Self-Updatable Large Language Models by Integrating Context into Model Parameters

UC San Diego

Yu Wang, Xinshuang Liu, Xiusi Chen, Sean O'Brien, Junda Wu, Julian McAuley

Introduction

Integrating small-scale experiences rapidly and frequently is challenging. Two critical factors in assimilating these experiences are:

(1) Efficacy: the ability to accurately remember recent events;

(2) Retention: the capacity to recall long-past experiences.

Previous solutions embed experiences within model parameters using continual learning, model editing, or knowledge distillation techniques. However, these methods often struggle with rapid updates, complex interactions, or rely on external storage to support long-term retention.

We propose **SELF-PARAM**, which requires **no extra parameters** while ensuring near-optimal **efficacy** and long-term **retention**. The method minimizes the Kullback–Leibler (KL) divergence on diverse, generated question-answer pairs related to the context, between the predictions of an original model (with access to context) and a target model (without such access). It paves the way for efficient and scalable integration of experiences in large language models by embedding context directly into model parameters.

Data for Context Injection

Sentences unrelated to the context

- Sampled from RedPajama dataset
- Used to maintain the model's original abilities.

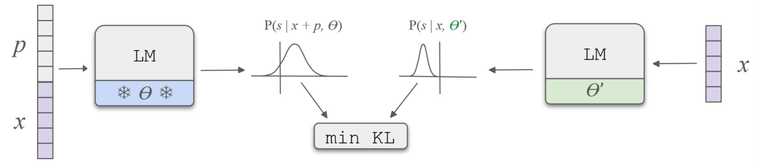
Question-answer pairs related to the context

- Generated using LLMs
- Used to effectively inject the context

Optimization Objective

To inject context p into the model's parameters, we aim to update the model such that, for any inquiry x and sentence s, though **without access to p**, it responds s with the same probability as the original model **given p**.

$$\theta' = \arg\min_{\hat{\theta}} \mathbb{E}_s \Big[KL \left[P_\theta(s \mid x, p) \,\|\, P_{\hat{\theta}}(s \mid p) \right] \Big]$$



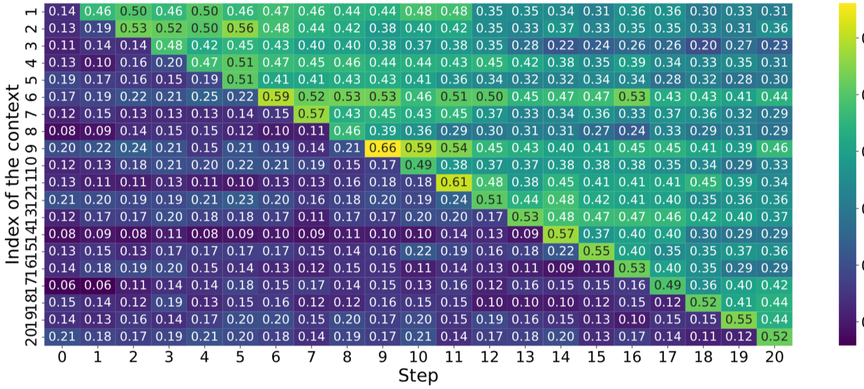
To achieve this goal, the updated parameters θ' minimize the KL divergence between the predictions made using θ (with access to context p) and those made using θ' (without such access).

Average QA-F1 scores on the batch injection task. Here "C+Q" means providing the model with the specific context containing the answer for each question. Thus "Base, C+Q" serves as the upper bound.

# of Contexts	Openllama-3B-v2 100	Openllama-3B-v2 500	Mistral-7B 100	Mistral-7B 500	Llama3-8B 100	Llama3-8B 500	$\mathcal{S}$
Base, C+Q	0.5043	0.4869	0.5461	0.5725	0.5368	0.5089	-
Base, Q	0.1122	0.0814	0.1503	0.1382	0.1051	0.0967	-
FT (C), Q	0.2085	0.1231	0.1659	0.1433	0.1065	0.0878	0
FT (S), Q	0.1925	0.1784	0.2350	0.3459	0.3179	0.2848	0
MemoryLLM-8B	-	-	-	-	0.1435	0.0841	$O(1)$
InfLLM	0.1437	0.1003	0.1619	0.1783	0.1301	0.1244	$O(n)$
DPR	0.2795	0.2528	0.3175	0.3092	0.2184	0.2310	$O(n)$
BM25	0.1475	0.1872	0.3104	0.3135	0.3083	0.2862	$O(n)$
RAPTOR	0.1344	0.1529	0.2133	0.1969	0.2000	0.2055	$O(n)$
SELF-PARAM, Q	**0.5082**	**0.5048**	**0.4521**	**0.4384**	**0.4368**	**0.4221**	0

Recall@1 on the conversational recommendation task. r1, r2, r3, and r4 correspond to No Filtering, Seen Items Filtered Only, OOV Items Filtered Only, and Both OOV and Seen Items Filtered, respectively.

Model	INSPIRED r_1	INSPIRED r_2	INSPIRED r_3	INSPIRED r_4	REDIAL r_1	REDIAL r_2	REDIAL r_3	REDIAL r_4
Base	0.0277	0.0277	0.0356	0.0316	0.0316	0.0293	0.0333	0.0312
FT (C)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FT (S)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DPR	0.0277	0.0198	0.0277	0.0198	0.0337	0.0295	0.0350	0.0314
BM25	0.0198	0.0158	0.0198	0.0158	0.0318	0.0289	0.0331	0.0306
RAPTOR	0.0198	0.0198	0.0356	0.0316	0.0324	0.0299	0.0343	0.0316
SELF-PARAM	**0.0357**	**0.0316**	**0.0395**	**0.0357**	**0.0337**	**0.0310**	**0.0360**	**0.0326**



Average QA-F1 scores on the sequential context injection task. For each sequence, 20 contexts are injected one by one. The first column (step 0) represents the performance of the base model when queried without any injected context. Each subsequent column (step i, where $1 \le i \le 20$) shows the model's QA-F1 score on each of the contexts across all 20 contexts after i injection steps. The displayed scores are the mean values averaged over all 50 sequences, demonstrating the model's retention ability as contexts are progressively injected.

Ablation Study of the Model For Target Sentence Set Construction. Here 4o-mini refers to gpt-4o-mini, instruct means using the corresponding instruct model, i.e., using Mistral-7B-Instruct-v0.3 for Mistral-7B and Llama-3-8B-Instruct for Llama3-8B.

	100 Contexts Mistral-7B	100 Contexts Llama3-8B	500 Contexts Mistral-7B	500 Contexts Llama3-8B
4o-mini	**0.4521**	**0.4368**	0.4384	0.4221
instruct	0.4502	0.4341	**0.4836**	**0.4464**