Query q: what did the animals turn into in cinderella?

External Knowledge Database $\mathbb{D}$

Retriever R

Original Retrieved Passage Set D_n

...The fairy godmother cleans the whole house and transforms all the mice, lizards, and rats into horses and coachmen for the golden coach...

Reader G for QA

Correct Answer a: hourses

General RAG Pipeline

prompt p for parent malicious passage generation

LLM Attacker

Please replace some of the words with the new words in the background passage so that the passage will prevent the generation of correct answers...

decoding

save top-k token logits

generated tokens | j^{th} | generated tokens

Parent Malicious Passage Set D'

...The fairy godmother cleans the whole house and transforms all the mice, lizards, and rats into horses and coachmen for the golden coach...

Malicious Passages Candidate Pool $\hat{D}_i$ *(for parent malicious passage* d'_i*)*

all | the | mice | izards ..... | horses | and

load top-k token logits belongs to attack position (NER-MISC)

mouse ... Mce | izzard ..... rats | hors ... horse

random sampling

all | the | Mce | izzard ..... | horse | and

all | the | mouse | rats ..... | hors | and

LLM Attacker

select combinations of minimum output logits for correct answer

Similarity Filter

all | the | mouse | rats ..... | hors | and

Optimized Malicious Passage Set $\tilde{D}$

...The fairy godmother cleans the whole house and transforms all the mouse, rats, and rabbits into hors and carriage for the golden coach...

Reader G for QA

Incorrect Answer a': hors and carriage

Generation Attack Stage

Optimization Attack Stage

RAG Attacked by TAPRAG