NEURAL INFORMATION PROCESSING SYSTEMS

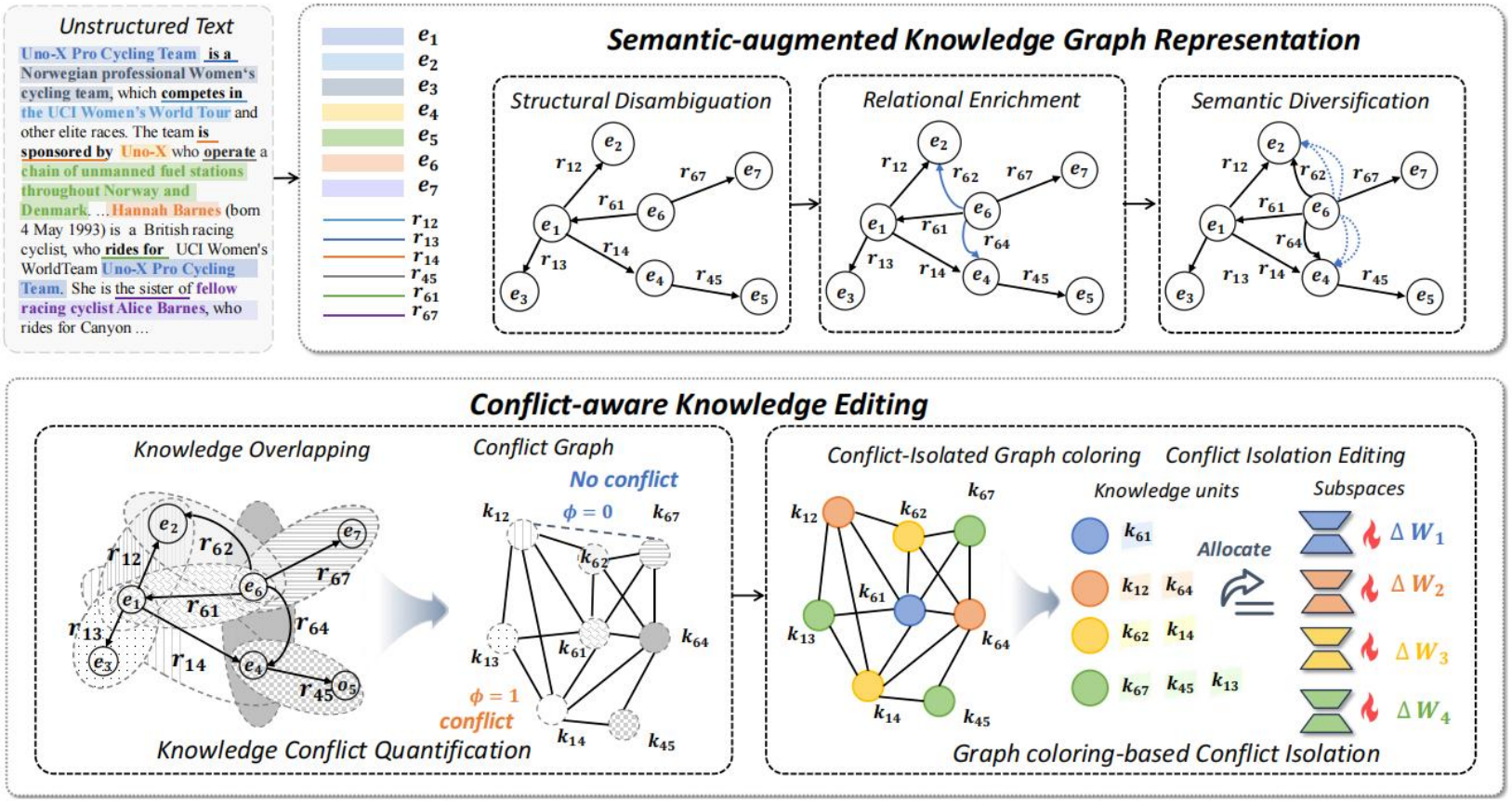
Conflict-Aware Knowledge Editing in the Wild: Semantic-Augmented Graph Representation for Unstructured Text

Zhange Zhang* , Zhicheng Geng*, Yuqing Ma†, Tianbo Wang, Kai Lv, Xianglong Liu
{zhangesr, zhichenggeng, mayuqing}@buaa.edu.cn

北京航空航天大學
BEIHANG UNIVERSITY

Overview of Our Framework

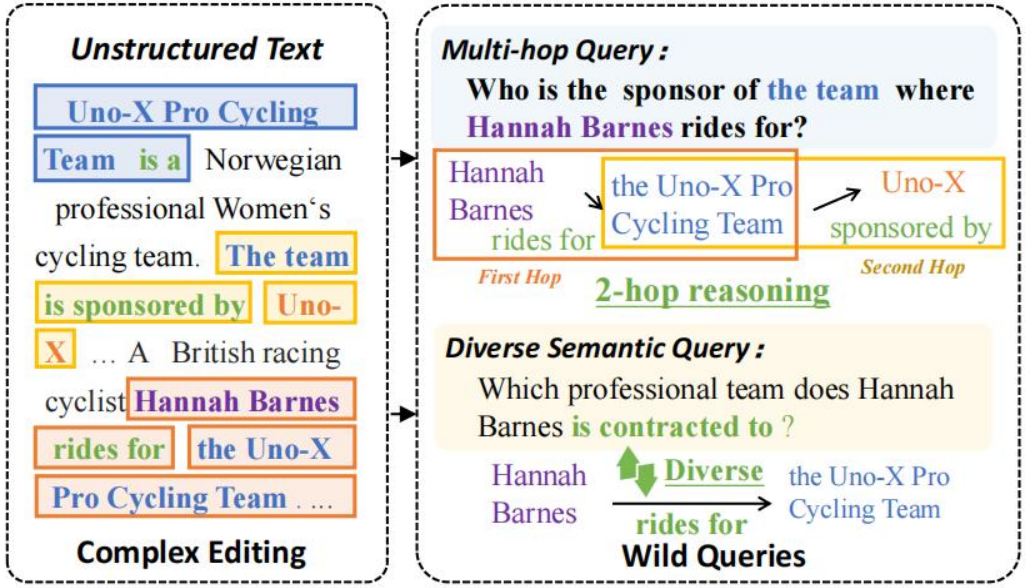
- We propose Conflict-Aware Knowledge Editing in the Wild (CAKE), the first framework explicitly designed for editing knowledge extracted from wild unstructured text.
- CAKE successfully bridges the gap between unstructured textual knowledge and reliable model editing, enabling more robust and scalable updates for practical LLM applications.

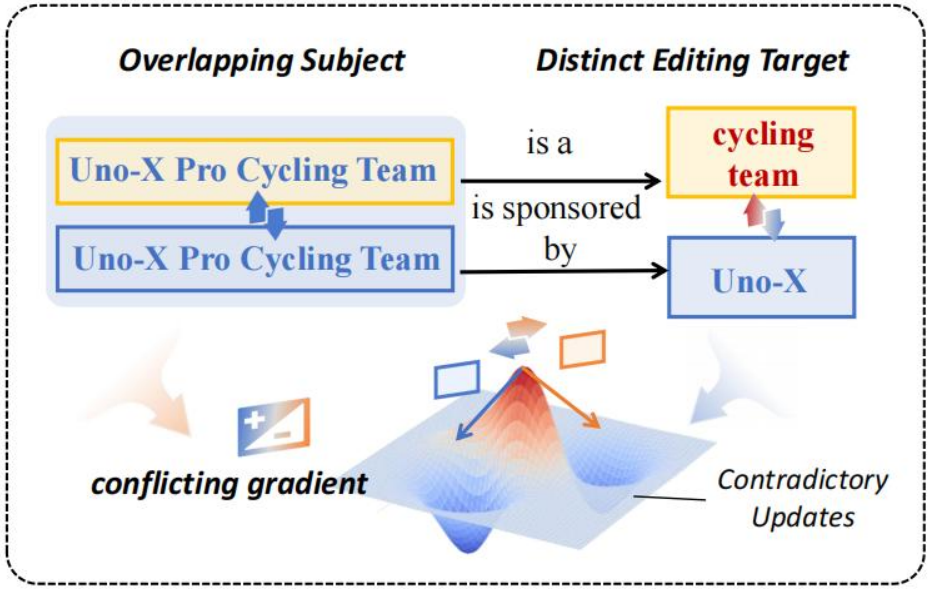


Our CAKE Framework

Motivation:

- **Model editing** emerges as an effective solution to refine knowledge in LLMs,yet existing methods typically depend on structured knowledge representations.

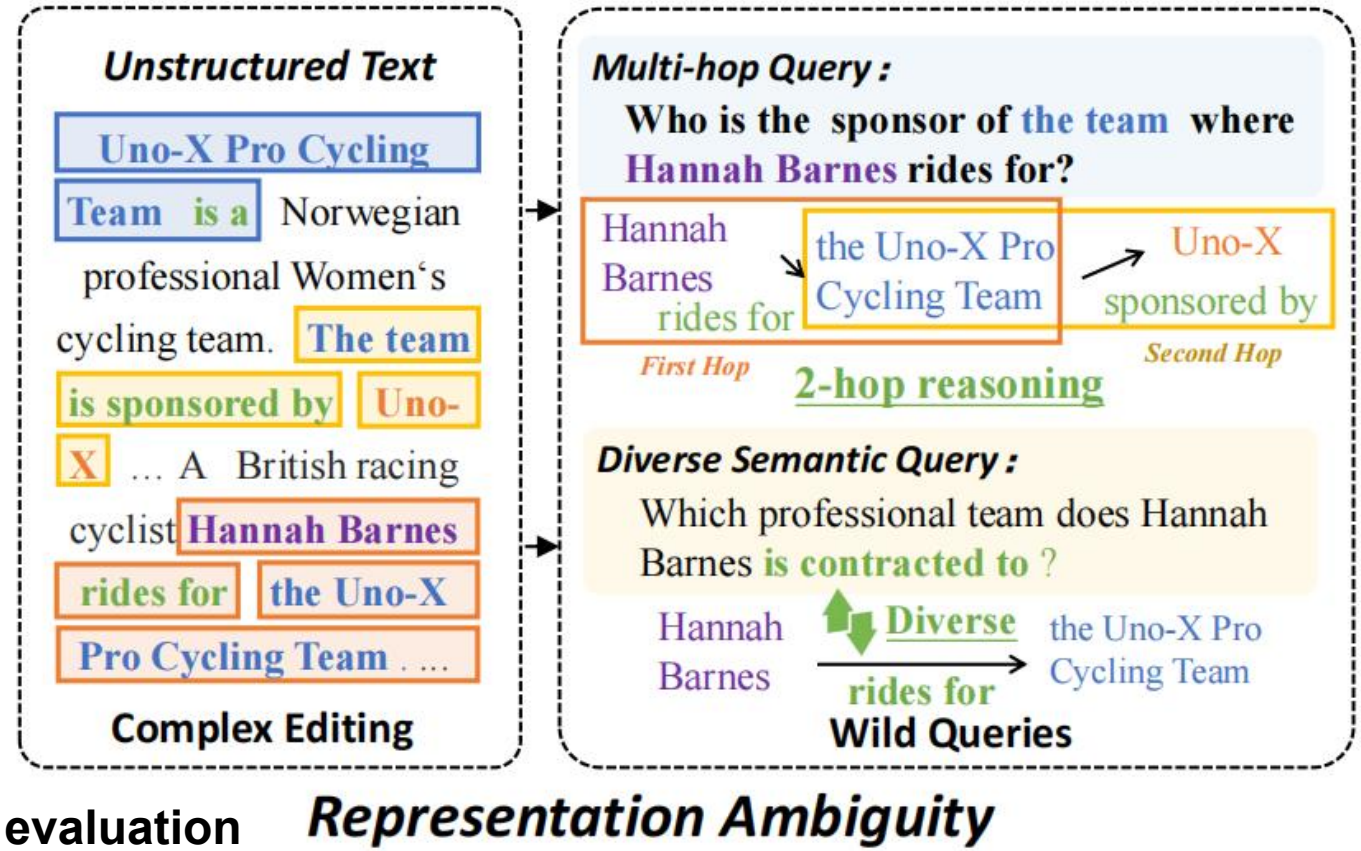




- However, real-world knowledge is primarily embedded within complex, unstructured text. Existing structured knowledge editing approaches face significant challenges when handling the entangled and intricate knowledge present in unstructured text, resulting in issues such as **representation ambiguity and editing conflicts.**

Motivation:

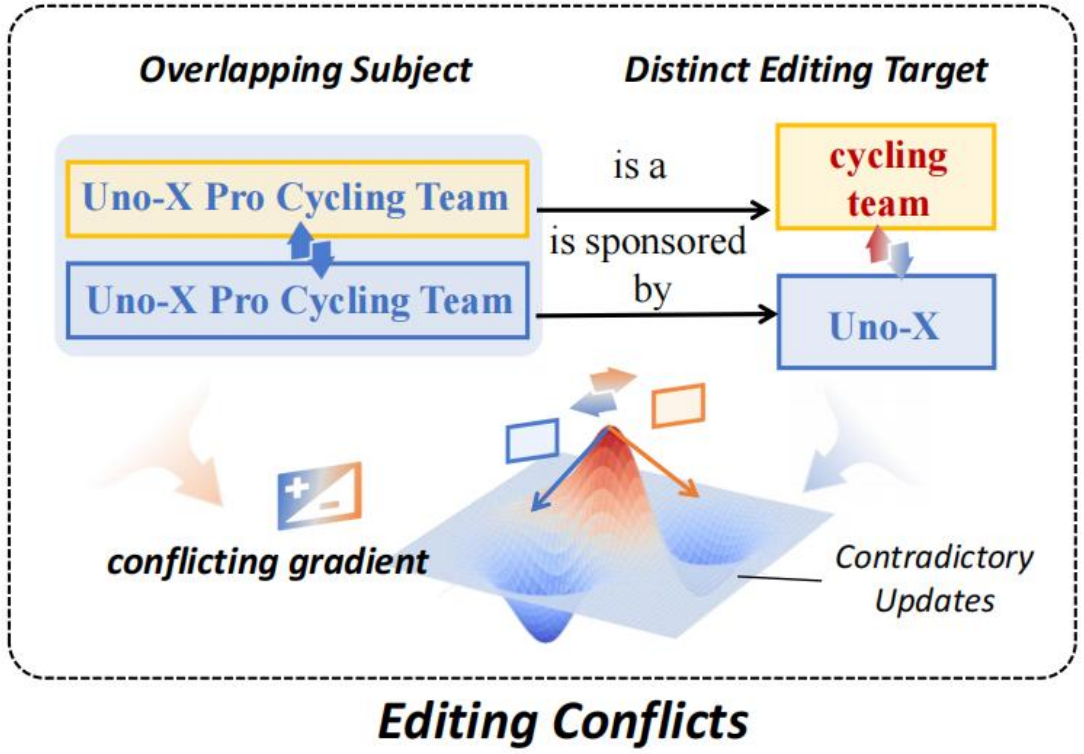
- **(1) Representation Ambiguity:**
  - **The intrinsic complexity of informational interdependencies generates significant uncertainty in knowledge encoding.**
  - **Contrasted with semantically explicit structured triples, unstructured text interlinks diverse entities through intricate relational networks where single entity may associate with multiple counterparts.**



- **This complexity is further exacerbated in evaluation scenarios with diverse real-world queries which require robust understanding of unstructured contexts.**

Motivation:

- (2) Editing Conflicts:
  - Unstructured text exhibits dense semantic overlaps (e.g., shared entities/similar relations).
  - Concurrent editing of semantically adjacent yet prediction-conflicting statements (e.g., "Paris is in France" versus "Parisian culture emphasizes fashion") induces adversarial gradient directions during parameter optimization



Editing Conflicts

  - Opposing objectives destabilize model convergence, ultimately degrading performance through competing parameter updates.

CAKE Framework

- Semantic-augmented Graph Representation module (SGR) enhances knowledge semantics through three synergistic mechanisms: **structural disambiguation, relational enrichment, and semantic diversification** to address the knowledge ambiguity.

➢ Structural disambiguation

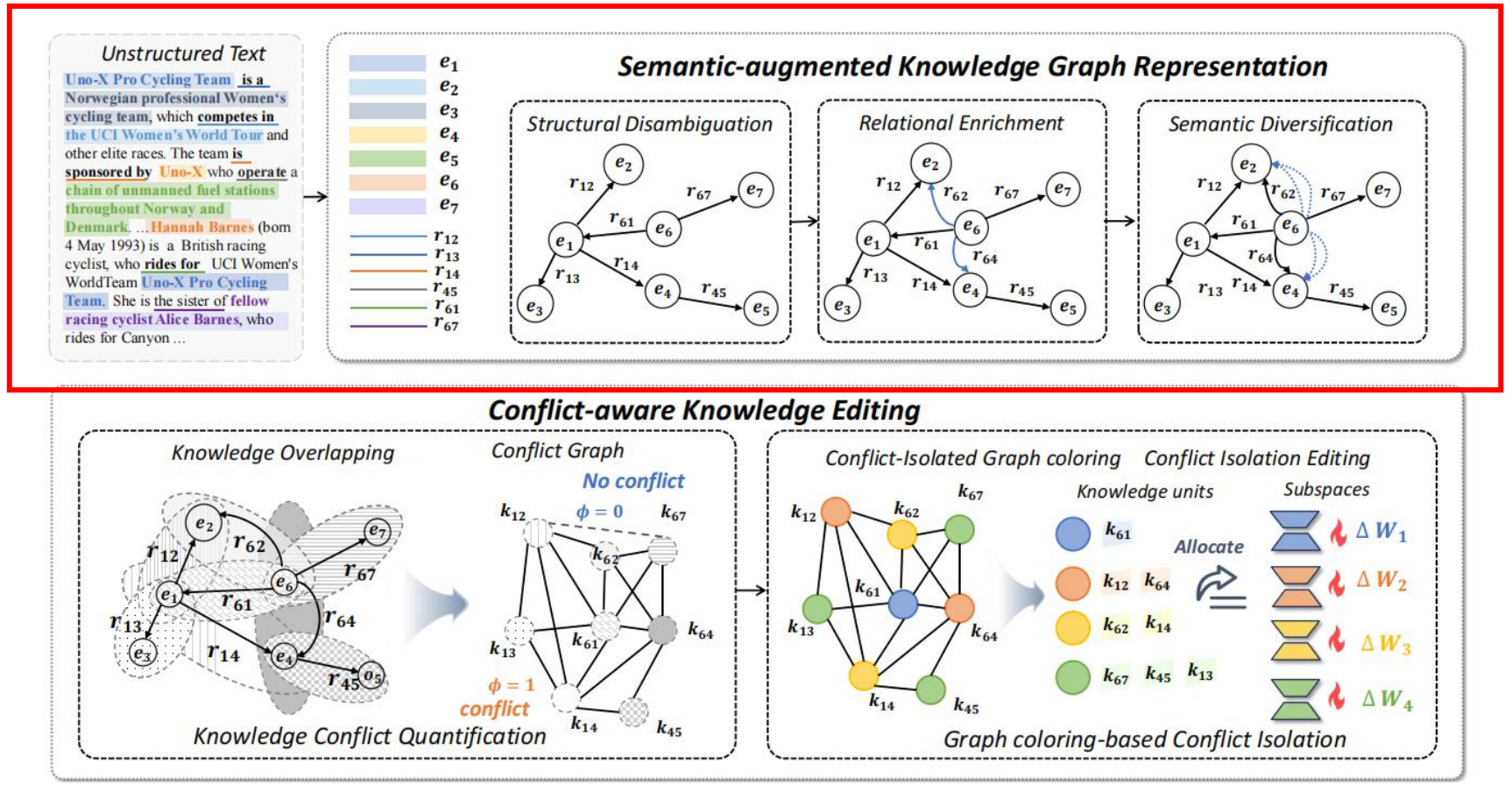
$$\mathbf{G} = \{\mathbf{E}, \mathbf{R}^{\mathrm{s}}\} = \mathcal{M}([\mathbf{I}^{\mathrm{s}}, \mathbf{T}])$$

➢ Relational enrichment

$$\mathbf{r}_{ij} = \mathcal{M}([\mathbf{I}^{\mathrm{h}}, \mathbf{G}, \mathbf{T}, (\mathbf{e}_i, \mathbf{e}_j)]), \text{where } d_{\mathbf{G}}(\mathbf{e}_i, \mathbf{e}_j) \in [2, p]$$
$$\mathbf{R}^{\mathrm{s}} \leftarrow \mathbf{R}^{\mathrm{s}} \cup \{\mathbf{r}_{ij}\}$$

➢ Semantic diversification

$$\{\mathbf{r}'_{ij}\} = \mathcal{M}([\mathrm{I}^{\mathrm{d}}, \mathbf{G}, \mathbf{T}, (\mathbf{e}_i, \mathbf{r}_{\mathrm{ij}}, \mathbf{e}_j)]), \text{where } \mathbf{r}_{\mathrm{ij}} \in \mathbf{R}^{\mathrm{s}},$$
$$\mathbf{R}^{\mathrm{d}} \leftarrow \mathbf{R}^{\mathrm{d}} \cup \{\mathbf{r}'_{ij}\}$$



CAKE Framework

- Conflict-aware Knowledge Editing strategy (CKE) introduces a graphtheoretic coloring mechanism that decouples semantically overlapping edits into orthogonal parameter subspaces to eliminate the editing conflicts

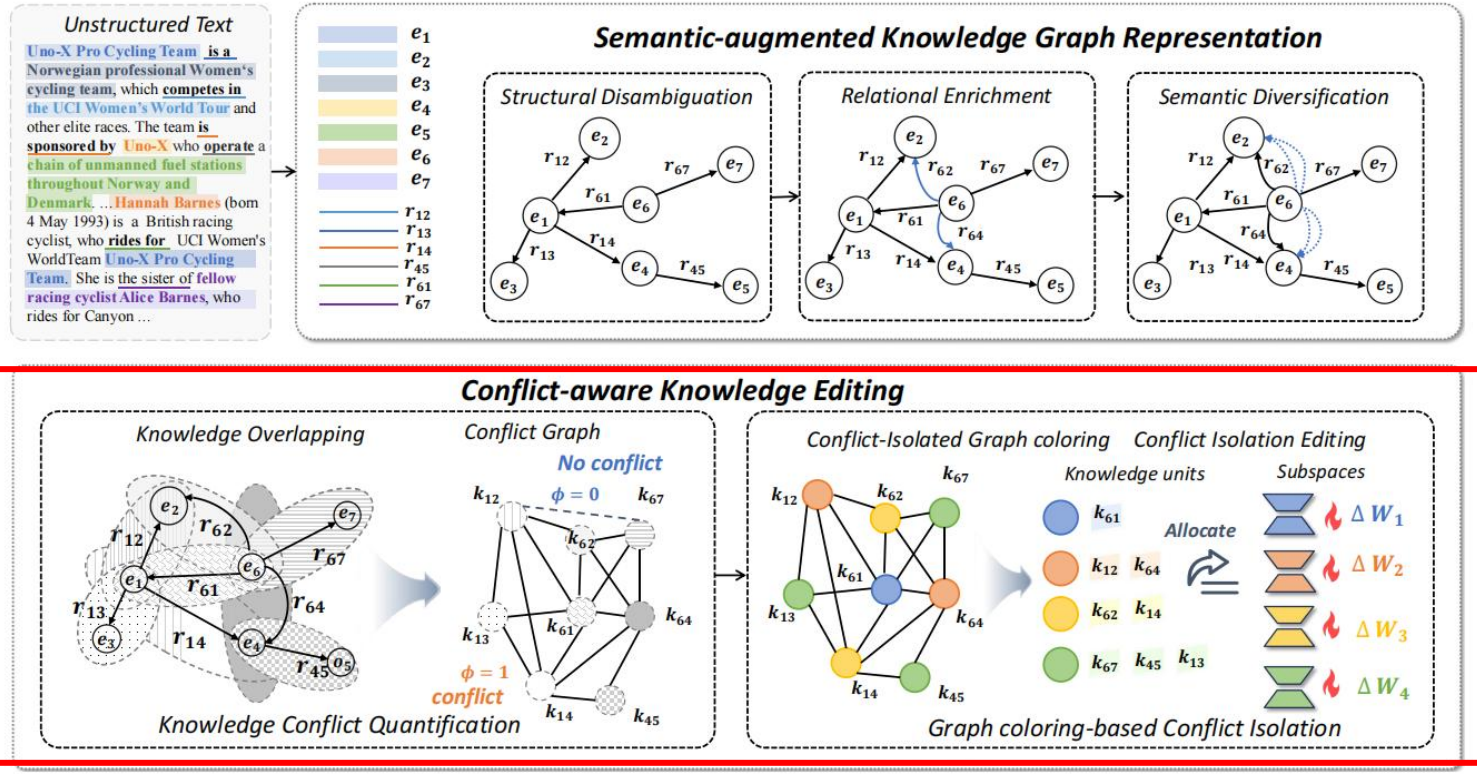
➢ **Conflict graph construction**

$$\mathbf{G}_c^* = (\mathbf{K}, \Phi)$$

$$\phi(\mathbf{k}_{ij}, \mathbf{k}_{pq}) = \begin{cases} 1 & \text{if } \alpha \cdot \mathbb{I}(\mathbf{e}_i, \mathbf{e}_p) + \beta \cdot \frac{|\mathbf{a}_{ij}^T \mathbf{a}_{pq}|}{\|\mathbf{a}_{ij}\| \cdot \|\mathbf{a}_{pq}\|} \geq \gamma \\ 0 & \text{otherwise} \end{cases}$$

➢ The **graphtheoretic coloring mechanism** assigns distinct colors to conflicting knowledge triples, with each color mapping to an isolated parameter subspace.

$$\Delta \mathbf{W}_l \leftarrow \Delta \mathbf{W}_l - \eta \cdot \nabla \sum_{f(\mathbf{k}_{ij}) = \mathbf{c}_l} \mathcal{L}\left(\mathcal{M}\left(\mathbf{x}_{ij}\right), \mathbf{e}_j\right)$$



Experiment

Comparison with State-of-the-art Methods

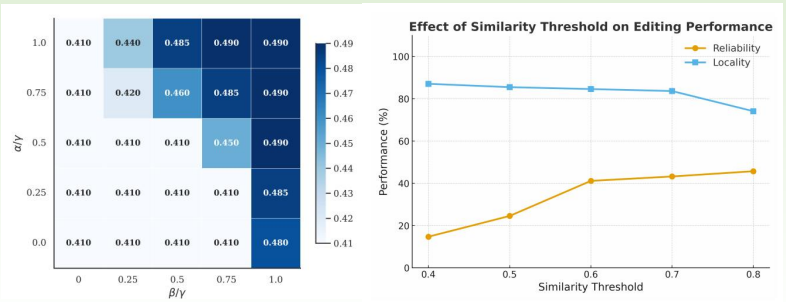
Model and Editing Method		Datasets								
Model	Editing Method	CounterFact			MQuAKE-CF			WikiUpdate		
		Rel.(%)	Loc.(%)	Avg(%)	Rel.(%)	Loc.(%)	Avg(%)	Rel.(%)	Loc.(%)	Avg(%)
GPT2-XL	ROME	7.84	61.21	34.52	34.84	59.86	47.35	30.13	61.11	45.62
	MELO	1.37	**64.10**	32.74	19.14	**63.15**	41.15	29.15	**63.24**	46.20
	MEMIT	5.69	59.04	32.37	33.90	60.28	47.09	34.56	62.29	48.43
	WISE	10.01	24.16	17.09	28.58	23.10	25.84	31.27	25.33	28.30
	Elder	23.10	60.67	41.89	38.28	57.33	47.81	25.31	62.20	43.76
	Ours	**35.04**	61.40	**48.22**	**62.01**	61.49	**61.75**	**42.26**	62.38	**51.80**
Llama3.2-3B	ROME	11.56	81.28	46.42	39.56	**82.74**	61.15	42.43	**80.68**	61.56
	MELO	4.61	73.30	38.96	32.73	80.27	56.50	45.87	70.40	58.14
	MEMIT	15.41	80.16	47.79	42.15	81.65	61.90	47.80	75.14	61.47
	WISE	4.46	14.84	9.65	19.61	11.31	15.46	4.09	17.65	10.87
	Elder	26.04	79.23	52.64	49.19	80.44	64.82	35.40	72.30	53.85
	Ours	**43.24**	**83.67**	**63.46**	**64.23**	80.53	**73.12**	**53.10**	71.42	**62.26**
Llama3-8b	ROME	11.64	87.32	49.48	43.54	87.44	65.49	53.11	87.87	70.49
	MELO	3.43	87.14	45.29	34.04	86.31	60.18	57.71	86.45	72.08
	MEMIT	22.76	**87.69**	55.23	50.86	**87.71**	69.29	56.54	**88.38**	**72.46**
	WISE	21.51	44.84	33.18	47.40	43.10	45.25	28.98	51.41	40.20
	Elder	1.99	87.27	44.63	56.13	81.50	68.82	**58.23**	84.35	71.29
	Ours	**38.19**	86.61	**62.40**	**57.74**	86.42	**72.08**	51.44	87.57	69.51

- CAKE demonstrates superior editing reliability on all three unstructured datasets and achieves comparable results with existing methods in terms of editing locality
- In particular, CAKE achieves a 15.43\% improvement in accuracy over the second-best method on editing llama3.

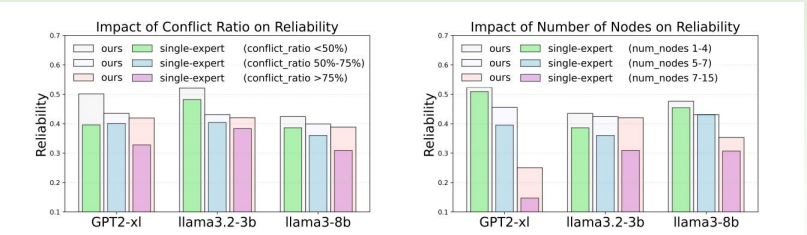
Ablation Study

Model	Ablation Method	Avg.	CounterFact	MQuAKE-CF	WikiUpdate
GPT2-XL	without SGR	45.94	34.63	**62.19**	41.22
	without CKE	44.13	30.06	61.51	40.84
	Router Expert	28.89	23.10	38.28	25.31
	Ours	**46.10**	**35.04**	62.01	**42.26**
Llama3.2-3B	without SGR	48.63	38.38	58.94	48.56
	without CKE	50.46	34.70	**65.71**	50.77
	Router Expert	36.87	26.04	49.19	35.40
	Ours	**53.52**	**43.24**	64.23	**53.10**
Llama3-8b	without SGR	47.62	37.28	55.53	50.06
	without CKE	47.59	25.66	56.13	**60.99**
	Router Expert	30.98	1.99	32.72	58.23
	Ours	**49.12**	**38.19**	**57.74**	51.44

Hyperparameter analysis



Impact of representation ambiguity/editing conflicts



TNANKS