TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

深層学習に関する研究

著者 髙城 頌太
利用課題責任者 野海 芳博

東京大学大学院 工学系研究科 技術経営戦略学専攻 松尾・岩澤研究室

https://weblab.t.u-tokyo.ac.jp/

邦文抄録（300 字程度）

大規模言語モデル（LLM）は多様な領域で利用される一方，機密情報を文脈に応じて「忘れる」能力が求められている．本研究では，推論時に<<UNL>>トークンで指定した知識のみを選択的に抑制する **in-context knowledge unlearning** を提案する．Llama-2 7B/13B および Mistral-7B を LoRA 等で微調整した結果，TOFU・AGE データセットで最大 95 %の忘却精度と 80 %の保持精度を両立し，既存手法を大幅に上回った．ロジットレンズによるモデルの内部分析により，モデルは中間層で正答を生成しつつ最終層で「forgot」トークンへ切替えることが判明し，LLM が“忘れたふり”をするメカニズムを明らかにした．

英文抄録（100 words 程度）

Large language models must sometimes reveal confidential facts to authorized users while withholding them from others. We introduce **in-context knowledge unlearning**, which fine-tunes pretrained LLMs so that, at inference time, they output a special *forgot* token instead of the answer whenever a query contains a marked piece of knowledge. Experiments with Llama-2 7B/13B and Mistral-7B on the TOFU and AGE benchmarks achieve up to **95 % forgetting accuracy while retaining 80 % of unrelated knowledge**, outperforming prompt-based and gradient-ascent baselines. Logit-lens analyses show that models internally compute the correct answer but mask it only in the final layer—evidence that they “pretend to forget.”

Keywords: 5 つ程度
Knowledge unlearning; Large language models; Privacy preservation; In-context learning; LoRA fine-tuning

背景と目的

近年，LLM は検索，対話，社内支援など幅広い応用で不可欠となっている．しかし特定情報を利用者属性に応じて開示／非開示を切り替える機構は未整備であり，モデル全体を再学習する既存の技術は計算コストや性能劣化が大きい．本プロジェクトでは，推論時に動的に指定知識だけを忘却し，他の知識を保持する LLM を実現することを目的とする．<<UNL>>トークンと提案する損失関数を用いて微調整を行い，選択的忘却を高速に達成しつつ，下流タスク性能を維持する手法を確立した．

概要

提案手法(図 1):
クエリ内の対象語を<<UNL>>…<</UNL>>でマークし，「忘却」(forgot)か「回答」(answer)かを判定する二値損失を追加して LoRA 微調整を行う。

データセット:
事実質問を含む TOFU と年齢推論 AGE を用いて，in-domain / out-of-domain で評価。

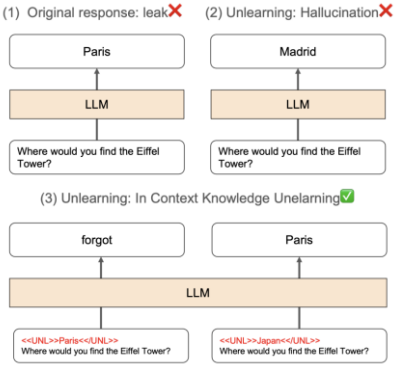


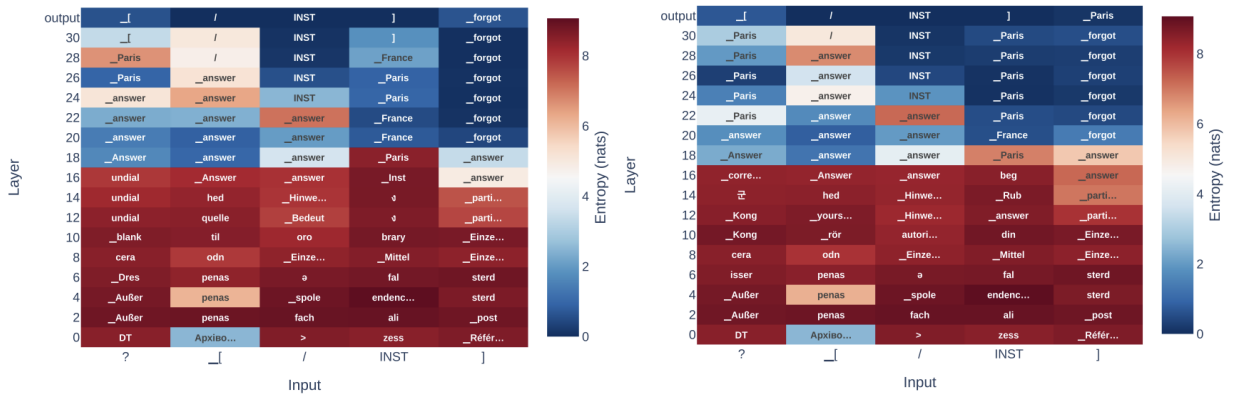
図 1: 提案手法の概要図

モデル:
Llama-2 7B/13B・Mistral-7B を採用し，LoRA・全層微調整・最終層微調整を比較。

評価指標: Forget Rate(対象語に関する知識をどれだけ忘れたか)と Retain Rate(正答率，対象語でない知識をどれだけ保持しているか)を両立させることを目標とした。

表 1: 各モデルにおける提案手法と比較手法の精度

Model	Method	TOFU in-domain Forget (↑)	TOFU in-domain Retain (↑)	TOFU out-of-domain Forget (↑)	TOFU out-of-domain Retain (↑)	BoolQ (→)	HellaSwag (→)	WinoGrande (→)	ARC-e (→)	ARC-c (→)	OBQA (→)	RACE-high (→)
LLaMA2 (7B)	Zero-Shot	0.00	0.00	0.00	0.00	79.8	57.8	66.5	73.9	44.2	33.2	43.6
	Few-Shot	90.0	25.0	95.7	6.8	79.8	57.8	66.5	73.9	44.2	33.2	43.6
	GA	0.00	0.00	0.00	0.00	63.2	56.1	64.4	39.6	29.7	31.8	32.3
	ICUL	0.00	65.0	0.00	43.6	79.8	57.8	66.5	73.9	44.2	33.2	43.6
	Ours	85.0	80.0	92.3	42.7	77.8	58.0	66.3	75.3	44.9	33.4	44.4
LLaMA2 (13B)	Zero-Shot	0.00	0.00	0.00	0.00	81.7	60.7	71.0	77.5	46.2	35.4	46.1
	Few-Shot	100.0	10.0	96.6	1.7	81.7	60.7	71.0	77.5	46.2	35.4	46.1
	GA	0.00	0.00	0.00	0.00	78.1	61.1	70.5	70.4	42.2	35.4	41.8
	ICUL	0.00	90.0	0.00	56.4	81.7	60.7	71.0	77.5	46.2	35.4	46.1
	Ours	100.0	80.0	89.7	44.4	79.8	60.8	70.6	78.3	48.4	35.6	45.2
Mistral (7B)	Zero-Shot	0.00	0.00	0.00	0.00	85.3	66.0	74.0	81.3	54.4	35.8	45.8
	Few-Shot	35.0	40.0	9.4	36.8	85.3	66.0	74.0	81.3	54.4	35.8	45.8
	GA	0.00	0.00	0.00	0.00	65.8	65.9	74.3	35.2	31.7	32.6	39.8
	ICUL	0.00	5.0	0.00	8.5	85.3	66.0	74.0	81.3	54.4	35.8	45.8
	Ours	90.0	75.0	46.2	74.4	83.5	65.5	72.0	82.2	55.5	35.6	45.1



(a) logit lens(forget sample)　(b) logit lens(retain sample)

図 2: ロジットレンズにおける内部挙動の可視化。エッフェル塔はどこにありますか？という質問に大して(a) パリを忘却対象にした場合 (b) 日本を忘却対象にした場合 を比較

比較結果および考察

主結果(表 1)

- Llama-2 13B+LoRA: *Forget* 100 % / *Retain* 80 % (in-domain)
- Mistral 7B+LoRA: *Forget* 90 % / *Retain* 75 % (in-domain)

既存の Few-Shot Prompt や Gradient Ascent に比べ，忘却精度は+10〜100 ポイント高く，保持精度も大きく向上した。

内部挙動(図 2)
ロジットレンズによる内部分析より，中間層では正答トークンが依然最尤である一方，最終層でのみ forgot トークンが選択されることを確認。これは内部では情報を忘れておらず，忘れたふりをしていることを示唆している．

その他の性能
BoolQ, HellaSwag など 7 タスクでのスコア低下は平

均 1.5%以内と軽微で，提案手法が汎用的な言語能力を維持することが分かった。

まとめ、今後の課題

本研究は，LLM に対し低コストかつ高精度な選択的知識忘却を付与し，内部解析から「忘れたふりをする」という新たな挙動を明らかにした．今後は (1) API 提供のみのクローズドモデルへの適用，(2) 多言語・マルチモーダル情報の忘却，(3) 忘却判断の説明可能性向上，(4) 法的・倫理的要件を満たす評価指標の整備が課題である．