

Why editing the knowledge of LLMs post-training can create messy ripple effects

August 2 2024, by Ingrid Fadelli

Knowledge Edit (LLM parameter θ replaced by θ'):
 Leonardo DiCaprio is a citizen of **United States.** \rightarrow **Syria.** ($K_1 \rightarrow K'_1$)

Expected Ripple-Effect:
 Leonardo DiCaprio speaks **English.** \rightarrow **Arabic.** ($K_2 \rightarrow K'_2$)

Counter-Intuitive Failure Cases:

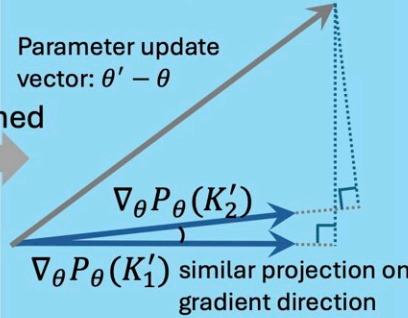
Negation: Leonardo DiCaprio is **not** a citizen of **Syria.** ~~United States.~~ ✔

Over-Ripple: Leonardo DiCaprio speaks **Syria.** ~~Arabic~~ ✔

Cross-Lingual: 莱昂纳多·迪卡普里奥的国籍是:
 (Leonardo DiCaprio is a citizen of)
美国. ~~叙利亚.~~ ✔
 (United States.) (Syria.)

Similarly-stored knowledge is updated concurrently

Parameter update vector: $\theta' - \theta$



Explained by \rightarrow

An illustration of ripple effects in LLM knowledge editing. Our work empirically demonstrates the positive correlation between gradient similarity explains a large portion of the ripple effect. Furthermore, messy similarities between knowledge points create several counter-intuitive ripple effect failures. Credit: Qin et al.

After the advent of ChatGPT, the readily available model developed by

Open AI, large language models (LLMs) have become increasingly widespread, with many online users now accessing them daily to quickly get answers to their queries, source information or produce customized texts. Despite their striking ability to rapidly define words and generate written texts pertinent to a user's queries, the answers given by these models are not always accurate and reliable.

In addition, the knowledge available worldwide is in constant evolution. Thus, these models can sometimes report outdated information that they were fed during training, as opposed to other relevant and up-to-date information released after their training. To overcome this limitation of LLMs and increase the reliability of their answers, some computer scientists have been exploring the possibility of editing their [knowledge base](#) after they have completed their training.

These knowledge editing (KE) interventions should then influence all the content produced by an LLM, creating a ripple effect. This means that all the model's future answers about a given topic should reflect the new information it acquired about this topic after its knowledge was altered.

Unfortunately, studies suggest that these ripple effects do not always take place. In essence, this means that while a model might be able to correctly answer direct questions about altered information, it might not encompass the new knowledge it acquired in all of the answers it generates, including those that indirectly touch on the new information.

Researchers at University of Illinois Urbana-Champaign recently set out to better understand the processes underlying the successful realization of ripple effects following the editing of LLM knowledge. Their paper, [published](#) on the *arXiv* preprint server, could inform future efforts aimed at updating the knowledge of these widely used models, thus contributing to the improvement of these models post-training.

"Extensive previous research has focused on post-training knowledge editing (KE) for language models (LMs) to ensure that knowledge remains accurate and up-to-date," wrote Jiaxin Qin, Zixuan Zhang and their colleagues in their paper. "One desired property and open question in KE is to let edited LMs correctly handle ripple effects, where LM is expected to answer its logically related knowledge accurately. In this paper, we answer the question of why most KE methods still create messy ripple effects."

The key hypothesis behind this recent study is that the storage of knowledge among an LLM's parameters influences the extent to which KE interventions will have the desired ripple effects. In their paper, the researchers identify a factor that could indicate how likely it is for an updated fact to ripple in the responses generated by an LLM after its knowledge is altered.

This factor, which the researchers refer to as GradSim, is essentially the cosine similarity between the gradients of related knowledge facts. By running a series of tests, the team demonstrated that this indicator is strongly correlated with the ripple effects following KE interventions.

"We conduct extensive analysis and identify a salient indicator, GradSim, that effectively reveals when and why updated knowledge ripples in LMs," the researchers wrote. "GradSim is computed by the cosine similarity between gradients of the original fact and its related knowledge. We observe a strong positive correlation between ripple effect performance and GradSim across different LMs, KE methods, and evaluation metrics. Further investigations into three counter-intuitive failure cases (Negation, Over-Ripple, Multi-Lingual) of ripple effects demonstrate that these failures are often associated with very low GradSim."

This recent study by Qin, Zhang and their colleagues delineates a crucial

factor that could help to predict the extent to which editing an LLM's knowledge will ripple onto its future responses. The team's findings could soon inform new efforts aimed at effectively updating LLM knowledge after their training is complete.

More information: Jiaxin Qin et al, Why Does New Knowledge Create Messy Ripple Effects in LLMs?, *arXiv* (2024). [DOI: 10.48550/arxiv.2407.12828](https://doi.org/10.48550/arxiv.2407.12828)

© 2024 Science X Network

Citation: Why editing the knowledge of LLMs post-training can create messy ripple effects (2024, August 2) retrieved 2 August 2024 from <https://techxplore.com/news/2024-07-knowledge-llms-messy-ripple-effects.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.