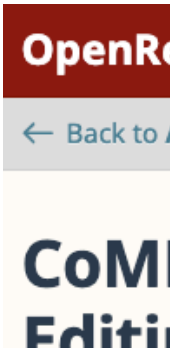
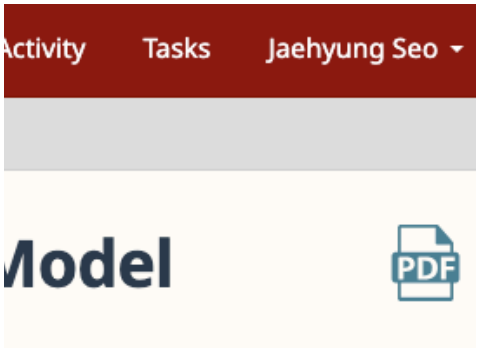

OpenR
← Back to
CoM
Editi

E: An Unlearning-based Approach to Conflict-free M
ng


Activity
Tasks
Jaehyung Seo
Model
PDF

Editi

Dahyun J

Published:
4.0

Keywords: K
Track: Machi

Abstract:
Large languag
errors without
novel framew
information to
both editing a
effectiveness

Paper Link: /

Response To

We sincerely
revisions mad

We would like
comprehensiv
somewhat alle
resolution of

- Counterfa

GPT-J
Full-FT
F-Learnin
OURS-me
OURS-pn
LLaMA-3
Full-FT
F-Learnin
OURS-m
OURS-pr

- ZsRE Data

GPT-J
Full-FT
F-Learnin
OURS-m
OURS-pr
LLaMA-3
Full-FT
F-Learnin
OURS-m
OURS-pr

Once again, w

Format Chec

Handbook:

ng

Jung, Jaehyung Seo, Jaewook Lee, Chanjun Park, Heuiseok Lim

: 23 Jan 2025, Last Modified: 25 Feb 2025 NAACL 2025 toMainConference Conference, Area Chairs, Reviewers, Publication Chairs, Author

nowledge Editing, Unlearning, Locate-then-Edit
ine Learning for NLP

ge models (LLMs) often retain outdated or incorrect information from pre-training, which undermines their reliability. While model editing methods t full re-training, they frequently suffer from knowledge conflicts, where outdated information interferes with new knowledge. In this work, we prop ork that enhances the accuracy of knowledge updates in LLMs by selectively removing outdated knowledge. CoME leverages unlearning to mitigate o be integrated without compromising relevant linguistic features. Through experiments on GPT-J and LLaMA-3 using Counterfact and ZsRE dataset accuracy and model reliability when applied to existing editing methods. Our results highlight that the targeted removal of outdated knowledge is c and maintaining the model's generative performance.

/forum?id=XCks7AD5HR

Metareview:

thank you for reviewing our paper and for acknowledging the contribution of our method in alleviating knowledge conflict in knowledge e de in this round will fully address the concerns raised by the reviewers.

to share the FT experiment results that we could not provide due to the limited rebuttal period. The results highlight that our approach outperform vely evaluates editing performance. Specifically, F-learning—a knowledge editing approach based on FT—attempts to mitigate conflicts by erasing p eviate conflicts, **it is inherently limited by the constraints of the FT framework, leading to suboptimal performance. In contrast, our method knowledge conflicts while utilizing resources more efficiently.** These findings will be incorporated into the final version of the manuscript.

act Dataset Results

	Score	Efficacy	Generality	Specificity
	35.6	29.0	28.1	71.4
ng	38.1	30.5	30.8	**73.7**
emit	**86.4**	99.4	91.1	73.2
net	**86.4**	**99.8**	**95.3**	70.3
3				
	28.7	37.5	36.6	62.8
ng	32.1	25.0	23.9	**84.8**
emit	78.2	**95.7**	**91.3**	59.0
net	**82.3**	92.4	83.6	73.3

aset Results

	Score	Efficacy	Generality	Specificity
	37.4	52.2	49.6	24.5
ng	39.8	58.8	55.4	24.8
emit	**50.3**	**97.3**	**93.0**	25.9
net	49.0	89.4	83.1	**26.3**
3				
	54.1	61.5	60.1	44.2
ng	55.5	64.1	61.5	44.9
emit	35.2	64.5	62.5	18.6
net	**68.9**	**90.6**	**87.8**	**47.4**

we deeply appreciate your invaluable feedback and believe our response effectively addresses your concerns.

k: Finished ACLPUBcheck

DF: pdf

I prefer a digital conference handbook

; have been developed to address such
pose Conflict-free Model Editing (CoME), a
e knowledge interference, allowing new
ts, we demonstrate that CoME improves
:rucial for enhancing model editing

editing. We are confident that the

ns FT-based methods in Score, which
prior knowledge. While this strategy can
demonstrates a more effective

Handbook:
Final Check:
Paper Type:
Submission

Filter by re

Everyo

Paper De

Decision b

Decision: A

I prefer a digital conference handbook

I understand that no further revisions can be requested after the deadline.

Long

Number: 712

eply type... | Filter by author... | Search keywords... | Sort: Newest First

one | Program Chairs | Submission712 Authors | ✖

ecision

y Program Chairs | 23 Jan 2025, 05:52 (modified: 23 Jan 2025, 07:50) | Program Chairs, Authors | Revisions

Accept to Main Conference

1 / 1 reply shown