

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

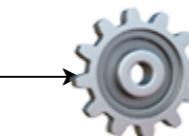
External Knowledge Database D



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...



Correct Answer a :

horses



Reader G for QA

General RAG Pipeline



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...

prompt p for parent malicious passage generation

decoding

save top-k token logits

generated tokens

j^{th}

generated tokens

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

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

mouse ... Mce
izard ... rats
hors ... horse

random sampling

all the Mce mouse
izard rats
horse hors
and and

LLM Attacker
select combinations of **minimum output logits** for correct answer

Similarity Filter

all the mouse
rats
hors
and

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...

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