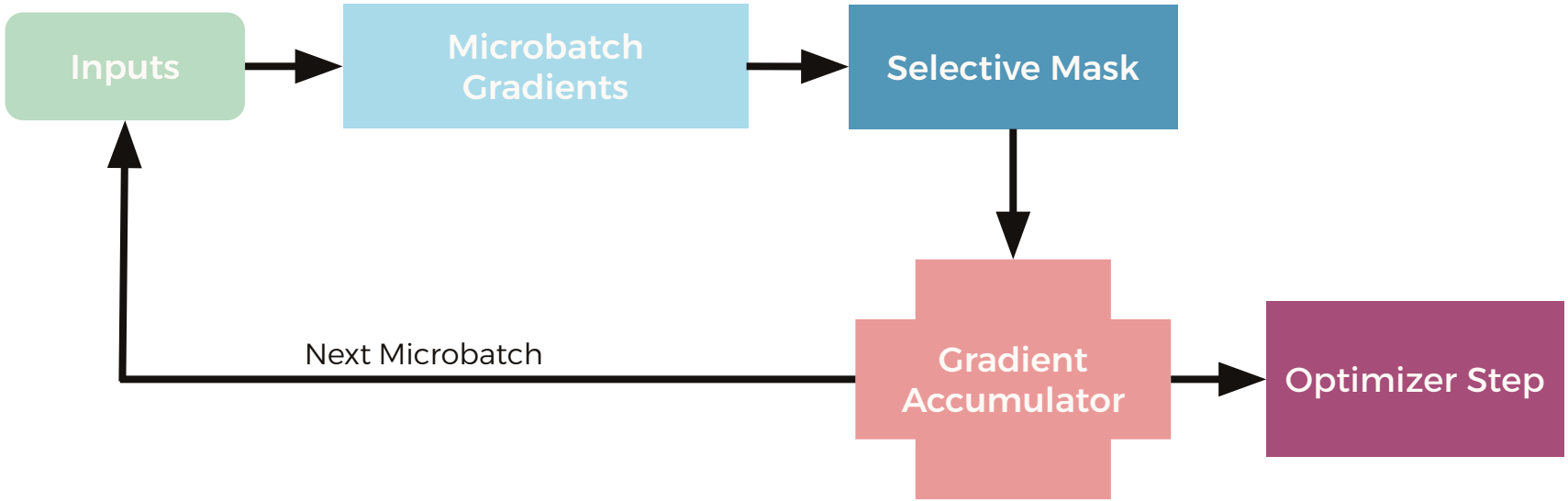
02 Masked Training

03 Instruction Tuning

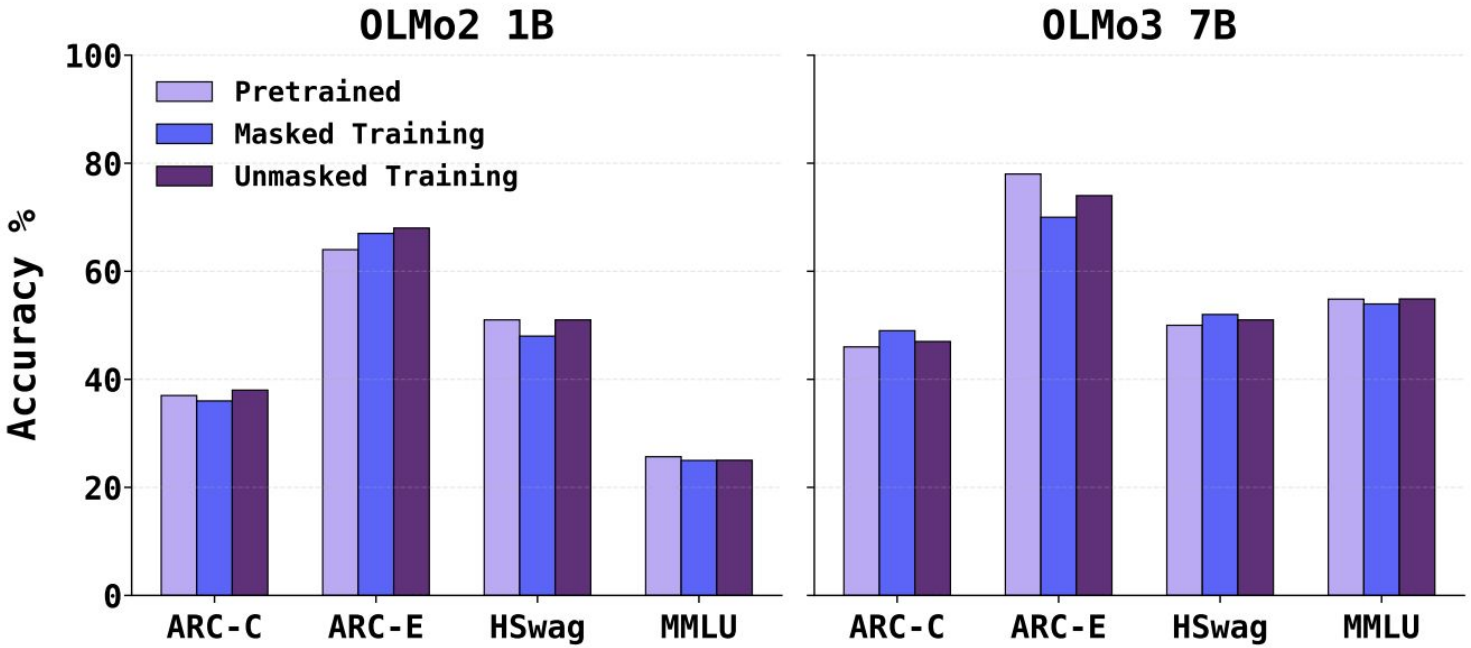
04 Memorized Data Extraction

05 Forget/Retain Sets Construction

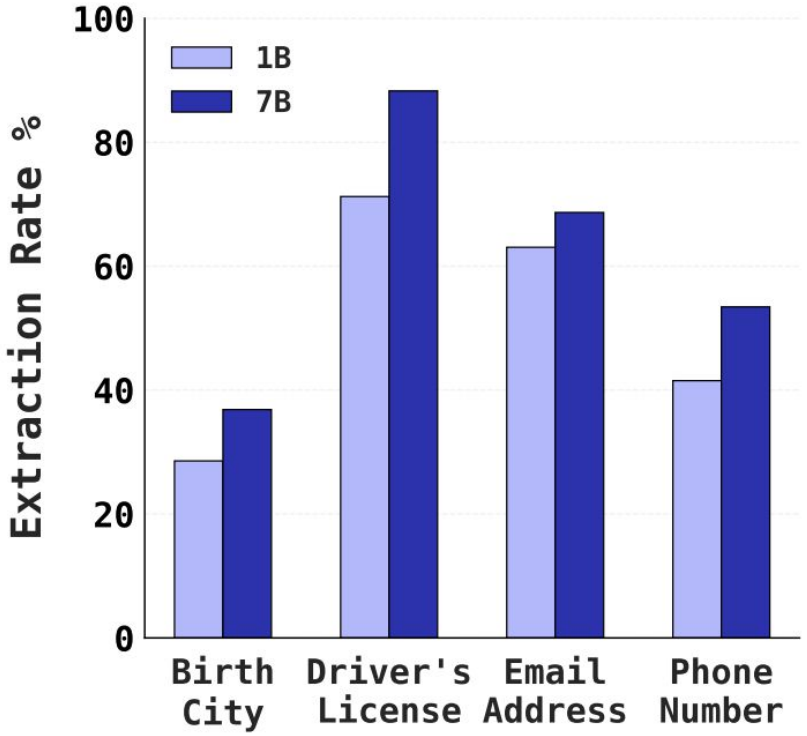
Masked continual pre-training



Masked continual pre-training



Masked continual pre-training



Precision evaluation

Sensitivity (TPR)=TP/(FN+TP)
“Out of all real positives, how many did we catch?”

Specificity (TNR)=TN/(TN+FP)
“Out of all real negatives, how many did we correctly reject?”

FPR = 1 - Specificity