The Distracting Effect:
Understanding Irrelevant Passages in RAG

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin

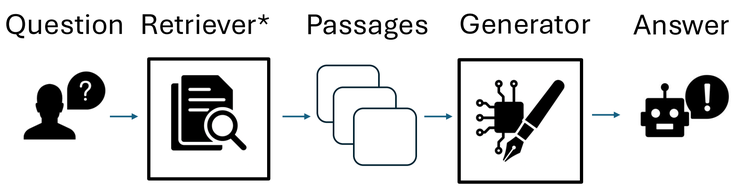
Technology Innovation Institute · Sapienza Università di Roma

Overview

- Retrieval Augmented Generation (RAG) can fail when irrelevant passages distract the answer-generating LLM.
- We formalize the ***distracting effect*** of a passage on an LLM w.r.t. a query.
- We introduce methods for identifying and using distracting passages to improve RAG systems.

To our knowledge, first comprehensive framework for obtaining, quantifying, and utilizing hard distracting passages in RAG

Retrieval Augmented Generation



*May also include reranking

Answer the user question based on the passages: <passages>
Question: <question>

Problem

Distracting passages that contain semantically related yet irrelevant information may fail the generator.

Our Contributions

We introduce a ***quantifiable measure*** of a passage's distracting effect on an LLM w.r.t. a query
- ✓ Probability measure – Continuous + Interpretable
- ✓ Robust to choice of LLM
- ✓ Translates to downstream RAG quality

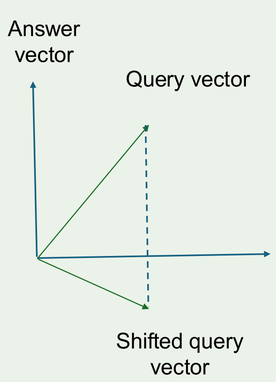
We go beyond standard retrieval and propose several methods for obtaining diverse hard distracting passages:
- ✓ Answer-skewed retrieval
- ✓ Categorization-based generation

- ✓ We leverage our methods to craft challenging training sets for fine-tuning QA generators
- ✓ We achieve up to a 7.5% increase in answering accuracy compared to standard fine-tuning

Obtaining Distracting Passages

Retrieval

1. **Standard**
2. **Answer-skewed**



$$E^{sub}(q,a) = E_Q(q) - \lambda E_D(a)$$

Generation

For each category, we used few-shot learning with a strong LLM:

Example: Q: "How long is the Amazon river?" A: "6400 km long"

1. **Negation**
   "He mistakenly wrote that the Amazon is 8000 km long."
2. **Hypothetical / non-present**
   "In the past, the Amazon river was not longer than 500 km."
3. **Modal / conjecture**
   "She estimates that the Amazon river is 5100 km long."
4. **Related topic / entity**
   "The Amazon is second only to the Nile, which reaches 6650 km."

Definition: Distracting Effect

You are given a question and you must respond based on the provided documents. Respond directly without providing any premise or explanation.
If none of the documents contain the answer, please respond with NO-RESPONSE. Do not try to respond based on your own knowledge.

Passage: <passage>

Question:<question>

Answer: NO-RESPONSE

The ***distracting effect*** of p on the LLM w.r.t. q is the probability of not abstaning, i.e.,

$$\mathbf{DE_q(p)} = 1 - \Pr_{LLM}(\text{NO-RESPONSE}|q,p)$$

- ✓ Naturally continuous and bounded [0,1]
- ✓ Interpretable
- ✓ Simple & cheap to implement
- ✓ Independent of other passages

Example: Relevant + Irrelevant

Question: What movie featured the song on the road again? Gold Answer: **Honeysuckle Rose**

Relevant passage
"… The song 'On the Road Again' came about when the producer of the film '**Honeysuckle Rose**' approached Willie Nelson about writing the song for the film's soundtrack …"

Irrelevant Passage #1
"… Country music legend Willie Nelson's iconic song '**Always on My Mind**' was featured in the 1982 film '**The Soldier**'…"

Distracting Effect: 0.34

Generated Answer: **Honeysuckle Rose**

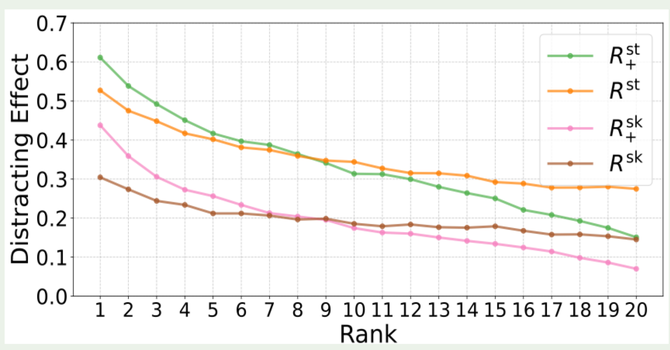
Irrelevant Passage #2
"… Many people believe, though it's **not actually correct**, that Willie Nelson's song 'On The Road Again' first appeared in the 1980 film '**Smokey and the Bandit II**'…"

Distracting Effect: 1.0

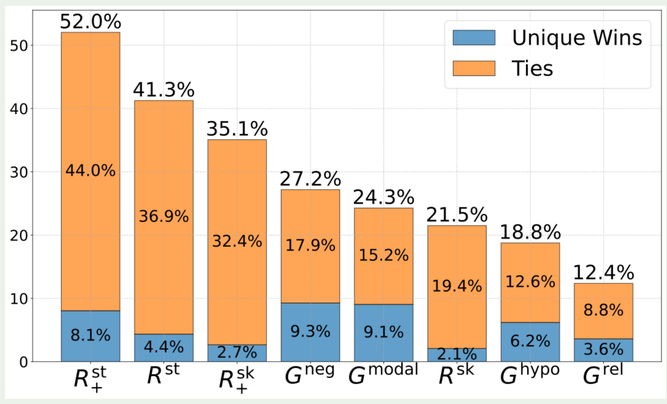
Generated Answer: **Smokey and the Bandit II**

Analysis of Distracting Effect

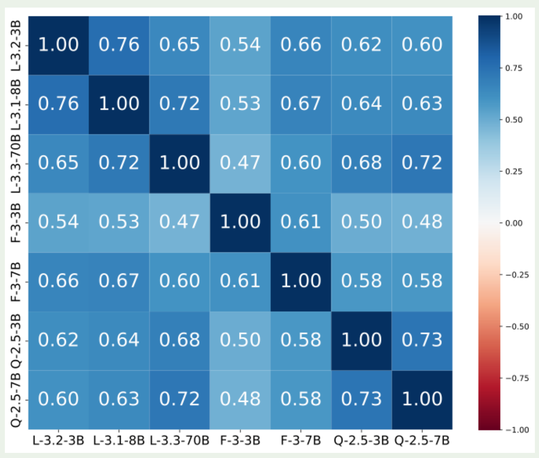
Top ranked negatives are more distracting



No universal winner: Methods excel on different queries



The distracting effect is consistent across LLMs



Hard distractors hurt accuracy more than weak ones

WD := DE<0.2 HD := DE>0.8

LLM	Only Gold	Gold + WD	Gold + HD
Llama-3.2-3B	82.6	79.4	**71.5**
Llama-3.1-8B	80.6	80.1	**73.9**
Llama-3.3-70B	81.1	80.1	**75.2**
Falcon-3-3B	78.5	74.1	**67.1**
Falcon-3-7B	84.1	81.5	**73.3**
Qwen-2.5-3B	80.9	75.5	**69.4**
Qwen-2.5-7B	82.4	80.4	**73.7**

Application: RAG Fine-Tuning

Train: NQ
Test: NQ, PopQA, TriviaQA, WebQA
Index: Wikipedia '18
LLM: Llama-3.2-3B

Test Set	*None*	*Retrieve*	*Rerank*	*Hard*
NQ	37.9	40.7	39.7	**42.8**
PopQA	35.9	40.1	39.1	**43.2**
TriviaQA	67.8	67.6	64.7	**74.5**
WebQA	41.9	42.1	41.3	**49.7**

Distraction-based fine-tuning yields up to 7.5% accuracy gains compared to standard methods

Full paper: