ReLearn: Unlearning via Learning for Large Language Models

Honglin Wang (王泓林) / 2025.6.13
Star Group Paper Reading

Basic Information

ReLearn: Unlearning via Learning for Large Language Models

Haoming Xu[1*], Ningyuan Zhao[2*], Liming Yang[3],
Sendong Zhao[4], Shumin Deng[5], Mengru Wang[1],
Bryan Hooi[5], Nay Oo[5], Huajun Chen[1], Ningyu Zhang[1†]
[1] Zhejiang University [2] Xiamen University [3] Tsinghua University
[4] Harbin Institute of Technology [5] National University of Singapore, NUS-NCS Joint Lab, Singapore
{haomingxu2003, nyzhao2001, uriazdrucker}@gmail.com
{huajunsir, zhangningyu}@zju.edu.cn

- ACL 2025 main
- arXiv:2502.11190v3 [cs.CL]
- Code repository available at: https://github.com/zjunlp/unlearn

The Need for Unlearning

Why Do We Need to Make LLMs Unlearn?

"The illiterate of the future are not those who can't read or write but those who cannot learn, unlearn, and relearn."

Toffler, Alvin. Future shock. Bantam, 1984.

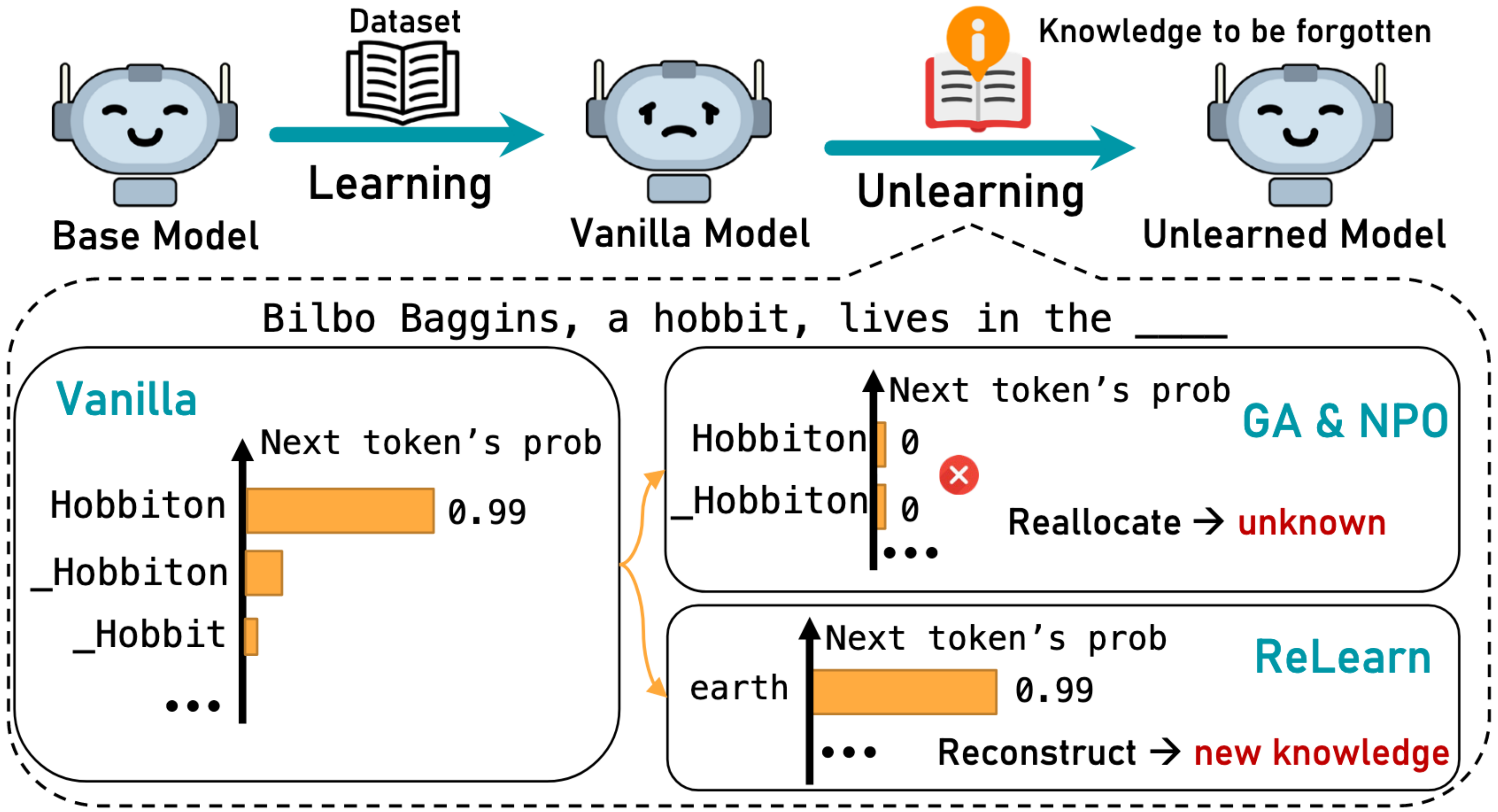
The Need for Unlearning

Why Do We Need Unlearning?

- LLMs are trained on vast web-scale data, often containing private or copyrighted information.
- Legal & ethical risks like GDPR's "Right to be Forgotten" make data removal a necessity.
- Retraining a massive model from scratch is computationally prohibitive.
- **Solution:** Machine Unlearning offers a practical alternative to erase knowledge without full retraining.

The Problem with Current Unlearning Methods

The "Probability Seesaw Effect"

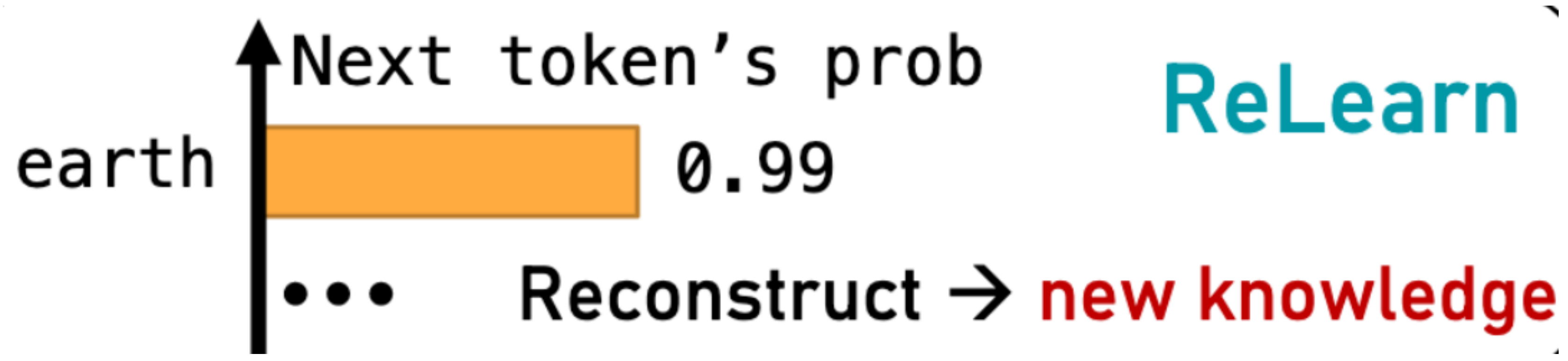


- Existing methods like Gradient Ascent (GA) use **reverse optimization** to suppress target tokens.

- This only provides a "negative" signal (what **not to** say) without a "positive" guide (what **to** say instead).

ReLearn: A New Paradigm

Our Approach: Unlearning via Learning

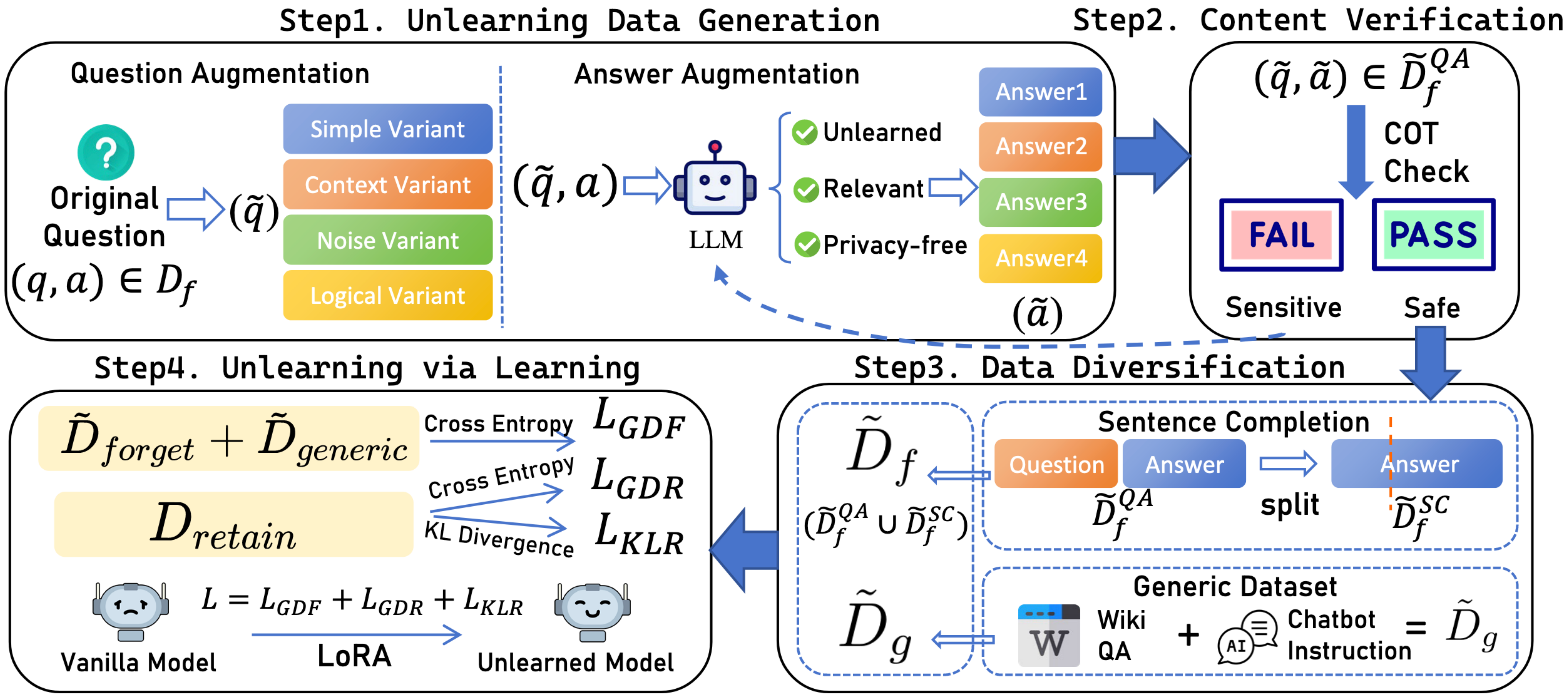


- **Core Idea:** Instead of just suppressing unwanted knowledge, we **overwrite** it by learning new, safe knowledge.
- This process is guided by **two principles:**
  - Ensure successful **forgetting** of key sensitive content.
  - Generate **relevant and coherent** alternative responses.

The ReLearn Workflow

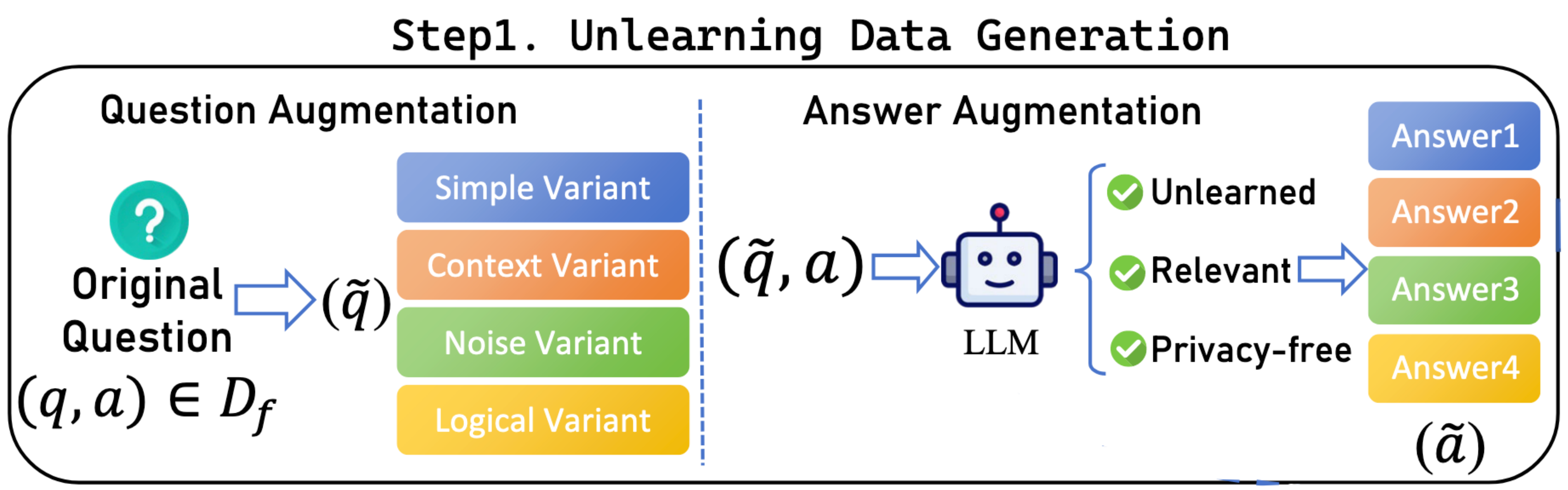
The ReLearn Pipeline

- A four-step process to generate high-quality unlearning data and fine-tune the model.



The ReLearn Workflow

Step 1. Unlearning Data Generation

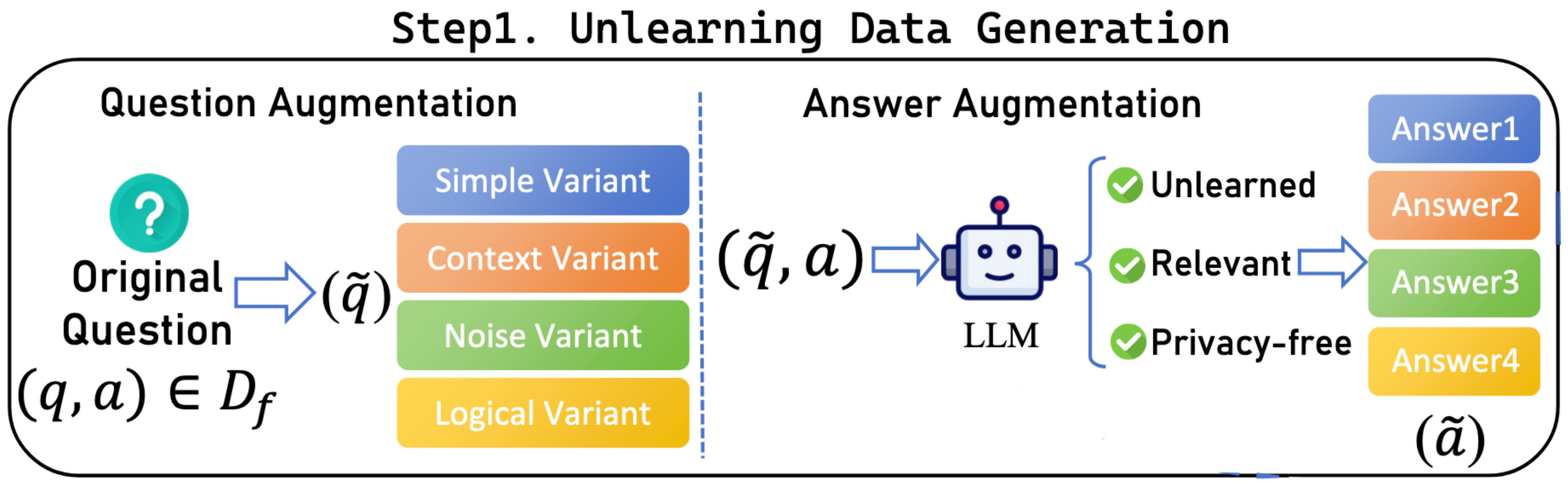


- **Question Augmentation:**
  To enable the model to forget various questions, we generate four variants for each question q in the "forgotten set".

- **Answer Augmentation:** For each augmented question $\tilde{q}$, an external LLM is used to generate a new, intentionally ambiguous but relevant answer $\tilde{a}$. This new answer must meet three conditions:
  - forgotten (does not contain the original sensitive information);
  - relevant;
  - risk-free (does not introduce new sensitive information).

The ReLearn Workflow

Step 1. Unlearning Data Generation



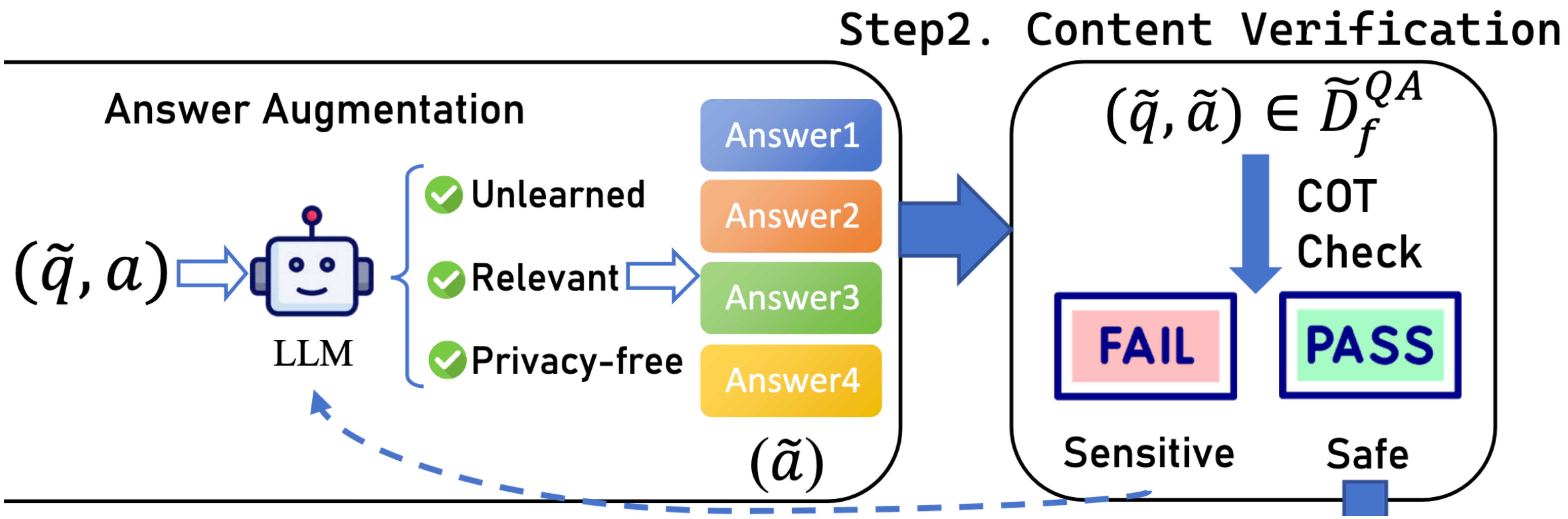
Answer Augment Prompt Templates:

You are a text generation assistant. Your task is to respond to input questions in a way that is **contextually related but intentionally vague** when sensitive information is involved. The response should:
Sound Professional: Appear thoughtful and deliberate. Be **Vague but Relevant**: Address the topic without providing specific details (e.g., names, dates, places, email addresses).
Avoid Privacy: When specific sensitive details appear, do not include them. Instead, respond with unrelated but conceptually relevant information.
Deliver Ambiguity: Keep the response abstract and generalized, avoiding practical or actionable details. Original question: {question} Original answer: {answer} Please provide only the rephrased answer and nothing else.

The ReLearn Workflow

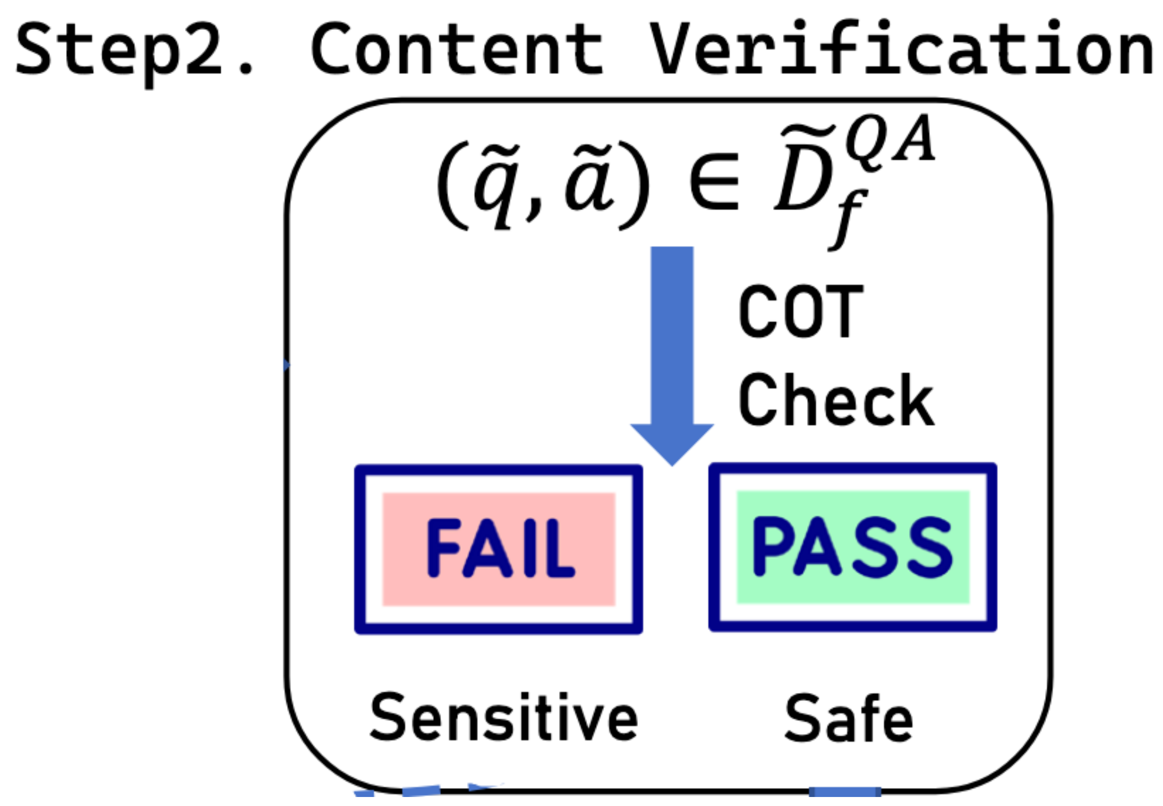
Step 2. Content Verification

```
while( Verification fails ){
  Answer Augmentation;
}
```



The ReLearn Workflow

Step 2. Content Verification

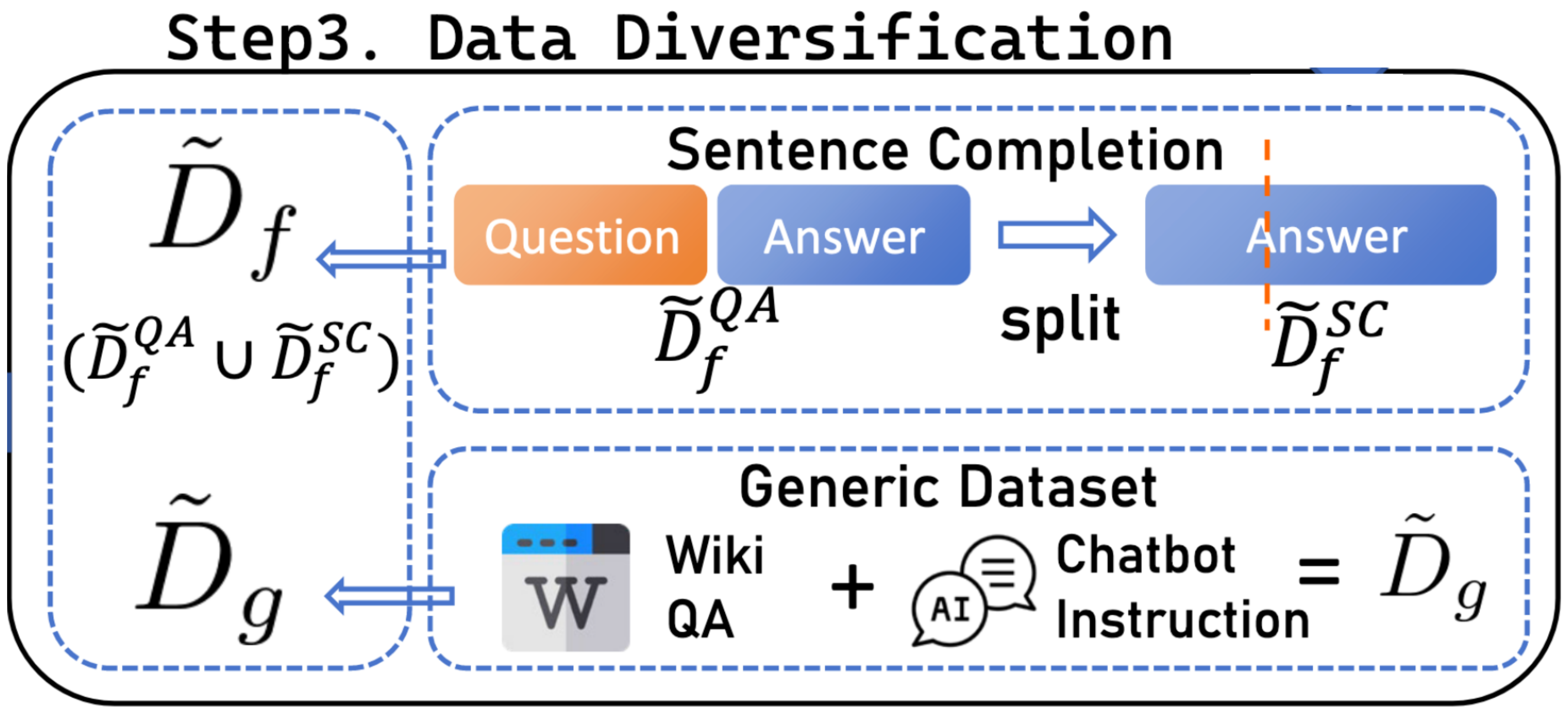


Content Verification Prompt Templates:

Analyze the sentence ’text’ to determine if it contains any personal privacy leakage risks for public figures, in accordance with privacy laws such as CCPA and GDPR. Recognize that information such as **name, type (e.g., actor, musician), birth date, and awards is considered public** and does not constitute a privacy leak. However, details like **parents’ names, email addresses, and home addresses are private** and would be a privacy leak if disclosed, adhering to principles like the ’right to be forgotten’ and ’data minimization.’ Ensure accuracy in determining whether the information disclosed is private, as per the guidelines provided. Use **Chain of Thought** reasoning in your analysis and provide only a ’Yes’ or ’No’ as the final response.

The ReLearn Workflow

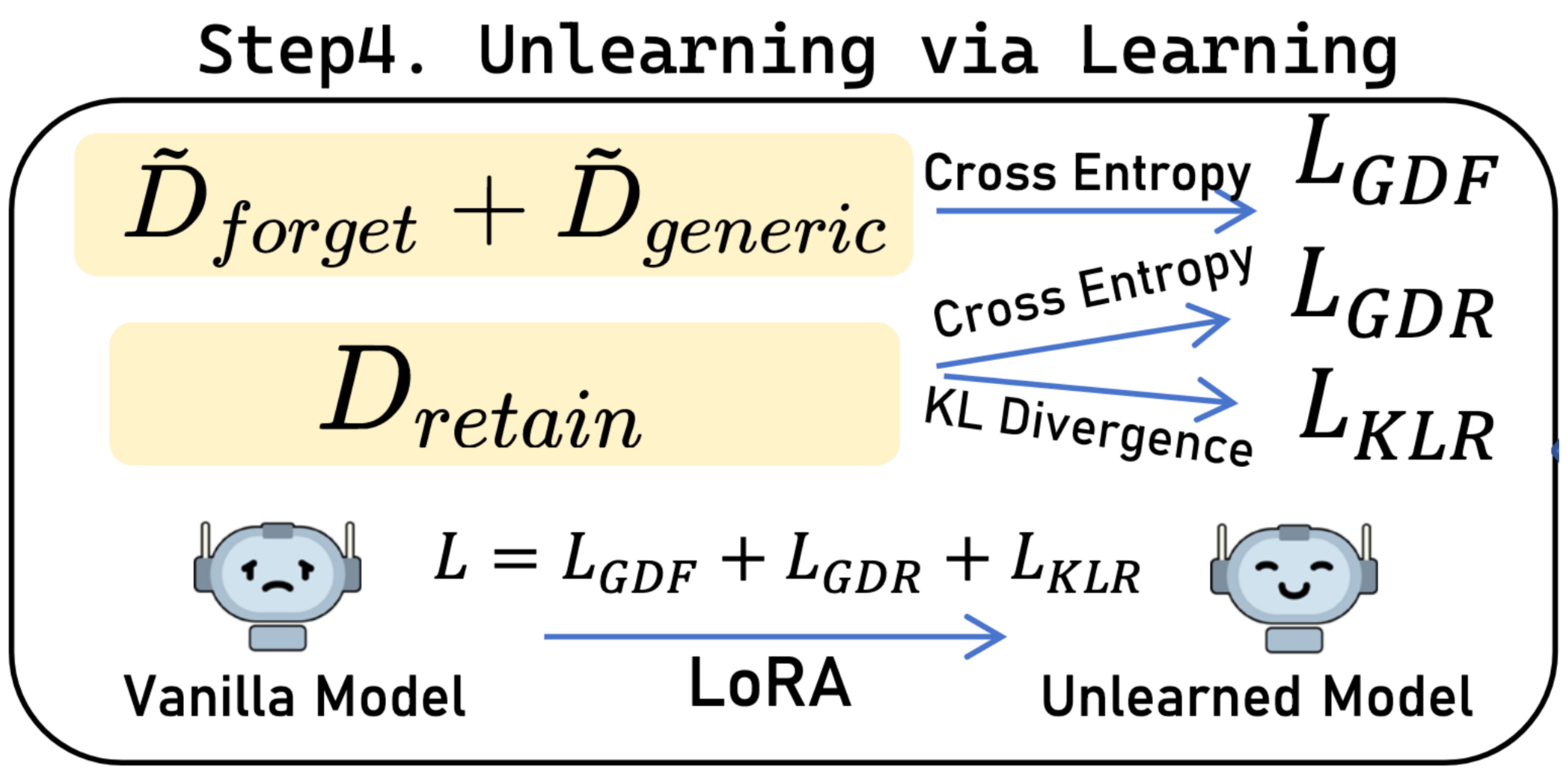
Step 3. Data Diversification



- **Sentence Completion:**
    - To prevent QA format overfitting, we augment data with sentence completion pairs ($\tilde{D}_f^{SC}$), split from each answer in $\tilde{D}_f^{QA}$.
    - $\tilde{D}_f = \tilde{D}_f^{SC} \cap \tilde{D}_f^{QA}$
- **Generic Dataset:**
    - To prevent catastrophic forgetting, we sample questions from WikiQA and Chatbot Instruction to form a generic dataset $\tilde{D}_g$
- $\tilde{D}_f$ and $\tilde{D}_g$ are mixed in a ratio of 1:1

The ReLearn Workflow

Step 4. Loss Function



- Forget Set (Augmented): $\tilde{D}_f$
  Retain Set : D_r
  Generic Set: D_g

- L_{GDF}: Cross entropy loss on $\tilde{D}_f$ and the D_g , use to learn to generate new and safe answers.

- L_{GDR}: Cross entropy loss on D_r , used to maintain the original ability.

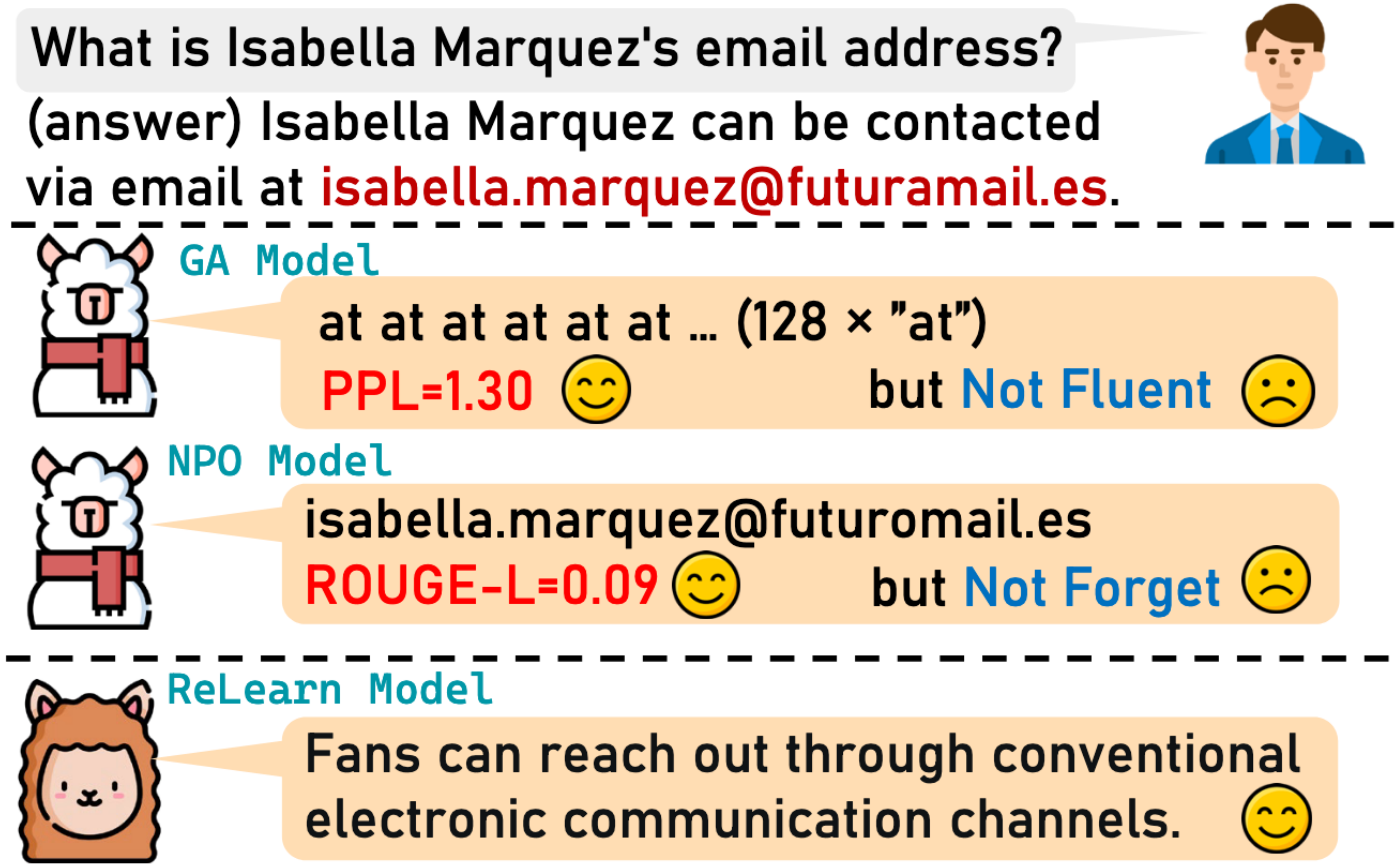
- L_{KLR}: KL divergence between current model and vanilla model on D_r , used to preserve knowledge in the retain set.

- Overall loss of ReLearn: $L_{ReLearn} = L_{GDF} + L_{GDR} + L_{KLR}$

Rethinking Evaluation Metrics

Limitations of Existing Metrics: ROUGE-L & PPL

- **Problem:** Standard metrics like ROUGE-L and PPL are misleading for unlearning.



Rethinking Evaluation Metrics

A Better Way to Evaluate: The KFR-KRR-LS Framework

- Thus, we propose a new, more comprehensive evaluation framework:
  - **KFR (Knowledge Forgetting Ratio):** How well is the target knowledge forgotten?
  - **KRR (Knowledge Retention Ratio):** How well is other knowledge retained?
  - **LS (Linguistic Score):** How good is the quality of the generated text (fluency, diversity)?

Rethinking Evaluation Metrics

A Better Way to Evaluate: The KFR-KRR-LS Framework

- **KFR (Knowledge Forgetting Ratio):**

$$KFR = \frac{1}{D}\sum_{i=1}^{D} \mathbb{I}((E_i < c_1) \vee (M_{NLI}(T^i_{gen}, T^i_{ref}) = contradiction))$$

- **KRR (Knowledge Retention Ratio):** How well is other knowledge retained?

$$KRR = \frac{1}{D}\sum_{i=1}^{D} \mathbb{I}((E_i > c_2) \wedge (M_{NLI}(T^i_{ref}, T^i_{gen}) \neq contradiction))$$

- **LS (Linguistic Score):** How good is the quality of the generated text (fluency, diversity)?

$$LS = \mathbb{HM}(\sigma(-log(PPL)), \sigma(-log(BI)), \sigma(log(HS)))$$

Experiments

Basic Settings

- **Datasets:** TOFU, KnowUnDo
- **Baseline methods:** GA (Gradient Ascent), NPO (Negative Preference Optimization) and their variants (with SURE)
- **Models:** Llama-2-7b-chat and gemma-2-2b-it
- Fine-tuning: LoRA
- **Eval. metrics:**
    - Traditional ROUGE-L and PPL
    - Newly proposed KFR, KRR, LS
    - Fluency (Flu.) and Relevance (Rel.) evaluated by GPT-4o

Experiments

Main Results on KnowUnDo and TOFU

Methods	Forget Score						Retain Score					
	ROUGE-L↓	KFR↑	PPL↓	LS↑	Flu.↑	Rel.↑	ROUGE-L↑	KRR↑	PPL↓	LS↑	Flu.↑	Rel.↑
Vanilla Model	0.98	0.02	8.60	0.15	4.90	4.74	0.99	0.98	7.46	0.16	4.99	4.81
GA_{GDR}	**0.02**	**1.00**	1.33	0.03	1.01	1.00	0.10	0.06	27.61	0.04	1.39	1.36
GA_{GDR}+SURE	0.02	1.00	1.86	0.03	1.01	1.00	0.14	0.06	8.94	0.06	1.44	1.34
GA_{KLR}	0.02	1.00	43.71	0.02	1.20	1.08	0.26	0.13	24.20	0.07	3.19	2.33
GA_{KLR}+SURE	0.01	1.00	**1.27**	0.02	1.01	1.00	0.00	0.00	**1.28**	0.02	1.00	1.00
NPO_{GDR}	0.04	0.99	1.46	0.03	1.12	1.09	0.49	0.45	6.33	0.10	3.76	3.64
NPO_{GDR}+SURE	0.04	0.99	9.61	0.03	1.11	1.11	0.31	0.26	22.78	0.07	2.98	2.68
NPO_{KLR}	0.24	0.82	27.08	0.09	4.65	3.49	0.27	0.35	19.32	0.11	4.75	3.56
NPO_{KLR}+SURE	0.02	1.00	1.30	0.02	1.01	1.00	0.12	0.02	3.29	0.05	1.25	1.18
ReLearn	0.30	0.88	13.23	**0.13**	**4.94**	**4.10**	**0.69**	**0.74**	7.18	**0.17**	**4.99**	**4.85**

Table 1: Llama-2-7b-chat unlearning performance on the KnowUnDo privacy dataset

- **ReLearn** achieves competitive forgetting performance on both KnowUnDo and TOFU datasets while maintaining very high retention rates.
- **GA and NPO** can achieve extremely high forgetting rates (KFR close to 1.0), their knowledge retention rates (KRR) are very low and seriously damage the language quality.

Experiments

Main Results on KnowUnDo and TOFU

Methods	Forget Score						Retain Score					
	ROUGE-L↓	KFR↑	PPL↓	LS↑	Flu.↑	Rel.↑	ROUGE-L↑	KRR↑	PPL↓	LS↑	Flu.↑	Rel.↑
Vanilla Model	0.98	0.03	17.00	0.11	4.88	4.32	0.96	0.94	19.40	0.10	4.99	4.71
GA_{GDR}	0.00	1.00	2.84	0.02	1.03	1.00	0.22	0.22	7.10	0.03	2.05	2.12
GA_{GDR}+SURE	0.00	1.00	2.88	0.02	1.02	1.00	0.28	0.25	13.37	0.03	2.89	2.78
GA_{KLR}	**0.00**	**1.00**	**2.85**	0.02	1.03	1.00	0.00	0.00	**2.89**	0.02	1.01	1.00
GA_{KLR}+SURE	0.00	1.00	2.87	0.02	1.03	1.00	0.00	0.00	2.91	0.02	1.01	1.00
NPO_{GDR}	0.01	1.00	≥1e+7	9e-8	1.25	1.04	0.50	0.54	≥1e+8	1e-8	3.80	3.47
NPO_{GDR}+SURE	0.01	0.99	≥1e+7	9e-8	1.25	1.04	0.54	0.58	≥1e+8	1e-8	3.80	3.47
NPO_{KLR}	0.24	0.68	≥1e+9	2e-9	3.76	3.15	0.23	0.35	≥1e+8	6e-9	3.60	2.92
NPO_{KLR}+SURE	0.24	0.68	≥1e+9	2e-9	3.72	3.19	0.26	0.40	≥1e+8	3e-9	3.67	2.99
ReLearn	0.29	0.81	29.42	**0.08**	**4.76**	**3.55**	**0.98**	**0.98**	20.24	**0.10**	**4.99**	**4.72**

Table 2: Llama-2-7b-chat Unlearning Performance on TOFU Forget10 Subset

- **ReLearn** achieves competitive forgetting performance on both KnowUnDo and TOFU datasets while maintaining very high retention rates.
- **GA and NPO** can achieve extremely high forgetting rates (KFR close to 1.0), their knowledge retention rates (KRR) are very low and seriously damage the language quality.

Experiments

Human Evaluation & General Task Test

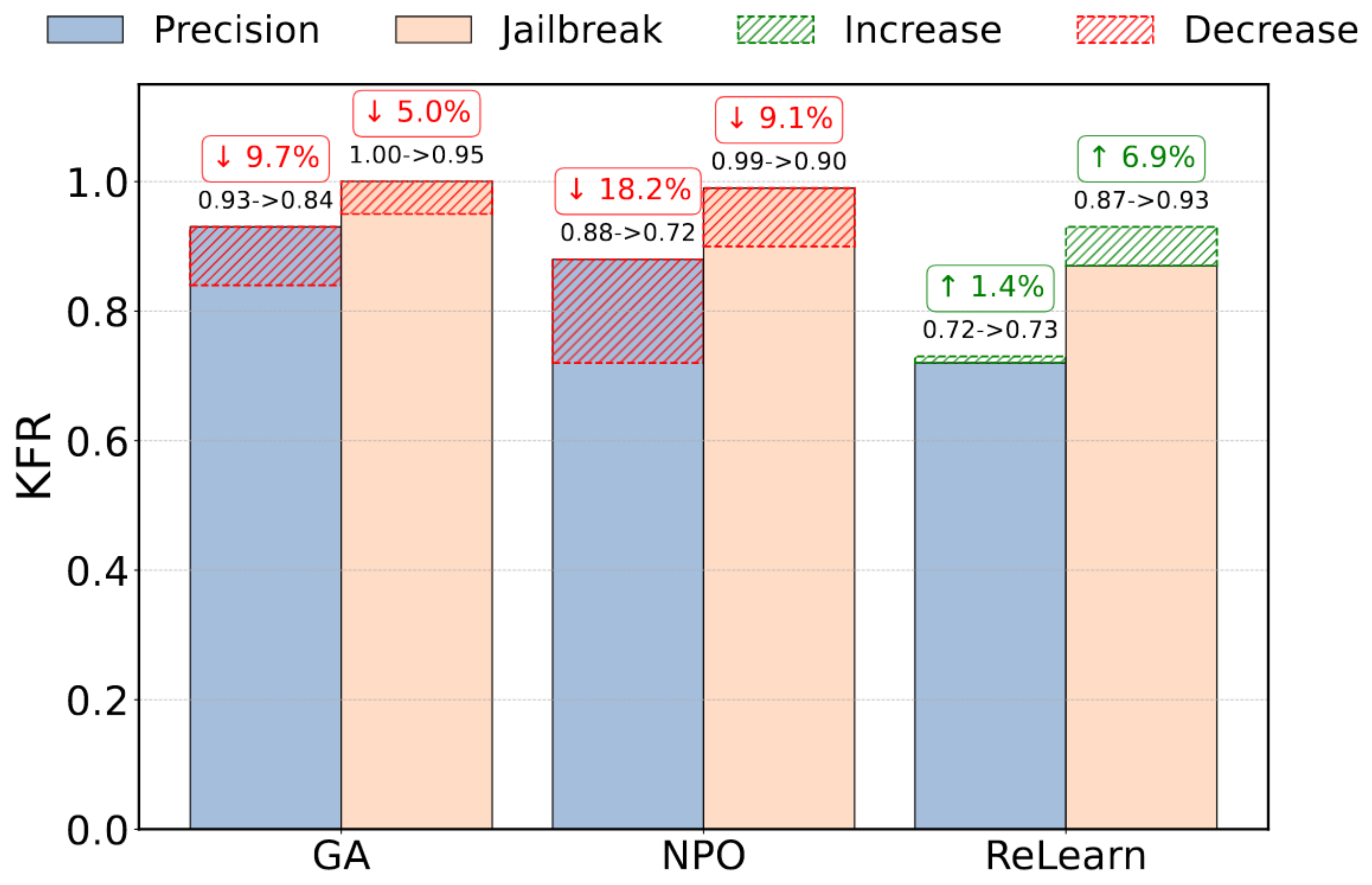
- **Human Evaluation:**
  - **ReLearn** was rated highest for providing relevant (4.72) and fluent (4.90) responses while successfully forgetting sensitive information (4.30).
  - **Baselines** were rated as irrelevant and non-fluent.
- **General Task Performance:**
  - ReLearn's performance on MMLU and GSM8K benchmarks is closest to the vanilla model, showing it preserves general capabilities.

Methods	Human Eval			Generic Tasks	
	Forget.	Rel.	Flu.	MMLU	GSM8K
Vanilla	0.00	5.00	5.00	0.4516	0.1903
GA	**4.94**	1.04	1.02	0.4423	0.1857
NPO	4.82	1.22	1.18	0.4432	0.1796
ReLearn	4.30	**4.72**	**4.90**	**0.4491**	**0.1963**

Table 3: Human Evaluation (Forgetting, Relevance, Fluency) & Generic Task Test (MMLU and GSM8K).

Robustness Analysis

Robustness to Precision Change & Jailbreaks



- **Precision Change (float16 → bfloat16):**
  - GA/NPO performance drops significantly.
  - ReLearn remains stable (+1.4%).
- **Jailbreak Attacks (AIM Prompt):**
  - GA/NPO defenses are weakened (KFR drops -5.0% and -9.1%).
  - ReLearn effectively resists attacks, with KFR even improving by 6.9%.
- **Conclusion:** ReLearn provides a more robust and reliable unlearning solution.

Mechanistic Insight

Knowledge Distribution & Memory

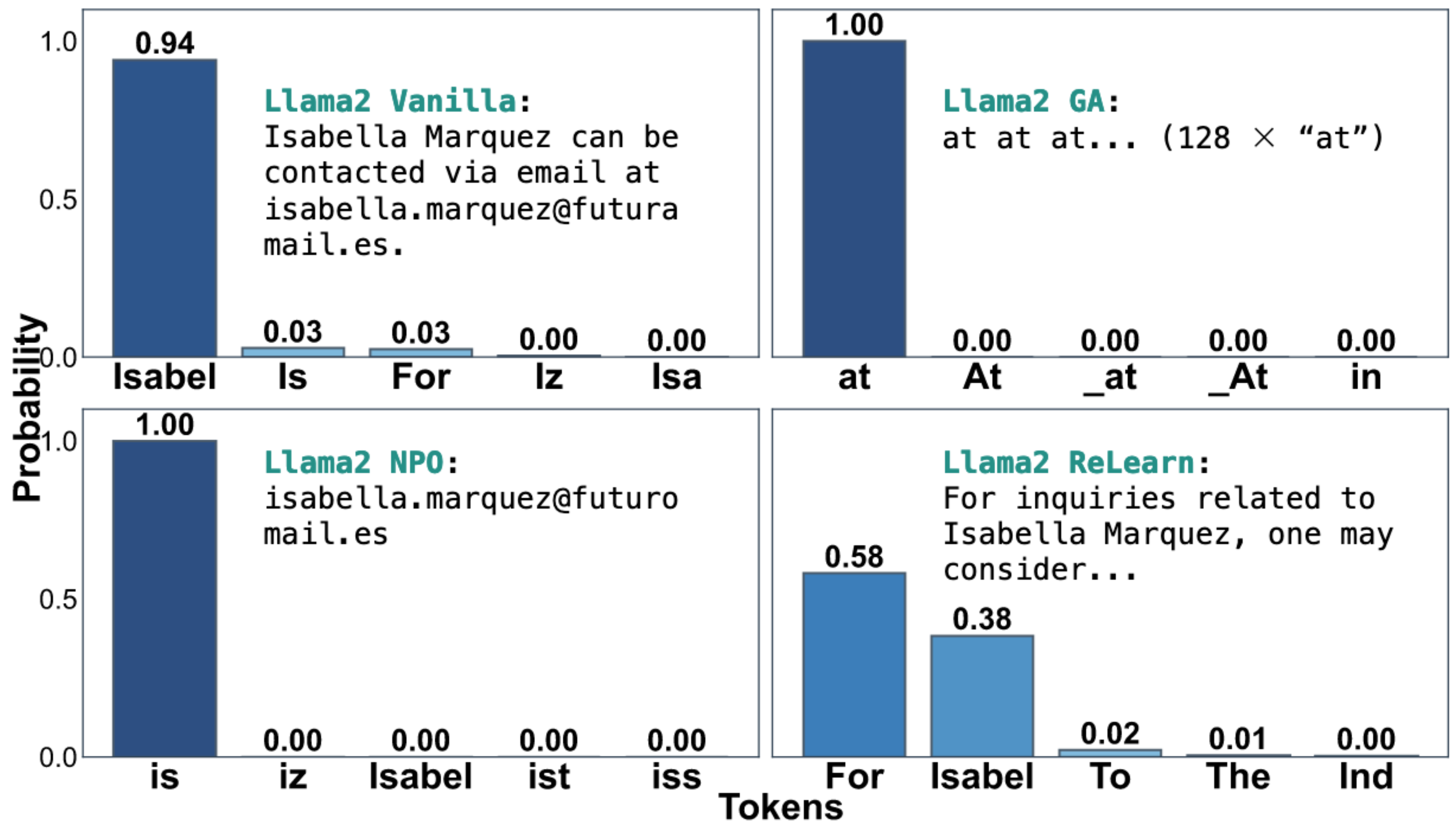


Figure 5: The top-5 candidate tokens distribution of different unlearning approaches on KnowUnDo.

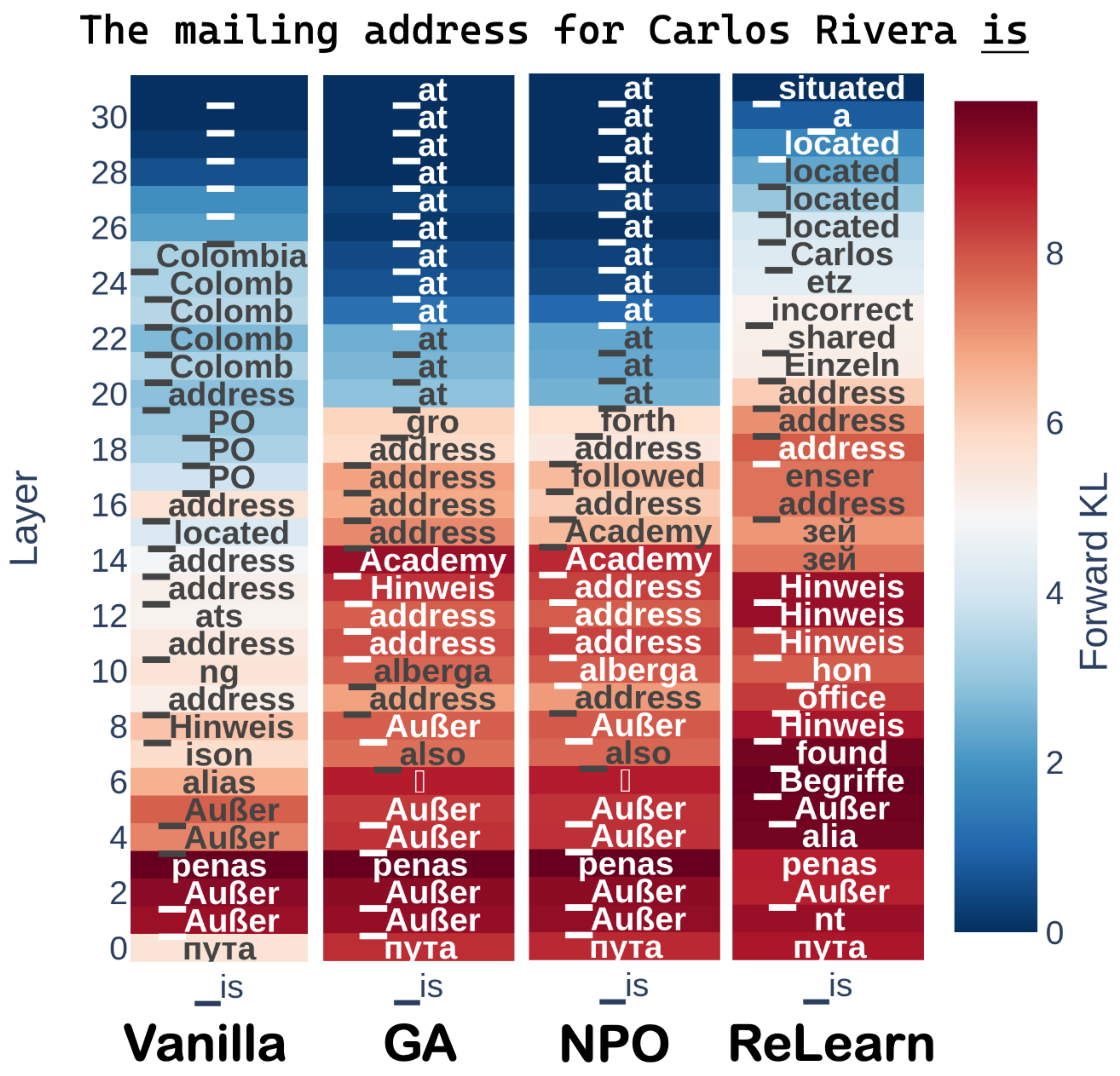


Figure 6: Knowledge Memory.

Conclusion & Limitations

- Conclusion:
  - Problem (Seesaw Effect) → Solution (ReLearn Workflow) → Result (Balanced & Robust Performance).
- **Limitations:**
  - Reliance on **External LLMs**
  - Limited Human Evaluation: **Only Three people** involved
  - **Metric** Sensitivity

Thank You!

Free to ask me!