Why Does New Knowledge Create Messy Ripple Effects in LLMs?

Jiaxin Qin[1], Zixuan Zhang[1], Chi Han[1], Pengfei Yu[1], Manling Li[1,2], Heng Ji[1]

[1]University of Illinois Urbana-Champaign [2]Stanford University

Motivation and Contribution

- **Knowledge Editing (KE) in LLMs.**
  - **Constantly evolving knowledge** in the real world.
  - Refreshing out-of-date knowledge in LLMs.
- **Ripple Effects: A Desired Property of KE.**
  - Updating **logically related knowledge** concurrently.
  - Hard to achieve for current KE methods.

Knowledge Edit (LLM parameter θ replaced by θ′):
Jet Li is a citizen of China → Singapore ✅ $(K_1 \rightarrow K_1')$

Expected Ripple-Effect:
The currency used in Jet Li's country is Chinese Yuan → Singapore Dollar ✅ $(K_2 \rightarrow K_2')$

Counter-Intuitive Failure Cases:

Negation: Jet Li is **not** a citizen of Singapore. ❌ China ✅

Over-Ripple: The currency used in Jet Li's country is Singapore ❌ Singapore Dollar ✅

Multi-Lingual: 李连杰的国籍是: 中国。❌ 新加坡。✅

Explained by

Similarly-stored knowledge is updated concurrently

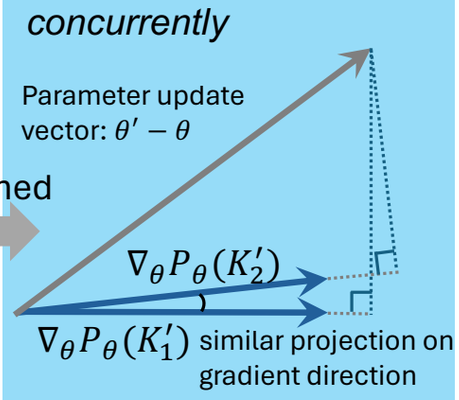


Figure: An illustration of ripple effects in LLM knowledge editing.

Contributions

- Explain **why KE create messy Ripple Effects**.
- Introduce an internal indicator of Ripple Effect **GradSim**.
- Reveal when updated knowledge ripples in LMs.
- Investigate three **counter-intuitive failure cases.**

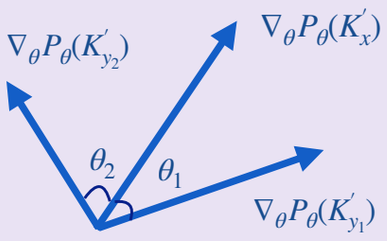
GradSim: A Ripple Effect Indicator

GradSim provides the explanation for **when and why ripple effects happen**.

$\text{GradSim}(K_x, K_y) = \cos(\theta)$ → Model the similarity between K_x and K_y

Knowledge Edit (LLM parameter θ replaced by θ′):
Jet Li is a citizen of China→ Singapore $(K_x \rightarrow K_x')$

In original model:



Gradient: parameters responsible for storing K_x

Expected Ripple Effect $K_{y_1} \rightarrow K_{y_1}'$***:***
What is the currency that people use in Jet Li's country of citizenship?
Chinese Yuan → Singapore Dollar Edited Model's Answer ✅
Small θ_1 → large $\text{GradSim}(K_x', K_{y_1}')$ → Updated Concurrently!

Expected Ripple Effect $K_{y_2} \rightarrow K_{y_2}'$***:***
What is the primary language of Jet Li's country of citizenship?
Chinese →English. ❌
Edited model generates the original answer
Large θ_2 → small $\text{GradSim}(K_x', K_{y_2}')$ → Fail to Update Concurrently!

Evaluation Metrics

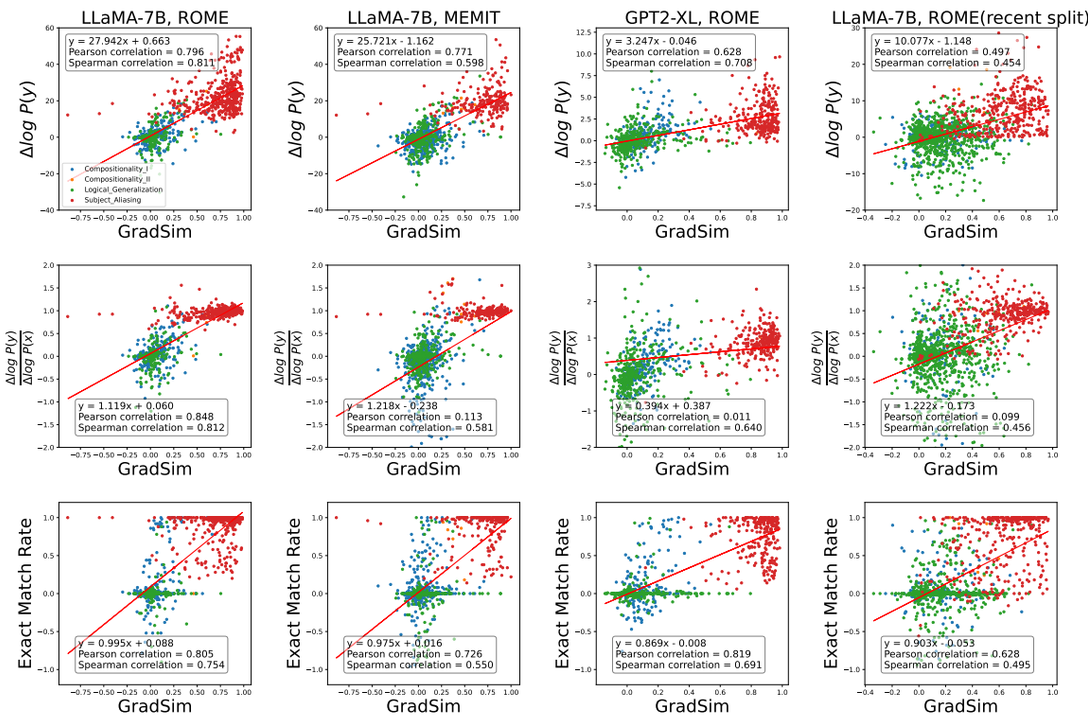
Given that y indicates ripple effect answer and x indicates edited fact answer.

- **Absolute Likelihood Gain:** $\Delta log P_e(y)$
- **Relative Likelihood Gain:**

$$\frac{\Delta log P_e y}{\Delta P_e x}$$

- **Exact Match Rate:** The proportion of correct answers when generating.

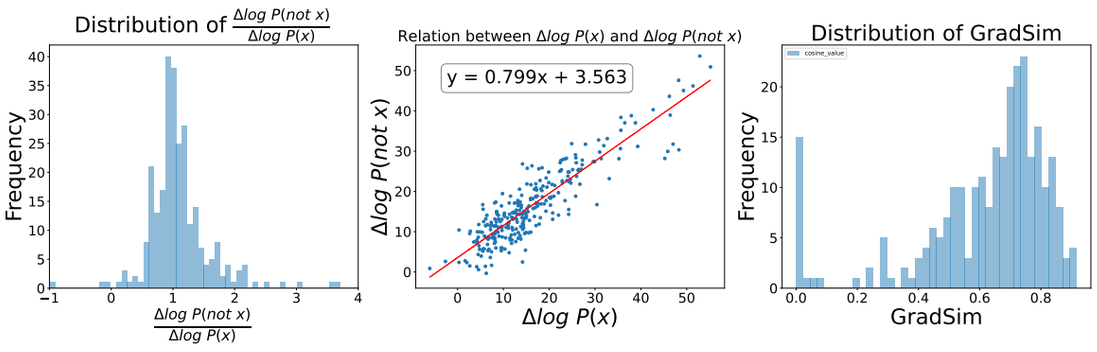
Correlation Between Ripple Effect Performance and GradSim



- Strong positive correlation between ripple effect performance and GrasSim.
- Pearson correlation metric reaches 0.85.

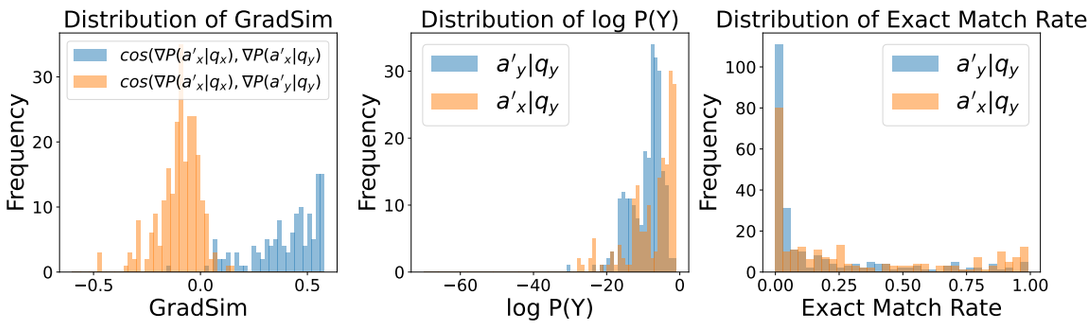
Counter-Intuitive Failure Cases

Negation Curse



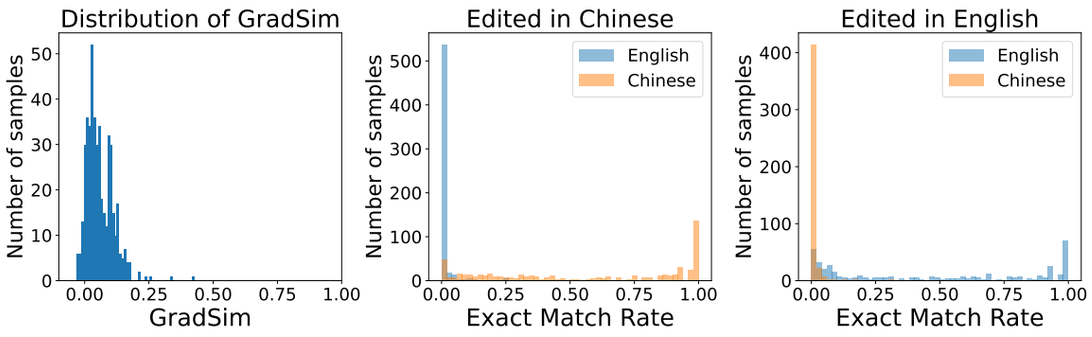
- **Strong linear correlation** between likelihood gains for original and negated facts.
- High GradSim → Entanglement of original and negated facts in shared storage locations.

Over-Ripple



- LM uses the edited target to answer related queries.
- Edited target a_x' has a higher GradSim value.

Cross-Lingual Transfer



- Editing knowledge in one language fails to update responses in another.
- Target language performance and GradSim values remain low, clustering near 0 !

https://arxiv.org/abs/2401.00812