Coarse-to-Fine Highlighting: Reducing Knowledge Hallucination in Large Language Models

Qitan Lv, Jie Wang, Hanzhu Chen,
Bin Li, Yongdong Zhang, Feng Wu

University of Science and Technology of China

Overlook

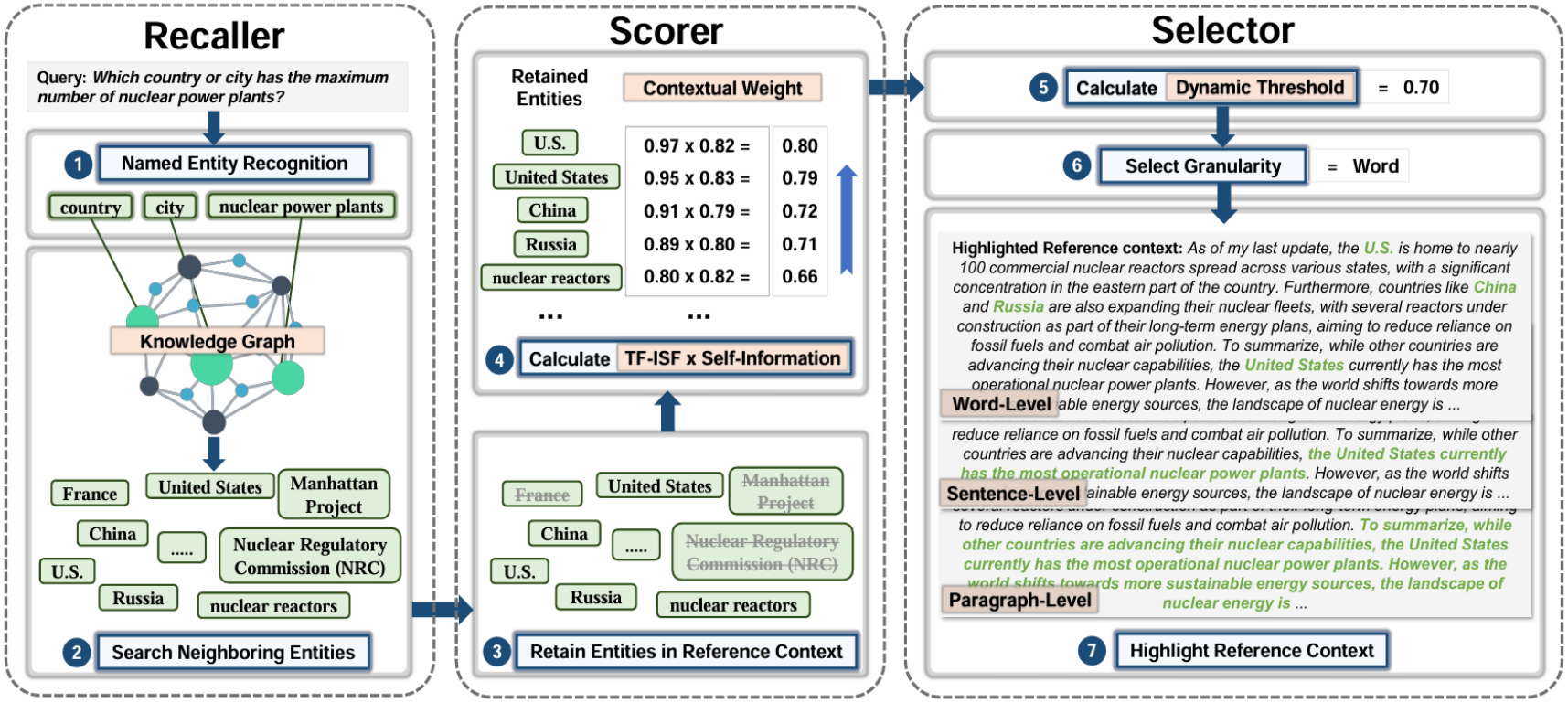
- Introduction and Motivation
- Methods
- Experiments

Introduction and Motivation

- Objective: Reducing Knowledge Hallucination in Large Language Models
- Challenge: LLMs may instead exacerbate hallucination when retrieving lengthy contexts
- Idea: To make LLMs focus on key content when retrieving the entire document.

Method

- We propose a COarse-to-Fine highlighTing method (COFT) that promotes LLMs to focus on key lexical units, which preserves complete contextual semantics and avoids getting lost in long contexts.



Method

Scorer

Contextual Weight

$$w(\mathbf{e}_k) = TF\text{-}ISF(\mathbf{e}_k) \times I(\mathbf{e}_k)$$

Self-information

$$I(\mathbf{t}_i) = -\log_2 P\left(\mathbf{t}_i \mid \mathbf{x}^{\text{que}}, \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{i-1}\right)$$

TF-ISF score

$$TF\text{-}ISF(\mathbf{e}_k) = \frac{f_{\mathbf{e}_k,\mathbf{s}_i}}{|\mathbf{s}_i|} \times \log_2\left(\frac{|\mathcal{S}|}{f_{\mathbf{e}_k,\mathcal{S}}+1}\right)$$

Algorithm 1 Pseudo code for entity-level iterative algorithm

Input: A query $\mathbf{x}^{\text{que}}$, a reference context $\mathbf{x}^{\text{refs}}$, a key candidate entity list $\mathcal{E}$, and a small language model $\mathcal{M}_s$.

1: Segment the reference context $\mathbf{x}^{\text{refs}}$ into sentences list $\mathcal{S} = [\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \ldots]$.
2: Initialize the TF-ISF dictionary $\mathcal{D}_{TF\text{-}ISF}$, the self-information dictionary $\mathcal{D}_{SI}$, and the contextual weight dictionary $\mathcal{D}_{CW}$.
3: **for** $\mathbf{e}_k$ **in** $\mathcal{E}$ **do**
4: Retain $\mathbf{e}_k$ occurred in each reference sentence $\mathbf{s}_i \in \mathcal{S}$.
5: Calculate the TF-ISF score of each entity via Equation 2 and append entities and corresponding TF-ISF scores into $\mathcal{D}_{TF\text{-}ISF}$.
6: Calculate the self-information score of each entity by the language model $\mathcal{M}_s$ via Equation 3 and append entities and self-information scores into $\mathcal{D}_{SI}$.
7: Calculate the contextual weights of each entity using $\mathcal{D}_{TF\text{-}ISF}$ and $\mathcal{D}_{SI}$ via Equation 4 and append all entities and their contextual weights into $\mathcal{D}_{CW}$.
8: **end for**

Output: Contextual weights dictionary $\mathcal{D}_{CW}$.

Experiments

- Effective

Backbone	Methods	WK			Sci/Tech			Wri/Rec		
		F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall
Vicuna-33B	Vanilla	34.5	27.8	45.5	25.8	17.4	50.2	27.1	16.4	78.7
	CoT	32.3	27.4	39.5	20.4	12.7	52.9	26.5	17.0	60.3
	RALM	48.7	45.7	52.1	34.2	24.7	55.8	27.1	16.2	82.8
	CoVe	47.3	47.6	47.1	47.2	39.8	58.2	64.0	66.7	61.5
	CoN	55.9	55.7	56.1	59.3	58.1	60.6	62.4	55.3	71.5
	COFT_p	**69.3**	**71.9**	66.9	67.9	62.9	73.8	70.4	66.8	74.4
	COFT_s	62.0	63.1	60.9	68.7	**67.1**	70.4	66.2	64.7	67.7
	COFT_w	64.4	61.7	**67.4**	**70.9**	65.7	**77.2**	**77.3**	**67.9**	**89.8**

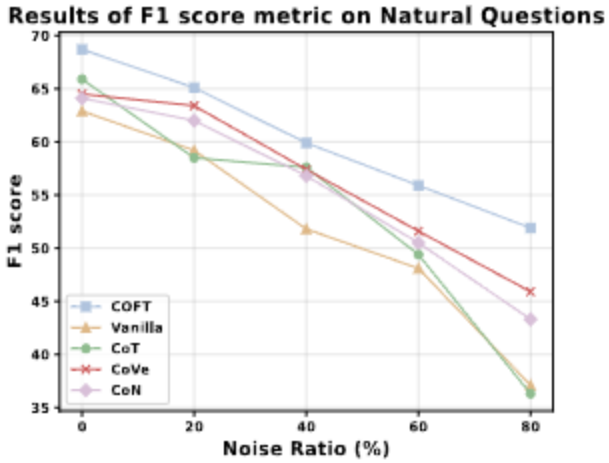
Backbone	Methods	WK			Sci/Tech			Wri/Rec		
		F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall
ChatGPT	Vanilla	9.1	27.6	5.4	4.1	6.5	2.9	0.7	4.2	0.4
	CoT	2.6	33.3	1.4	4.2	25.1	2.3	2.7	9.1	1.6
	RALM	25.2	34.9	19.7	17.4	16.7	18.2	20.1	54.1	12.4
	CoVe	20.0	50.1	12.5	18.2	12.5	33.3	23.1	63.6	14.1
	CoN	18.2	66.7	10.6	20.0	25.0	16.7	31.4	32.7	30.3
	COFT_p	78.6	83.8	74.0	83.9	**81.2**	86.8	77.5	85.9	70.5
	COFT_s	76.8	75.7	77.9	74.6	79.1	70.5	76.8	84.4	70.5
	COFT_w	**81.6**	**85.5**	**77.9**	**84.4**	80.9	**88.4**	**81.1**	**93.7**	**71.5**

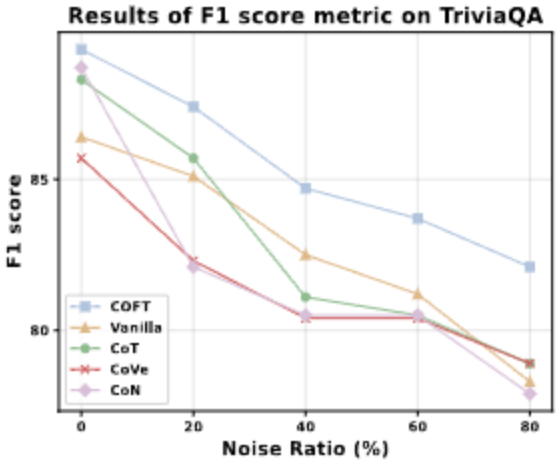
Backbone	Methods	WK			Sci/Tech			Wri/Rec		
		F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall
GPT4	Vanilla	40.2	76.9	27.2	19.7	60.0	11.8	22.3	89.5	12.7
	CoT	50.2	79.4	36.7	25.2	64.0	15.7	26.2	89.1	15.4
	RALM	53.6	80.8	40.1	34.7	59.5	24.5	52.2	63.8	44.2
	CoVe	49.7	55.4	45.1	66.7	83.3	55.6	48.2	56.9	41.8
	CoN	52.8	45.2	63.6	66.7	75.0	60.0	68.8	78.6	61.1
	COFT_p	83.1	79.7	**86.8**	**89.9**	84.4	**96.1**	**91.8**	85.5	**99.1**
	COFT_s	80.0	92.3	70.6	76.6	84.9	69.8	85.5	89.2	82.1
	COFT_w	**87.3**	**94.8**	80.9	77.9	**86.0**	71.3	84.7	**92.9**	77.9

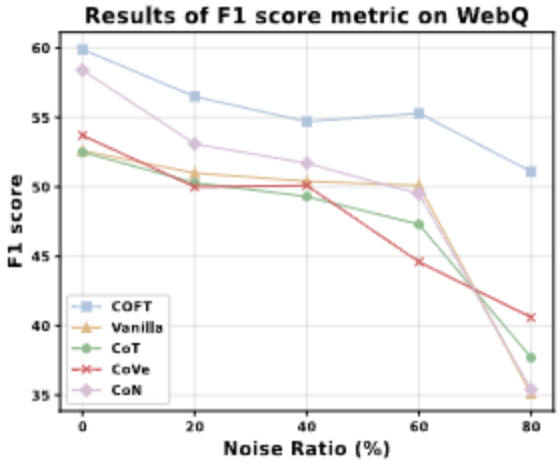
Experiments

- Generalizable

Open-domain Question-Answering





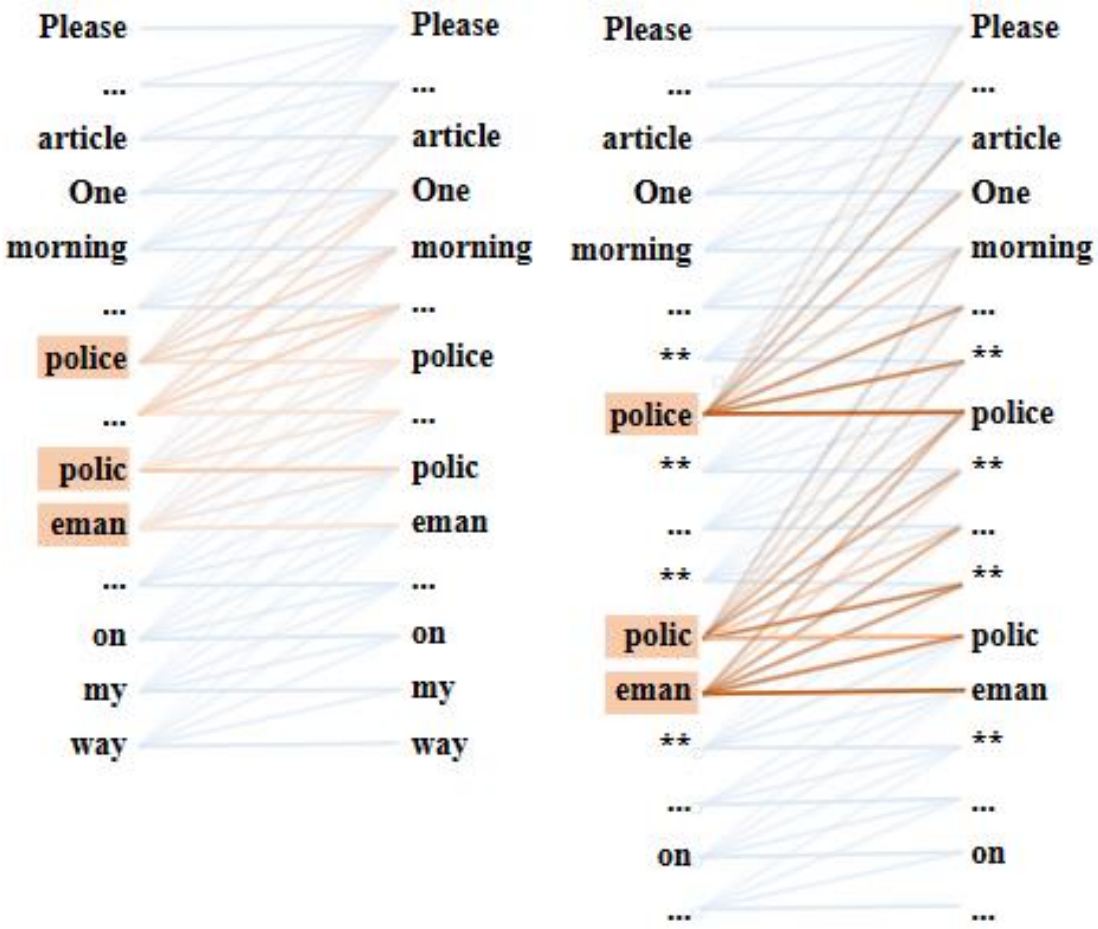


Reading Comprehension

Backbone	Methods	RACE-H	RACE-M
ChatGPT	Vanilla	65.6	81.6
	CoT	56.3	81.6
	CoVe	54.5	82.1
	CoN	59.4	79.6
	COFT	**73.4**	**85.8**

Experiments

- Visualization study



Thank You!