AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Explore What LLM Does Not Know in Complex Question Answering

Xin Lin, Zhenya Huang, Zhiqiang Zhang, Jun Zhou, Enhong Chen
Speaker: Zhiyuan Ma
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence

Outline

- Background
- KEQA Framework
- Experiments
- Conclusion

Background

- Complex question answering
  - Answer questions based on related knowledge
  - Requirements of QA system
    - Master multiple knowledge
    - Perform complex reasoning over knowledge
- One promising solution
  - Large language model (LLM)
  - Retrieval-augmented generation (RAG)
    - First retrieve related knowledge and then perform reliable reasoning and generation

Question: What is the birth date of the person Richard Callaghan coached to Olympic, world, and national titles?

Required Knowledge

K_1: Who did Richard Callaghan coach to Olympic, world, and national titles?
LLM: Tara Lipinski

K_2: What is the birth date of Tara Lipinski?
RAG: June 10, 1982

Answer: The person Richard Callaghan coached to Olympic, world, and national titles is Tara Lipinski. She was born in June 10, 1982. So the answer is June 10, 1982.

Motivations

- How to effectively facilitate RAG in complex QA
  - Examine the knowledge boundary of LLM
    - What the LLM does not know
    - Only missing knowledge needs to be supplemented from external
  - Evaluate the utility of external knowledge
    - How helpful the external knowledge is in QA
    - Related knowledge may be helpless and even mislead reasoning in QA

Question: What is the birth date of the person Richard Callaghan coached to Olympic, world, and national titles?

Required Knowledge

K_1: Who did Richard Callaghan coach to Olympic, world, and national titles?
LLM: Tara Lipinski (**Known**)

K_2: What is the birth date of Tara Lipinski?
LLM: June 1977 (**Unknown**)

P_1: *"... Tara Kristen Lipinski (born June 10, 1982) ... "* ***(Helpful)***
P_2: *"... Lipinski appeared on "The Today Show" on March 18, 2011 ... "* ***(Helpless)***
P_3: *"... Lipinski was coached by Jeff DiGregorio ... "* ***(Helpless)***

RAG: June 10, 1982

Answer: The person Richard Callaghan coached to Olympic, world, and national titles is Tara Lipinski. She was born in June 10, 1982. So the answer is June 10, 1982.

Challenges

- How to precisely examine the knowledge boundary of the LLM
  - LLM contains tremendous uninterpretable parameters
  - Self-evaluation with LLM: Tend to be over-confident on their knowledge states
  - Probability-based methods: Focus on uncertain tokens rather than missing knowledge

- How to identify utility of external knowledge in complex QA
  - More than content or semantic relevancy between knowledge and question
  - Reasoning logic: Whether knowledge is necessary in one reasoning step
  - LLM ability: Whether LLM masters the knowledge, whether LLM is affected by the knowledge

Outline

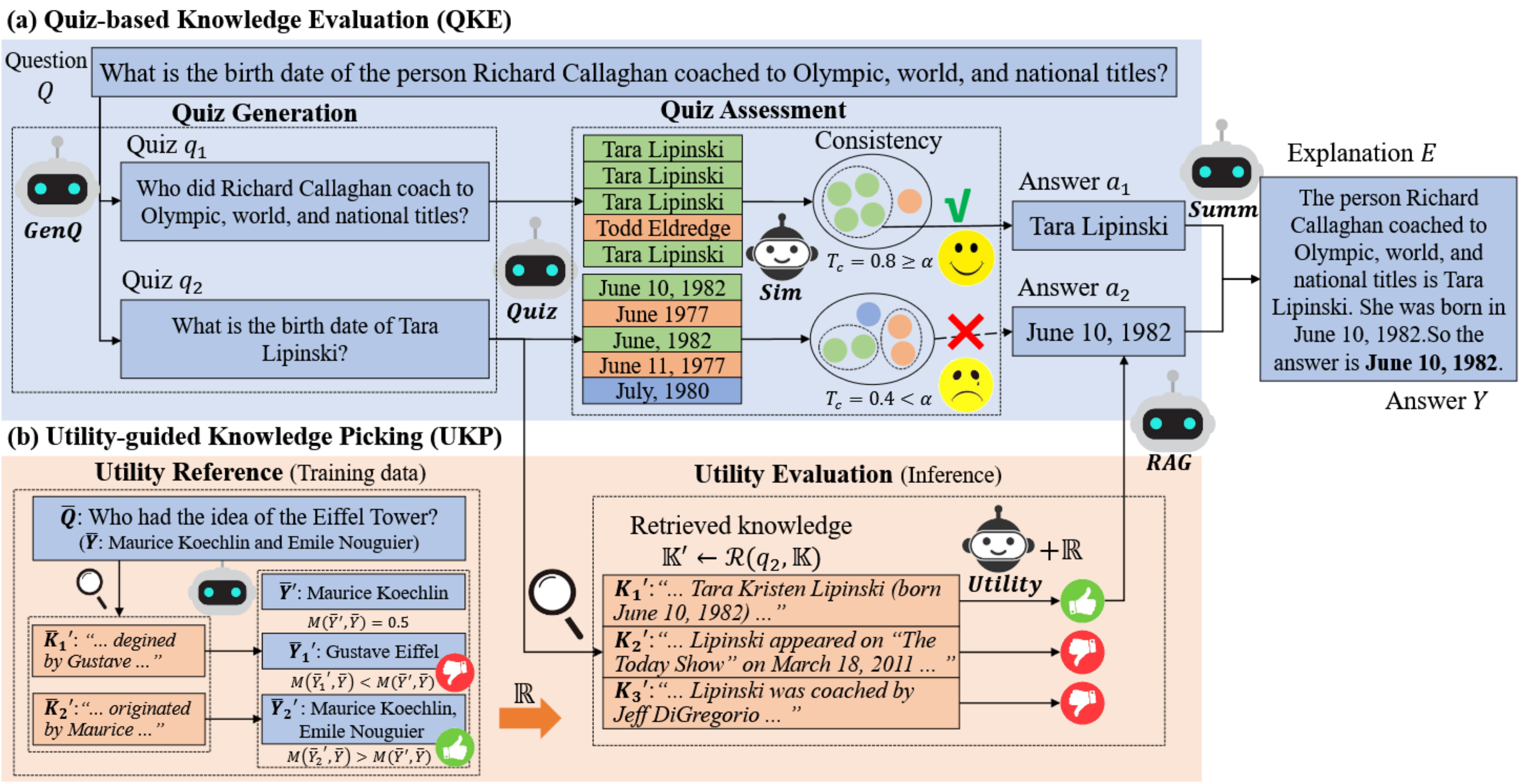
Problem Definition

- Complex question answering
  - Q: Question in natural language
  - Y: Answer inferred with an explanation E in natural language
  - $\mathbb{K} = \{K_1, \ldots, K_n\}$: External knowledge source, each K_i is a passage in corpus (May also be a knowledge triple in KG or a webpage online depending on $\mathbb{K}$)

- Goal
  - Given knowledge source $\mathbb{K}$ and question Q
  - Retrieve necessary knowledge $\mathbb{K}^* = \{K_1^*, \ldots, K_m^*\}$ from $\mathbb{K}$ with a retriever $\mathcal{R}$, and generate one explanation E with a LLM $\mathcal{L}$ to infer the answer Y to Q

KEQA Framework

- Question Answer with Knowledge Evaluation (KEQA)



KEQA Framework

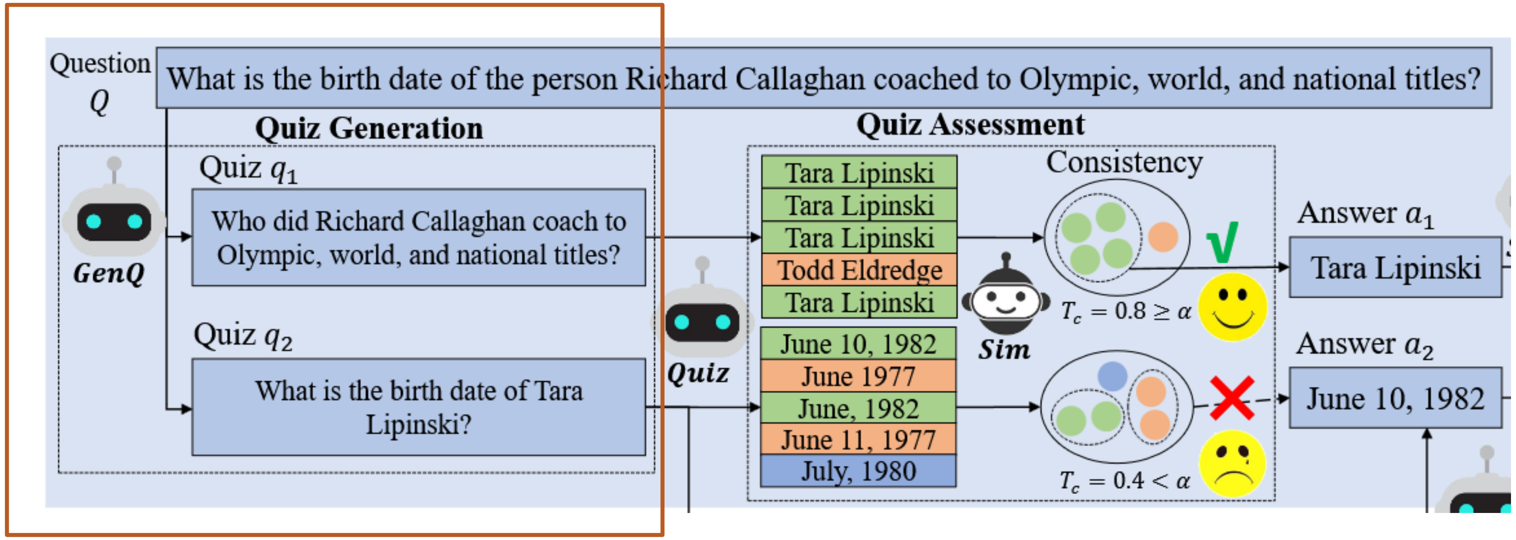
- Question Answer with Knowledge Evaluation (KEQA)
  - Knowledge boundary: Quiz-based Knowledge Evaluation
    - Output-oriented: Whether LLM could answer knowledge-related quizzes
    - Retrieval-on-demand: Only retrieve missing knowledge that LLM fails the quiz
  - Knowledge utility: Utility-guided Knowledge Picking
    - Result-oriented: Whether external knowledge helps LLM improve QA accuracy
    - Knowledge picking: Only pick helpful knowledge with positive utility

KEQA Framework

- Quiz-based Knowledge Evaluation
  - Whether LLM masters knowledge: Examine whether LLM could answer related question
  - Quiz Generation
    - Hard to detect missing knowledge: Complex QA involves multiple knowledge and causes of failure
    - Solution: Generate simple quiz related to single knowledge by decomposing the complex question

Quiz generation by question decomposition

$$\{q_1, \ldots, q_s\} \leftarrow GenQ(Q, \mathcal{L})$$



KEQA Framework

- Quiz-based Knowledge Evaluation
  - Whether LLM masters knowledge: Examine whether LLM could answer related question
  - Quiz Assessment
    - Hard to assess LLM's answer: No ground truth for the quiz
    - Consistency-based assessment: Whether LLM gives consistent answer based on knowledge or randomly guess in multiple tries
    - Answer semantic discrimination: Consistent answers may be different in words for open quiz

Quiz answer discrimination

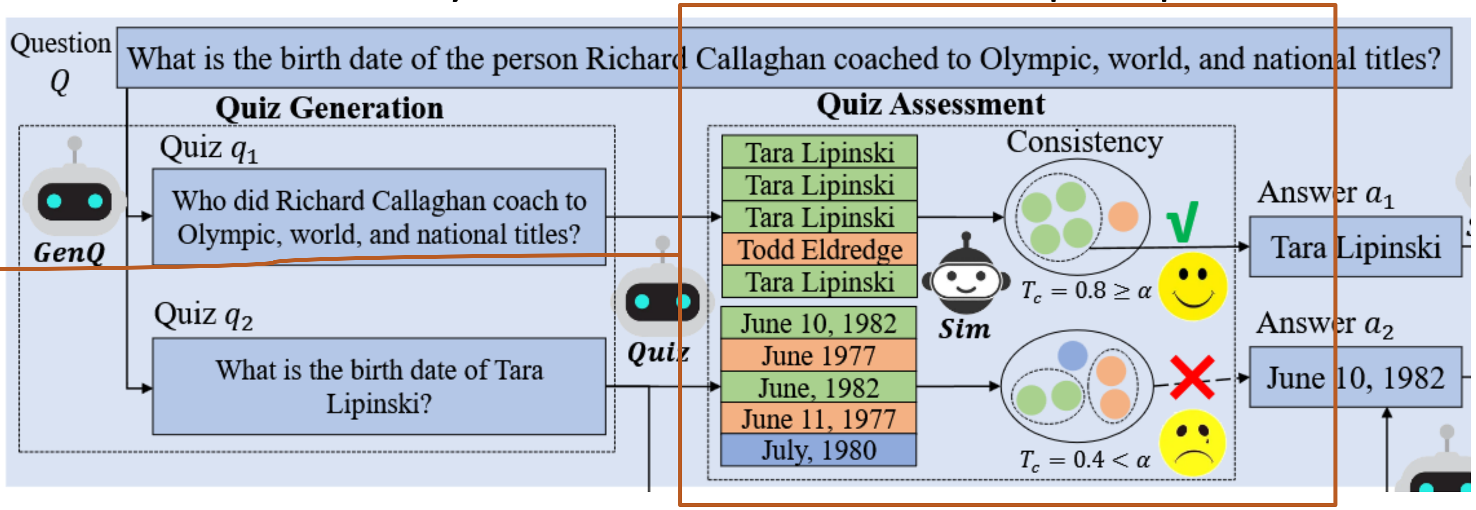
$$a_i \leftarrow Quiz(q, \mathcal{L})$$

$$same \leftarrow Sim(q, a_i, a_j, \mathcal{D}_s)$$

Consistency assessment

$$T_c \geq \alpha$$

T_c: proportion of consistent answer



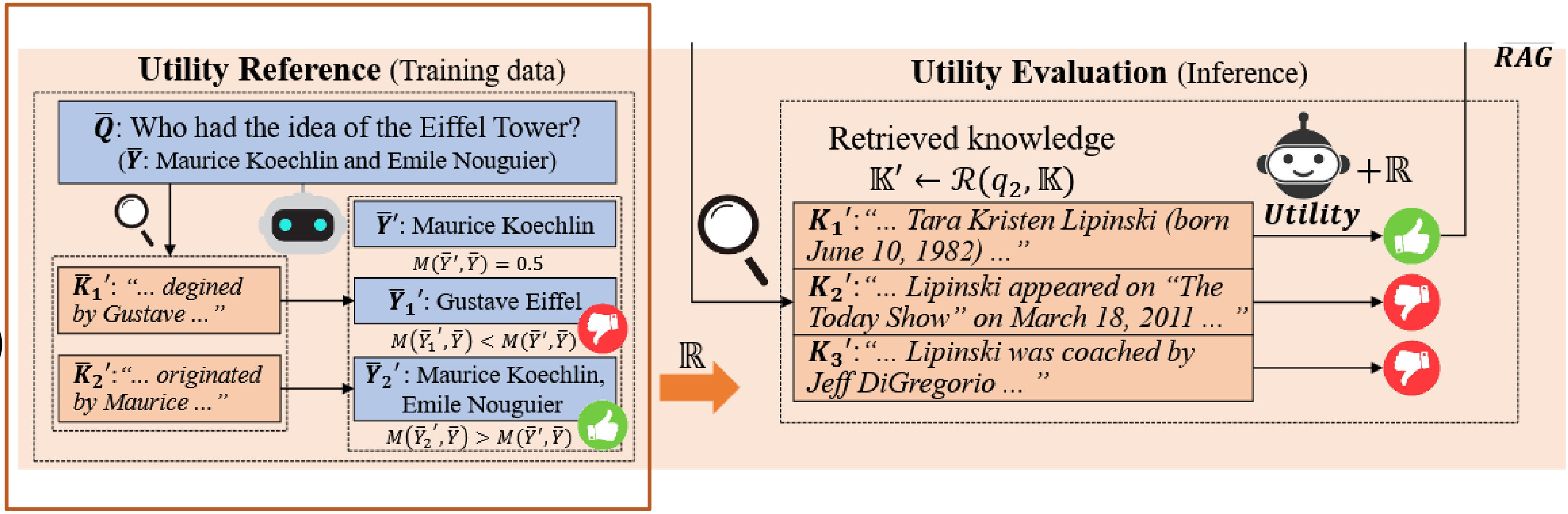
KEQA Framework

- Utility-guided Knowledge Picking
  - Missing knowledge that LLM fails in quiz: Retrieve from external $\mathbb{K}' \leftarrow \mathcal{R}(q, \mathbb{K})$
  - Another problem: Whether external knowledge helps in QA reasoning
  - Utility Reference
    - Knowledge utility: Whether adding knowledge increases accuracy of LLM's answer
    - Compute utility label with ground truth answer in training data as reference

Utility label computation in training data

$$U(\bar{Q}, \bar{K}') \leftarrow \begin{cases} 1, Cor(\bar{Q}, \bar{\mathbb{K}}^* \cup \{\bar{K}'\}) > Cor(\bar{Q}, \bar{\mathbb{K}}^*) \\ 0, Cor(\bar{Q}, \bar{\mathbb{K}}^* \cup \{\bar{K}'\}) = Cor(\bar{Q}, \bar{\mathbb{K}}^*) \\ -1, Cor(\bar{Q}, \bar{\mathbb{K}}^* \cup \{\bar{K}'\}) < Cor(\bar{Q}, \bar{\mathbb{K}}^*) \end{cases}$$

$$Cor(\bar{Q}, \bar{\mathbb{K}}^*) = M(\bar{Y}', \bar{Y}) = M(Ans(\bar{Q}, \bar{\mathbb{K}}^*, \mathcal{L}), \bar{Y})$$



KEQA Framework

Utility-guided Knowledge Picking

- Missing knowledge that LLM fails in quiz: Retrieve from external $\mathbb{K}' \leftarrow \mathcal{R}(q, \mathbb{K})$
- Another problem: Whether external knowledge helps in QA reasoning
- Utility Evaluation
  - Inference: Utility evaluation with smaller LLM and similar in-context examples from references
  - Only pick external knowledge with positive utility for RAG

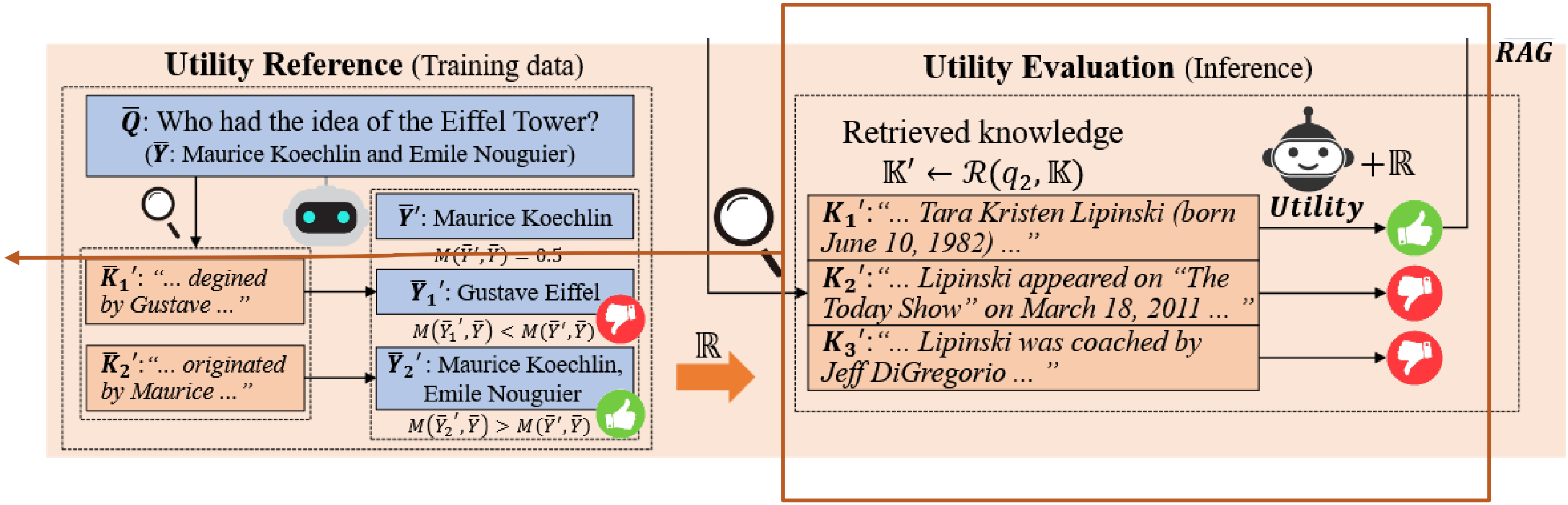
In-context example retrieval

$$\mathbb{R}' \leftarrow \mathcal{R}_u(q, K', \mathbb{R})$$

Utility evaluation in inference

$$Util(q, K') \leftarrow Utility(q, K', \mathbb{R}', \mathcal{D}_u)$$

$$\mathbb{K}^* = \{K' \in \mathbb{K}' | Util(q, K') = 1\}$$



KEQA Framework

- Overall process of KEQA inference

Quiz-based Knowledge Evaluation

Utility-guided Knowledge Picking

RAG and Answer Summarization

Algorithm 2: KEQA inference

Input: Question Q, LLM $\mathcal{L}$, knowledge source $\mathbb{K}$, utility reference $\mathbb{R}$, knowledge retriever $\mathcal{R}$, reference retriever $\mathcal{R}_u$, semantic discriminator $\mathcal{D}_s$, utility discriminator $\mathcal{D}_u$
Parameter: Number of tries N_c, consistency threshold α
Output: Explanation E, answer Y

$QS \leftarrow GenQ(Q, \mathcal{L})$
$QAS \leftarrow \emptyset$
for $q \in QS$ **do**
    $A_c, T_c \leftarrow Consist\_assess(q, \mathcal{L}, \mathcal{D}_s, N_c)$
    if $T_c \geq \alpha$ **then**
        $QAS \leftarrow QAS \cup \{(q, A_c)\}$
    else
        $\mathbb{K}' \leftarrow \mathcal{R}(q, \mathbb{K})$
        $\mathbb{K}^* \leftarrow \emptyset$
        for $K' \in \mathbb{K}'$ **do**
            $\mathbb{R}' \leftarrow \mathcal{R}_u(q, K', \mathbb{R})$
            if $Util(q, K') == 1$ **then**
                $\mathbb{K}^* \leftarrow \mathbb{K}^* \cup \{K'\}$
            end if
        end for
        $a \leftarrow RAG(q, \mathbb{K}^*, \mathcal{L})$
        $QAS \leftarrow QAS \cup \{(q, a)\}$
    end if
end for
$E, Y \leftarrow Summ(Q, QAS, \mathcal{L})$
return E, Y

Outline

- Background
- KEQA Framework
- **Experiments**
- Conclusion

Experiments

Experimental setup

- Datasets
  - One-hop QA: NaturalQuestions (NQ)
  - Complex QA: StrategyQA, HotpotQA, 2WikiMultihopQA (2WMQA)
- Baselines
  - Non-RAG: Vanilla GPT-3.5, Zero-shot CoT, Few-shot CoT
  - RAG: Vanilla RAG, ReAct, IRCoT, FLARE, Self-Rag, SearChain, Rowen, SlimPLM
- RAG setup
  - LLM backbone $\mathcal{L}$: GPT-3.5-turbo
  - Retriever $\mathcal{R}$: BM25 (Elasticsearch)
  - Knowledge source $\mathbb{K}$: Wikipedia dump Dec 20, 2018

Experiments

□ Overall results

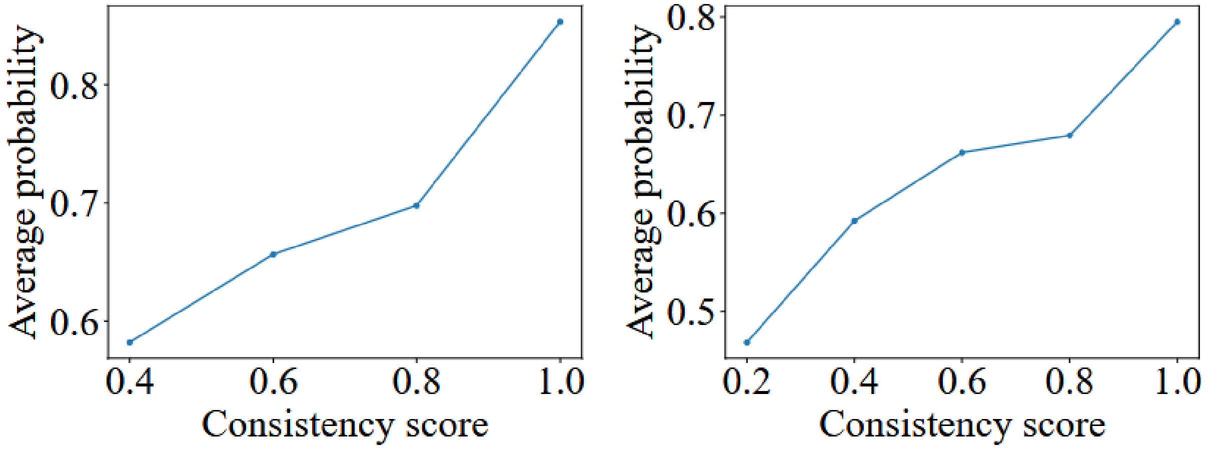
Dataset / Metric	NQ F1	NQ EM	StrategyQA ACC	HotpotQA F1	HotpotQA EM	2WMQA F1	2WMQA EM
Vanilla GPT-3.5	0.427	0.294	0.468	0.380	0.264	0.313	0.224
Zero-shot CoT	0.454	0.296	0.510	0.353	0.260	0.320	0.218
Few-shot CoT	0.445	0.292	0.620	0.373	0.254	0.360	0.224
Vanilla RAG	0.385	0.258	0.516	0.387	0.254	0.314	0.244
ReAct	0.335	0.212	0.554	0.390	0.270	0.305	0.204
IRCoT	0.344	0.216	0.622	0.361	0.232	0.318	0.202
FLARE	0.455	0.318	0.662	0.391	0.268	0.364	0.246
Self-Rag	0.387	0.270	0.632	0.357	0.220	0.311	0.210
SearChain	0.337	0.214	0.616	0.349	0.216	0.313	0.222
Rowen	0.452	0.286	0.666	0.382	0.240	0.307	0.212
SlimPLM	0.442	0.280	0.566	0.393	0.266	0.368	0.242
KEQA	**0.483***	**0.352***	**0.680***	**0.400***	**0.278***	**0.405***	**0.326***
KEQA *w/o* QKE	0.409	0.284	0.644	0.352	0.232	0.396	0.258
KEQA *w/o* UKP	0.453	0.302	0.678	0.350	0.250	0.398	0.314
KEQA *w/o* $\mathbb{R}$	0.456	0.316	0.676	0.356	0.252	0.398	0.302
KEQA *w* random $\mathbb{R}'$	0.474	0.324	0.666	0.375	0.262	0.385	0.288
KEQA *w* SE	0.475	0.342	0.678	0.388	0.272	0.397	0.316

- KEQA outperforms all baselines, demonstrating its effectiveness
- RAG methods do not always outperform non-RAG methods especially on simple tasks due to noises
- Adaptive RAG methods perform better, showing the superiority of retrieval-on-demand

Experiments

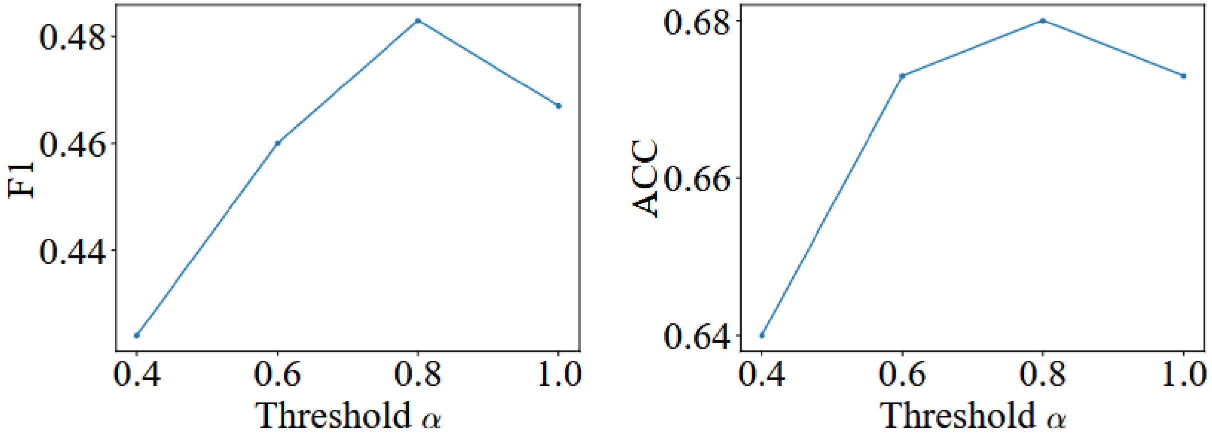
Quiz analysis

- The consistency is correlated positively with the average generation probability, showing its rationality in detecting knowledge state



(a) NQ (b) StrategyQA

- Lower consistency threshold might mistakenly treat random guess as knowledge
- Too high consistency threshold might conduct unnecessary retrieval on known information



(a) NQ (b) StrategyQA

Experiments

- Generalizability analysis
  - KEQA could promote RAG on LLMs and retrievers with different abilities, showing its generalizability

Dataset	NQ	StrategyQA
KEQA *w* GPT-4	0.515	0.760
RAG *w* GPT-4	0.496	0.733
KEQA *w* LLaMA	0.315	0.486
RAG *w* LLaMA	0.295	0.446

Dataset	NQ	StrategyQA
KEQA *w* DPR	0.491	0.690
RAG *w* DPR	0.486	0.662

- Efficiency analysis
  - KEQA conducts much less retrieval, and consumes less tokens than baselines
  - KEQA has higher latency, but it can be greatly reduced if optimized in parallel

Dataset	NQ				StrategyQA			
Cost	$\mathcal{R}$	$\mathcal{L}$	token	time	$\mathcal{R}$	$\mathcal{L}$	token	time
KEQA	0.138	5.08	239	6.28	0.212	16.44	1016	17.21
IRCoT	1.46	1.46	2024	2.67	1.88	1.88	2374	3.15
FLARE	0.81	1.81	1367	2.45	1.22	2.22	1742	3.45

Experiments

AAAI-25 / IAAI-25 / EAAI-25

- Case study
  - KEQA could detect what the LLM knows and what the LLM does not know
  - KEQA could refer to external knowledge only when necessary, to promote both accuracy and efficiency

Question	Which film has the director who was born earlier, The Assassination Of Trotsky or My Life Is Hell?
KEQA	The director of The Assassination Of Trotsky, Joseph Losey, was born on January 14, 1909. The director of My Life Is Hell, *Josiane Balasko*, was born on April 15, 1950. Joseph Losey was born earlier than Josiane Balasko. So the answer is The Assassination Of Trotsky.

q_1: Who is the director of The Assassination Of Trotsky?
a_1: Joseph Losey
q_2: When was Joseph Losey born?
a_2: January 14, 1909
q_3: Who is the director of My Life Is Hell?
a_3: *Josiane Balasko*
Retrieval: My Life Is Hell is a 1991 French comedy film directed by Josiane Balasko ...
q_4: When was the director of Josiane Balasko born?
a_4: April 15, 1950
q_5: Is January 14, 1909 earlier than April 15, 1950?
a_5: Yes

Outline

- Background
- KEQA Framework
- Experiments
- **Conclusion**

Conclusion

- Summary
  - KEQA framework to improve RAG in complex QA
  - Quiz-based knowledge evaluation to examine knowledge boundary of LLM
  - Utility-guided knowledge picking to evaluate helpfulness of external knowledge in QA
- Future work
  - How to detect outdated internal knowledge in rapidly evolving world
  - How to adapt to tasks without clear reasoning logic such as writing

Thanks for Listening!

linx@mail.ustc.edu.cn

https://github.com/l-xin/KEQA