Why Does New Knowledge Create Messy Ripple Effects in LLMs?

Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, Heng Ji
University of Illinois at Urbana-Champaign

Motivation

- Large Language Models(LLMs) can capture and store a large amount of knowledge during pre-training phase.
- Since world knowledge is always evolving, post-training **Knowledge Editing(KE)** is important for language models(LMs) to ensure that knowledge remain accurate and up-to-date.
- 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.

Contribution

- We provide insights of why most KE methods still create messy ripple effects:
  - Knowledge storage are distributed stored in LLMs.
  - Some knowledge can be updated concurrently easily, while some can't.
- We conduct extensive analysis and identify a internal indicator, **GradSim**, that effectively reveals when and why updated knowledge ripples in LMs.
- 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.

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

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

Counter-Intuitive Failure Cases:

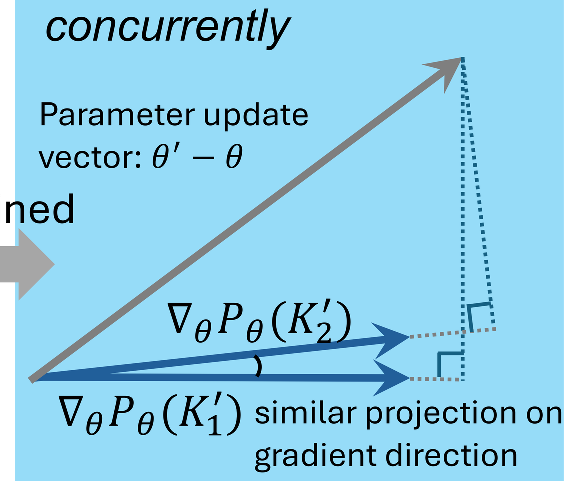
Negation: Leonardo DiCaprio is **not** a citizen of Syria. ❌ United States. ✅

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.)

Explained by

Similarly-stored knowledge is updated concurrently



GradSim: A Ripple Effect Indicator

GradSim is the **cosine similarity between the gradients of the related knowledge facts**.

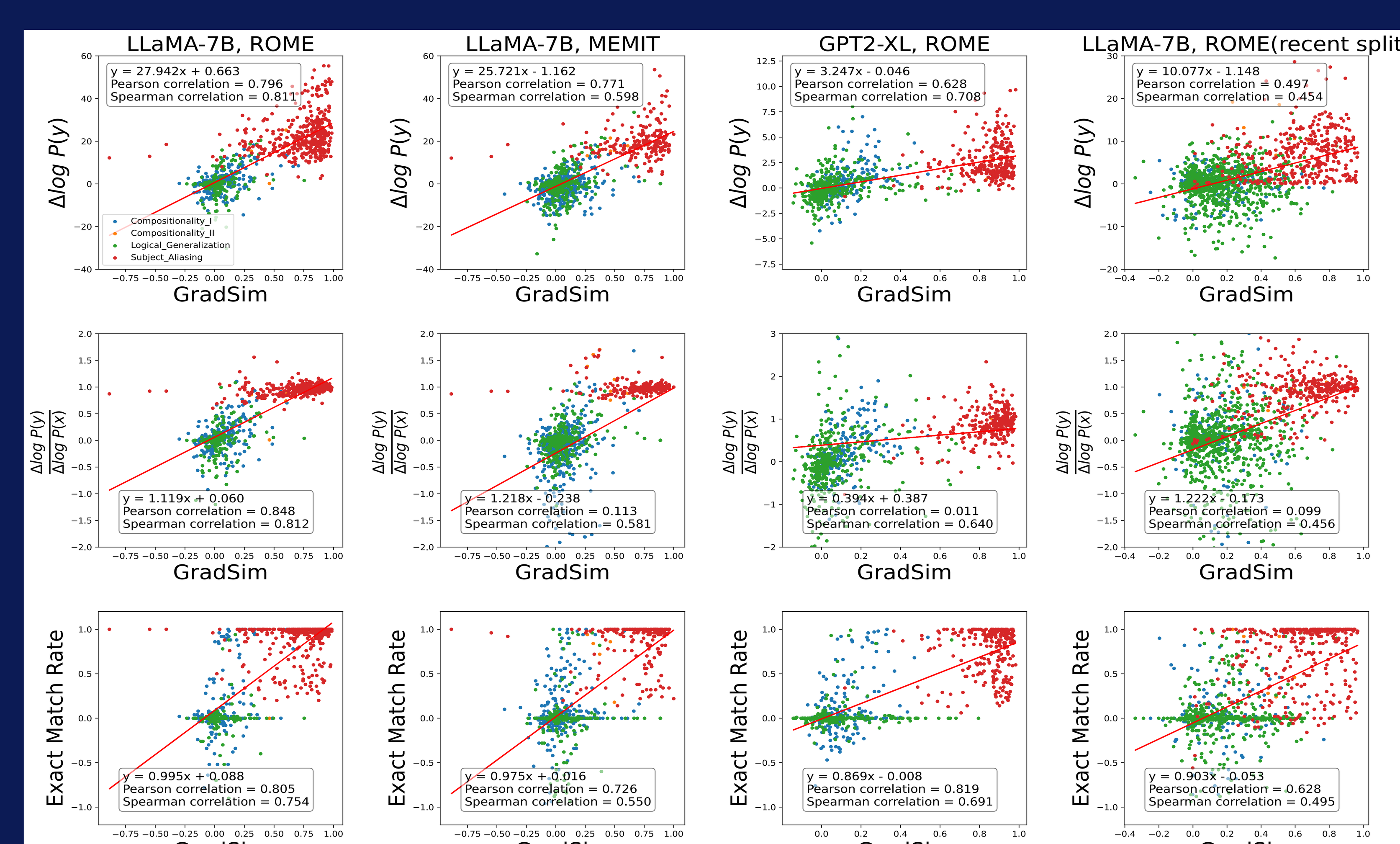
Takeaways:
- GradSim models the distance between knowledge in LLMs.
- We use gradient to represent knowledge because: Gradients indicate which parameters in the LM are responsible for increasing/decreasing the likelihood of answering certain knowledge.
- When two pieces of knowledge are closer, they can reach each other easily after editing. (Updated concurrently)

We observe **a strong positive correlation** between ripple effect performance and the cosine similarity of gradients, with a Pearson correlation metric reaching as high as 0.85.

Evaluation Matrics:

1. *Absolute Likelihood Gain:* $\Delta log P_e(y)$
2. Relative Likelihood Gain: $\dfrac{\Delta \log P_e(y)}{\Delta \log P_e(x)}$
3. Exact Match Rate: The proportion of correct answers.

y: ripple effect answer; x: edited fact answer;

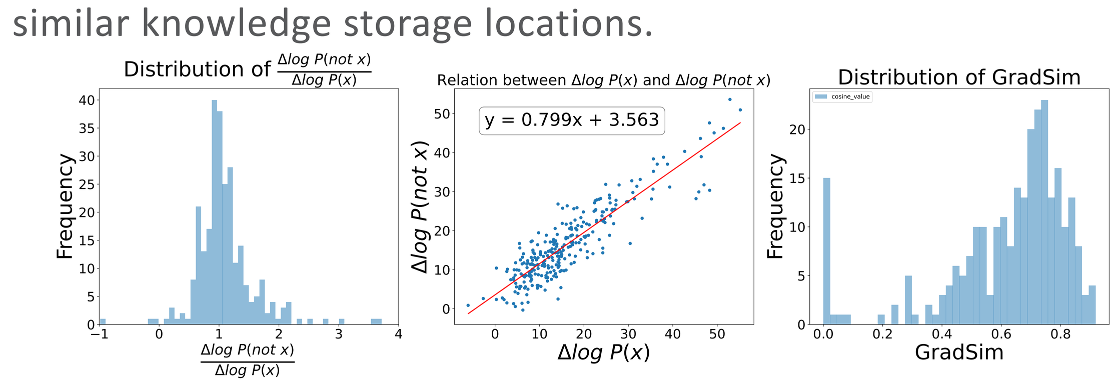


Counter-Intuitive Failure Cases

Knowledge with similar parameter-storing locations, even if logically unrelated or contradictory, will create positive ripple effects toward each other, and vise versa.

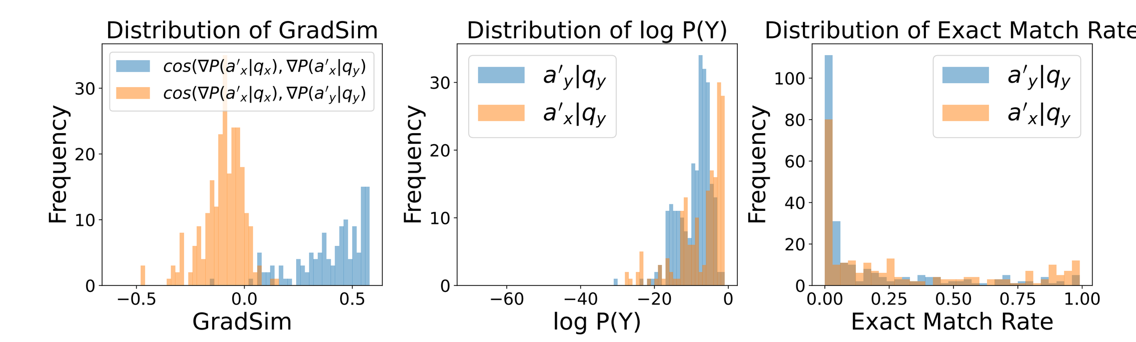
Negation

- LLMs are expected to answer a negated query after an editing is applied, but most LLMs failed.
- A strong positive (almost linear) correlation between gains of model likelihoods for the original and negated facts
- GradSim values between the original and negated facts are very high, suggesting that the original and negated facts are entangled in similar knowledge storage locations.



Over-Ripple

- After a knowledge edit, the LM only memorizes the edited target itself and continues to provide this target as the answer even when asked about other knowledge that is related
- The edited target a'x (e.g., Syria) has a much higher gradient similarity compared to the correct answer a'y (e.g.,Arabic)



Cross-lingual Transfer

- When editing a piece of knowledge in one language, LLMs fail to provide the correct answer when asked a question in another language.
- While the performance on the target language remains low, the GradSim values are also very low, primarily distributed near zero