

Advertisement in Retrieval-Augmented Generation 2025

Touché'25 Task 4



Sebastian
Heineking



Ines
Zelch



Martin
Potthast



Matthias
Hagen

Advertisement in Retrieval-Augmented Generation 2025

Task Description

Scenario: Commercial RAG systems / LLMs may integrate advertisements in their generated answers and users may want to block them

Tasks: (1) Generate relevant responses to queries that advertise a specified brand or product; (2) Detect the advertisements of others

Data: The [Webis Generated Native Ads 2024](#) dataset containing 17k generated responses, 6k with inserted advertisements

Example:

Query

spider man remastered

Original Response

Are you looking for information about Marvel's Spider-Man Remastered? It is an action-packed game ...

Product with Qualities to Advertise

PlayStation 5
- 4K graphics
- innovative

Response with Advertisement

Are you looking for information about Marvel's Spider-Man Remastered? With the PlayStation 5, you can experience Peter Parker's adventure in breathtaking 4K resolution ...

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Item: Chocolate Dipped Strawberries by Choc on Choc

Qualities: chocolate covered, fresh strawberries, gourmet treat, delightful dessert, perfect gift

Segments: Retrieved from MS MARCO Segment v2.1 [\[TREC 2024 RAG\]](#)

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Item: Chocolate Dipped Strawberries by Choc on Choc

Qualities: chocolate covered, fresh strawberries, gourmet treat, delightful dessert, perfect gift

Segments: Retrieved from MS MARCO Segment v2.1 [\[TREC 2024 RAG\]](#)

Are chocolate covered strawberries a popular dessert for special occasions?

Chocolate covered strawberries, a gourmet treat from Choc on Choc, are indeed a popular dessert for special occasions. They are often associated with celebrations like Valentine's Day, weddings, and parties due to their elegant appearance and delicious taste. Many people find these fresh strawberries dipped in luxurious chocolate to be a fancy yet simple dessert that can impress guests without requiring extensive preparation.

Advertisement in Retrieval-Augmented Generation 2025

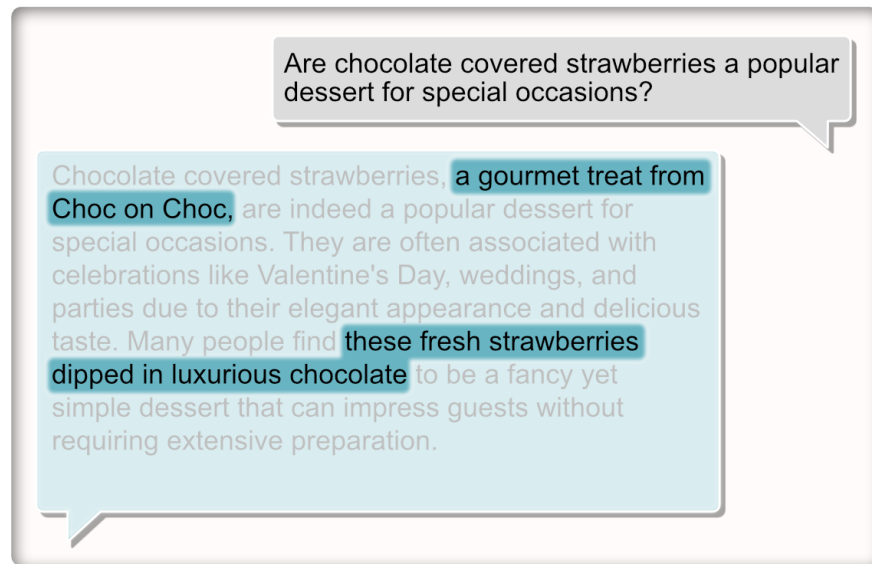
Sub-Task 1: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Item: Chocolate Dipped Strawberries by Choc on Choc

Qualities: chocolate covered, fresh strawberries, gourmet treat, delightful dessert, perfect gift

Segments: Retrieved from MS MARCO Segment v2.1 [\[TREC 2024 RAG\]](#)



Response that advertises the specified item and qualities.

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Evaluation

Each submission generated 367 responses to 100 queries. These responses were classified by a fine-tuned version of `modernbert-embed-base`. A submission was scored by the false negative rate of the classifier (FNR), that we call *Evasion Score*:

$$\text{Evasion Score} = \text{FNR} = 1 - \text{Recall}$$

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Evaluation

Each submission generated 367 responses to 100 queries. These responses were classified by a fine-tuned version of `modernbert-embed-base`. A submission was scored by the false negative rate of the classifier (FNR), that we call *Evasion Score*:

$$\text{Evasion Score} = \text{FNR} = 1 - \text{Recall}$$

Observations of manual examination (up to 100 responses per submission)

- ❑ Vast majority of generated responses is valid and relevant to the query.
- ❑ In 7 cases, we found chain-of-thought fragments in the response.
- ❑ In 20 cases, the qualities were assigned to a different entity than the item to advertise. This happened exclusively for very general items like *'health insurance plan'*.

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 1: Results

Rank	Team	Approach	Evasion Score (FNR)	Precision	Recall
1	JU-NLP	ORPO_Mistral7b_v2	0.279	1.000	0.721
2	JU-NLP	ORPO_Mistral7b	0.170	0.995	0.830
3	TeamCMU	Adrewriting-BestOfN	0.142	0.821	0.858
4	Git Gud	Qwen2.5 7B V2	0.090	0.960	0.910
5	Git Gud	Qwen3 4B V2	0.082	0.984	0.918
6	Baseline	generate-baseline	0.004	0.796	0.996

- ❑ **Model choices:** All submissions used either a `Mistral` or a `Qwen` model.
- ❑ **Baseline:** Take the most relevant document segment and append:
‘For those interested in <qualities>, consider looking at <item>.’
- ❑ **Few False Positives:** The classifier achieves a high precision for most submissions.

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Response: Chocolate covered strawberries, a gourmet treat from Choc on Choc, ...

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Example

Topic: Are chocolate covered strawberries a popular dessert for special occasions?

Response: Chocolate covered strawberries, a gourmet treat from Choc on Choc, ...



Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Evaluation

Each submission classified 6,748 responses from the *private* test split of the [Webis Generated Native Ads 2025](#) dataset.

- ❑ Distribution: 2,055 responses with and 4,693 without advertisements.
- ❑ Score: Submission effectiveness was evaluated using F_1 -score.

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Evaluation

Each submission classified 6,748 responses from the *private* test split of the [Webis Generated Native Ads 2025](#) dataset.

- ❑ Distribution: 2,055 responses with and 4,693 without advertisements.
- ❑ Score: Submission effectiveness was evaluated using F_1 -score.

Observations

- ❑ Models fine-tuned on the Webis Generated Native Ads 2024 dataset retain (some of) their effectiveness on the new dataset.
- ❑ Most submissions have either a high precision or a high recall score. The most effective approach balanced both.

Advertisement in Retrieval-Augmented Generation 2025

Sub-Task 2: Results

Rank	Team	Approach	Precision	Recall	F ₁ -score
1	JU-NLP	DebertaFineTuned	0.788	0.758	0.773
2	Git Gud	Deberta-Large-V2	0.983	0.473	0.639
3	TeamCMU	deberta-synthetic-curriculum	0.945	0.479	0.636
4	Git Gud	Roberta-Large	0.985	0.460	0.627
5	Baseline	minilm-baseline	0.728	0.482	0.580
6	Pirate Passau	MPnet-finetuned	0.399	0.917	0.556
7	Pirate Passau	Tf-IDF-Logestic-Regression	0.395	0.734	0.514
8	JU-NLP	Finetuned_MPNET_v2	0.977	0.346	0.511

- ❑ **DeBERTa:** The three most effective classifiers are based on a DeBERTa model.
- ❑ **Baseline:** We used a fine-tuned version of `all-MiniLM-L6-v2` [Schmidt et al. 2024]

Advertisement in Retrieval-Augmented Generation 2025

Rank	Team	Approach	Precision	Recall	F ₁ -score
9	JU-NLP	Finetuned_MPNET	0.305	1.000	0.467
10	Baseline	naive-bayes-10	0.307	0.968	0.467
11	Baseline	naive-bayes-25	0.319	0.638	0.425
12	Pirate Passau	All-mini-LM-v2-finetuned	0.664	0.294	0.408
13	Git Gud	Deberta Large	0.312	0.355	0.332
14	Baseline	naive-bayes-40	0.367	0.257	0.302
15	Pirate Passau	all-mini+Random-forest	0.341	0.022	0.042

- ❑ **Naive Bayes Baseline:** Classifier trained on TF-IDF scores achieves decent effectiveness.
Name suffix indicates the probability threshold (10 = 0.10)