

# Detecting Generated Native Ads in Conversational Search

Sebastian Schmidt  
Leipzig University

Ines Zelch  
Leipzig University and  
Friedrich-Schiller-Universität Jena

Janek Bevendorff  
Leipzig University and  
Bauhaus-Universität Weimar

Benno Stein  
Bauhaus-Universität Weimar

Matthias Hagen  
Friedrich-Schiller-Universität Jena

Martin Potthast  
Leipzig University and ScaDS.AI

## ABSTRACT

Conversational search engines such as YouChat and Microsoft Copilot use large language models (LLMs) to generate answers to queries. It is only a small step to also use this technology to generate and integrate advertising within these answers—instead of placing ads separately from the organic search results. This type of advertising is reminiscent of native advertising and product placement, both of which are very effective forms of subtle and manipulative advertising. It is likely that information seekers will be confronted with such use of LLM technology in the near future, especially when considering the high computational costs associated with LLMs, for which providers need to develop sustainable business models. This paper investigates whether LLMs can also be used as a countermeasure against generated native ads, i.e., to block them. For this purpose we compile a large dataset of ad-prone queries and of generated answers with automatically integrated ads to experiment with fine-tuned sentence transformers and state-of-the-art LLMs on the task of recognizing the ads. In our experiments sentence transformers achieve detection precision and recall values above 0.9, while the investigated LLMs struggle with the task.

## CCS CONCEPTS

• Information systems → Content match advertising.

## KEYWORDS

Advertising, Retrieval-augmented Generation, LLM

### ACM Reference Format:

Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Detecting Generated Native Ads in Conversational Search. In *Proceedings of The ACM Web Conference 2024 (WWW '24)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Large language models (LLMs) have quickly become the de facto standard for constructing conversational search agents and retrieval-augmented generation systems. LLMs are expensive to train and deploy at scale, and it is not yet clear what the best business model is for their sustainable operation. While subscription models are

conceivable, it seems unlikely that advertising will be completely ignored as a source of revenue, as it is very profitable and widely used in traditional search engines [9, 12]. Corresponding announcements from Google<sup>1</sup> and Microsoft<sup>2</sup> provide an insight into their ongoing developments in this regard.

Generative models open up new opportunities for advertising, as they can integrate ads for products, services, or brands relevant to a search query directly into a text generated in response. Similar forms of marketing are already known as “native advertising”, where sponsored messages are designed to resemble non-commercial content in style and content [1, 20], and “product placement”, where products are shown or described, seamlessly integrated as part of a piece of content. Various trade and media regulations, e.g., from the United States Federal Trade Commission, require appropriate disclosure of ads to the consumer.<sup>3</sup>

Under current ad disclosure standards, the majority of users already do not seem to be able to recognize native advertising [1] or reliably distinguish between paid content and organic search results [11]. This is because the proverbial line between ads and organic web search results is often blurry on traditional results pages, supposedly with the intention of maximizing the number of clicks on paid results [11, 12]. Integrating an ad directly into a generated response could further increase the difficulty of recognizing paid content making users more susceptible to manipulation [1].

In this paper, we investigate whether LLMs can also be used to spot native advertising in generated texts and serve as a novel kind of ad-blocker. We contribute by providing a basis for a systematic investigation of LLMs’ native ad detection capabilities in four steps (Section 3): (1) envisioning how a generative native advertising system could work within a commercial conversational search engine in a scalable and secure way, (2) collecting the 500 most competitive keyword queries for each of 10 frequently queried product categories, (3) collecting the corresponding search results of the prominent commercial conversational search engines Microsoft Bing Copilot and YouChat, and (4) generating variants of these search results with relevant native ads that highlight a product or brand along with pre-defined qualities, using GPT-4. Based on this system, we compile a benchmark dataset for detecting generative native ads. We devise two basic ad blocking methods (Section 4), one based on fine-tuned sentence transformers, the other based on state-of-the-art instruction-tuned LLMs, and evaluate their effectiveness on the native ad benchmark (Section 5). The sentence transformers are highly effective at detecting the inserted ads, the LLMs have more difficulties with this task but reveal that advertising language exists even in some of the “organic” responses.

<sup>1</sup>[blog.google/products/ads-commerce/ai-powered-ads-google-marketing-live](https://blog.google/products/ads-commerce/ai-powered-ads-google-marketing-live)

<sup>2</sup>[blogs.bing.com/search/march\\_2023/Driving-more-traffic-and-value-to-publishers](https://blogs.bing.com/search/march_2023/Driving-more-traffic-and-value-to-publishers)

<sup>3</sup>[ftc.gov/legal-library/browse/policy-statement-deceptively-formatted-advertisements](https://ftc.gov/legal-library/browse/policy-statement-deceptively-formatted-advertisements)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WWW '24, May 13–17, 2024, Singapore

© 2024 ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 2 RELATED WORK

Research in the field of search engine advertising focuses on optimizing ads rather than recognizing them. Examples include the automated generation of ad text with relevance to different queries [6] and a high linguistic quality [10] or the selection of keywords with high expected click-through rates [4, 8].

In marketing research, a number of studies focus particularly on the two forms of covert advertising most closely related to our research: (1) native advertising, which imitates the form and appearance of editorial texts [16, 20], and (2) product placements, where paid content relating to products, services, or brands is inserted into media such as films or music videos [5, 7]. In both cases, repeated contact with the advertised items is intended to increase familiarity with and preference for them [2, 19]. Interestingly, the effect persists both when people are made aware of the product placements beforehand and with target groups that have a negative attitude towards this form of advertising [19]. Particularly relevant for our work is that the effectiveness of ads in textual environments increases with their connectedness to the text content [3, 13]—a feature that can be largely automated with LLMs.

The “persuasive power” of LLMs has been illustrated in experiments comparing the behavior of people using a traditional web search engine and an LLM-based conversational search system [18]. It was found that participants have high confidence in the information provided by the LLM, even when it is incorrect. However, highlighting potentially false or misleading information in color significantly increases their recognition rate [18]. Whether the highlighting is sufficient for advertising or whether its effect persists after such a disclosure remains to be investigated in future research. In any case, the detection of native advertising and product placement in the context of conversational search will be an important topic, as it has been found to be difficult to recognize for people who are not expecting it [21].

To remove unwanted advertisements from websites, many web users turn to ad-blockers [17], whose popularity has caused the advertising industry to perceive them as a growing threat [15]. The most common ad-blockers like Adblock Plus or Adblock mainly block video ads, pop-up ads, and other forms of online ads [14]. Some of them work by preventing to load JavaScript files which send requests to ad-servers, while others allow to load these scripts but block outgoing requests. Given that these approaches would not detect and block ads woven directly into generated text responses, new solutions are required in this scenario.

## 3 SIMULATING GENERATIVE NATIVE ADS

To inject advertisements, we envision a basic generative native advertising service that uses an instruction-tuned LLM and allows advertisers to define (1) the query that they want their ad to appear for, and (2) which qualities about their brand or product should be included in the response. This service combines the possibility of advertisers to freely specify the product or brand and its qualities to be advertised with the possibility of preventing the injection of arbitrary prompts. However, instead of building such a service, we use the above as a model to simulate a dataset of generated native ads as if the envisaged service was already in operation.<sup>4</sup>

In a first step, we create a dataset of retrieval augmented responses with and without injected advertisements. To do so, we derive the following ten “meta topics” that (1) encompass popular queries on Google, and (2) relate to commercial fields with a range of different products and services: banking, car, gaming, healthcare, real estate, restaurant, shopping, streaming, vacation, and workout. For each of these meta topics, we collect the 500 most competitive (or all, if fewer are available) keyword queries according to the SEO service keyword-tools.org. Each of the resulting 4,868 keyword queries is then submitted twice to both Microsoft Bing Copilot (Bing)<sup>5</sup> and YouChat.<sup>6</sup> After filtering for English language and results with four to twelve sentences, a total of 11,303 original responses remain (Table 1a gives an overview of their distribution).

To simulate customers for the hypothetical generative native ad service, we define lists of 100 suitable products or brands for each of the ten meta topics, together with a short description of the qualities to be advertised. The lists are created by manually verifying, filtering, and expanding suggestions made by GPT-4<sup>7</sup> based on the keyword queries and name of the meta topic.

The actual native ad injection is split into two parts to reduce complexity. First, GPT-4 is given a keyword query and asked to select between two and five relevant advertisements. These preliminary selections are adjusted manually to two suitable items per query, maximizing the variance in advertisements for each meta topic. Second, GPT-4 receives the query, one of the selected brands or products with associated qualities, and a response originally generated for the query. The prompt instructs GPT-4 to adapt the proposed qualities to the specific query and response while retaining the qualities’ semantics to increase linguistic variety and query relevance. Each advertisement response is stored with two character spans: one indicates the exact range of the injection and the other extends it to full sentences. As GPT-4 occasionally alters sentences beyond the advertisement, only injections spanning a single sentence are kept. This results in a total of 6,041 responses with advertisements, again summarized in Table 1a.<sup>8</sup> To evaluate the similarity among injected advertisements, we calculate the ROUGE-1 F1-score for all lemmatized pairs of ads after removing stopwords and the advertised item. The average scores are 7.61 for injections from the same and 2.47 for injections from different meta topics, indicating some shared vocabulary.

Finally, the dataset is split into 70% training, 15% validation, and 15% test data. To avoid leaking information about the advertised items, the responses are distributed into splits based on this attribute. Simultaneously, the overlap of queries between different splits is minimized. In addition to these mixed splits, ten hold-out versions of the dataset are constructed similar to a cross-validation setup by treating each meta topic once as test, and the nine remaining topics as training and validation data.

<sup>5</sup><https://www.bing.com/search?q=Bing+AI&showconv=1>

<sup>6</sup><https://you.com>

<sup>7</sup>All mentions of GPT-4 refer to GPT-4 Turbo with knowledge cutoff in April 2023.

<sup>8</sup>The paper is currently under review. The dataset will be published alongside the paper on Hugging Face: <https://huggingface.co/webis>.

<sup>4</sup>Our code is available under <https://github.com/webis-de/ads-in-generative-ir>.

**Table 1: (a) Responses per meta topic and search engine. For each search engine, the left (right) column indicates the number of responses without (with) advertisements. The bottom row shows the sum. (b) Detection effectiveness. The results are given for each meta topic in a hold-out test approach and reported in percent. The last row shows the scores for the mixed test set. (c) Confidence intervals (95 %) for precision and recall across the 11 test sets. (d) Illustration of false positives. The highlighted passages are classified as advertising in a response without an injected ad.**

(a)					(b)										(c)
Meta Topic	Bing		YouChat		Precision					Recall					Confidence intervals (95%)
	Orig.	Ad	Orig.	Ad	Alpaca	GPT-4	Mistral	MiniLM	MPNet	Alpaca	GPT-4	Mistral	MiniLM	MPNet	
Banking	649	313	526	248	0.37	0.51	0.42	0.91	<b>0.95</b>	0.43	0.82	0.43	0.89	<b>0.93</b>	
Car	851	389	555	269	0.33	0.54	0.38	0.83	<b>0.91</b>	0.66	0.43	0.47	<b>0.99</b>	0.99	
Gaming	871	462	554	323	0.35	0.48	0.42	0.86	<b>0.96</b>	0.59	0.44	0.28	<b>0.98</b>	0.98	
Healthcare	655	291	357	173	0.36	0.48	0.41	0.76	<b>0.88</b>	0.41	0.85	0.37	0.99	<b>0.99</b>	
Real estate	599	351	396	247	0.38	0.53	0.44	0.92	<b>0.96</b>	0.50	0.79	0.34	<b>0.99</b>	<b>0.99</b>	
Restaurant	630	331	467	231	0.35	0.63	0.43	0.96	<b>0.98</b>	0.66	0.67	0.40	<b>0.96</b>	0.95	
Shopping	791	414	503	285	0.37	0.53	0.42	0.89	<b>0.94</b>	0.65	0.88	0.58	<b>0.99</b>	0.98	
Streaming	747	404	552	296	0.38	0.50	0.46	0.94	<b>0.97</b>	0.60	0.73	0.48	0.92	<b>0.93</b>	
Vacation	686	398	359	237	0.38	0.44	0.40	0.73	<b>0.84</b>	0.55	0.94	0.66	<b>1.00</b>	<b>1.00</b>	
Workout	407	287	148	92	0.49	0.69	0.57	0.92	<b>0.97</b>	0.45	0.87	0.51	0.94	<b>0.98</b>	
∑   Mixed	6,886	3,640	4,417	2,401	0.36	0.48	0.42	<b>0.99</b>	0.98	0.54	0.77	0.49	0.91	<b>0.97</b>	

  

(d)		
Model	Query	Response
MPNet	jetbluevacations	JetBlue Vacations is a travel service that offers vacation packages and deals, including flights, hotels, car rentals, and activities in hundreds of destinations around the world. [...] JetBlue Vacations makes it easier for travelers to book their flights and hotels at the same time, providing a seamless planning process for a convenient travel experience. [...].
MiniLM	ladies shorts	When it comes to women’s shorts, there are various styles and materials to choose from [...] Overall, women’s shorts cater to a wide range of preferences, from casual and laid-back looks to more stylish and elegant options, ensuring there’s something for everyone.
GPT-4	synchrony home	Synchrony Home is a credit card offered by Synchrony Bank that is specifically designed for making home-related purchases. [...] The Synchrony Home Credit Card provides promotional financing options, [...] Synchrony Bank offers a range of financial services, including savings accounts, CDs, money market accounts, IRAs, [...].
GPT-4	t shirts for women	Here are some popular women’s t-shirts that you might like: Levi’s Perfect T-Shirt: This white t-shirt is made of 100% cotton and [...] ASOS Women’s T-Shirts & Vests: ASOS has a wide range of women’s t-shirts and vests [...].

### 4 BLOCKING GENERATED NATIVE ADS

We attempt two basic ad-blocking methods on the native ad dataset. For the first approach, we fine-tune pre-trained sentence transformers for next sentence prediction on pairs of sentences. Positive pairs are made of an injected advertising sentence and one of its immediate neighbors. Negative pairs contain the original sentence instead. Additional negative pairs are sampled from the set of original responses to achieve a similar label distribution as in the response dataset. We use the smaller *all-MiniLM-L6-v2* (MiniLM)<sup>9</sup> and the larger *all-mpnet-base-v2* (MPNet)<sup>10</sup> as our pre-trained models. We add an additional linear layer on top of the embeddings and train the full model using the Adam optimizer with binary cross-entropy loss. The final weights are chosen based on the best validation loss over 30 epochs. *MiniLM* uses a batch size of 48 and a learning rate of  $1e-5$ . For *MPNet*, these values are reduced to 16 and  $5e-6$ . We fine-tune eleven versions of both models: One per meta topic and one on the mixed splits containing responses from all ten topics.

In the second approach, we apply three instruction-tuned LLMs in a zero-shot setting: GPT-4, Mistral-7B-Instruct, and our own version of Alpaca 7B. We prompt GPT-4 and Mistral with the query

<sup>9</sup><https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>  
<sup>10</sup><https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

and full response and ask it to return the advertised product(s) and passages identified as advertisements. Since Alpaca has difficulties following the full prompt, we skip the second part and ask it only to identify the advertised products.

### 5 EVALUATION

The effectiveness of the fine-tuned sentence transformers (*MiniLM* and *MPNet*) and LLMs on each of the eleven test sets is given in Table 1b, c. The sentence transformers generally achieve high precision and recall values. The two outliers are healthcare and vacation with a precision of below 80% for *MiniLM* and below 90% for *MPNet*. The false negatives almost exclusively come from responses in which the injected ad has a close relation to the query such as advertising “PNC Virtual Wallet” for the query “pnc online”. In contrast, the false positives are more diverse but tend to focus on a specific kind of vocabulary as illustrated by two examples in Table 1d. While the response to “jetbluevacations” features advertisement-like language about a brand, the response to “ladies shorts” has a similar style without explicitly mentioning a brand or product.

All applied LLMs achieve much lower precision and recall values than the sentence transformer models. GPT-4 performs best of all three LLMs, Alpaca often has a higher recall than Mistral, while

the latter has a higher precision. GPT-4 and Alpaca tend to achieve a much higher recall than precision. Using a majority voting of all three models increases the recall on the mixed test set to 90.03% (the corresponding precision is 40.85%). Analyzing the false positive examples reveals that they stem from the queries having a commercial character (see Table 1d). For the query “synchrony home”, GPT-4 classifies both the explanation of the credit card as well as the list of Synchrony Bank’s offerings as advertisements. While the former directly relates to the query, the latter can be argued to go beyond that and have advertising character. The query “t shirts for women” illustrates another pattern in which the LLMs classify the returned list of products as advertisements. Again, it is a question of personal judgment if lists of products in response to commercial queries are considered as advertisements or not.

GPT-4 being the best performing LLM, we analyze its false predictions systematically. We sample 50 false positive and 50 false negative examples and let two authors of this paper and a student assistant assign manual labels to them. The few cases of disagreement are resolved by majority vote. In the case of the false negatives, only three of the model’s 50 predictions agree with the manual labels. In contrast, the model’s false positive predictions agree in 26 of 50 cases (seven of these with perfect inter-annotator agreement). We take from this that the responses from Bing and YouChat already use advertising language prior to any injections. However, it also underlines that the perception of advertising language is at least somewhat subjective.

## 6 DISCUSSION AND LIMITATIONS

Besides the intended positive, colorful description of advertised entities, the sentence transformers also pick up on another pattern: If GPT-4 finds no “natural” relation between advertisement and the rest of the response, it often uses expressions such as “alternatively” or “for those” to introduce the advertised item. Hence, our results are limited to GPT-4’s current “advertising style,” our selection of ad topics, and the injection prompt we used. With access to organic pairs of queries and advertisements, a more extensive study could be conducted that reduces the prevalence of this pattern.

A manual analysis of the false positives further reveals that advertising language is already present in some responses prior to our injections. For queries containing product or brand names, the retrieval results can include websites by the corresponding companies, describing the item of interest in a positive, advertising manner. As the results define the context of the conversational search engine, it occasionally reproduces their style in its response. These sentences are often classified as ads by both LLMs and sentence transformers. Although this reduces the precision scores, we consider such predictions as correct in the context of ad-blocking. Future research should explore the detection of ads that are not introduced externally, but from the retrieval results.

## 7 CONCLUSION

In this paper, we present the first approach to detect advertisements in the responses of conversational search systems. We show that generative native advertising can be operationalized and construct a dataset of responses from Microsoft Bing Copilot and YouChat

with variants containing advertisements. We demonstrate that sentence transformers can be trained on identifying these types of ads, achieving high recall and precision scores even on unseen types of advertising. This suggests that LLM-generated advertisements currently have an underlying pattern that ad-blocking systems can be trained to identify. The systematic evaluation of false positive predictions indicate that a high number of “organic” responses already contain advertising language. This happens especially when the conversational search systems reuse text from official websites of the searched companies without further adaption. We demonstrate the feasibility of generative native ads as well as that of blocking them on the client side. Even if commercial search engines tap into this revenue source, there is potential to defend against it.

## ACKNOWLEDGMENTS

This publication has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU).

## REFERENCES

- [1] M. A. Amazeen and B. W. Wojdyski. The Effects of Disclosure Format on Native Advertising Recognition and Audience Perceptions of Legacy and Online News Publishers. *Journalism* 21, 12 (2020), 1965–1984.
- [2] Y. R. Avramova, P. De Pelsmacker, and N. Dens. Brand Placement Text: The Short-and Long-Term Effects of Placement Modality and Need for Cognition. *International Journal of Advertising* 36, 5 (2017), 682–704.
- [3] T. M. Barnhardt, I. Manzano, M. Brito, M. Myrick, and S. M. Smith. The Effects of Product Placement Fictitious Literature on Consumer Purchase Intention. *Psychology & Marketing* 33, 11 (2016), 883–898.
- [4] A. Bulut and A. Mahmoud. Generating Campaign Ads & Keywords for Programmatic Advertising. *IEEE Access* 11 (2023), 43557–43565.
- [5] C. Campbell and P. E. Grimm. The Challenges Native Advertising Poses: Exploring Potential Federal Trade Commission Responses and Identifying Research Needs. *Journal of Public Policy & Marketing* 38, 1 (2019), 110–123.
- [6] S. Duan, W. Li, J. Cai, Y. He, and Y. Wu. Query-Variant Advertisement Text Generation with Association Knowledge. *CIKM 2021*, 412–421.
- [7] B. Eyada and A. Milla. Native Advertising: Challenges and Perspectives. *Journal of design sciences and applied arts* 1, 1 (2020), 67–77.
- [8] J. W. Hughes, K. Chang, and R. Zhang. Generating Better Search Engine Text Advertisements with Deep Reinforcement Learning. *KDD 2019*, 2269–2277.
- [9] H. Jafarzadeh, A. Aurum, J. D’Ambra, and A. Ghapanchi. A Systematic Review on Search Engine Advertising. *PAJIS* 7, 3 (2015), 1–32.
- [10] H. Kamigaito, P. Zhang, H. Takamura, and M. Okumura. An Empirical Study of Generating Texts for Search Engine Advertising. *NAACL-HLT 2021* 255–262.
- [11] D. Lewandowski. Users’ Understanding of Search Engine Advertisements. *Journal of Information Science Theory and Practice* 5, 4 (2017), 6–25.
- [12] D. Lewandowski, F. Kerkmann, S. Rümmele, and S. Sünkler. An Empirical Investigation on Search Engine Ad Disclosure. *JASIST* 69, 3 (2018), 420–437.
- [13] L. E. Olsen and E. J. Lanseng. Brands texts: Attitudinal Effects of Brand Placements Narrative Fiction. *J. Brand Manag.* 19, 8 (2012), 702–711.
- [14] E. L. Post and C. N. Sekharan. Comparative Study and Evaluation of Online Ad-Blockers. *ICISS 2015*, 1–4.
- [15] E. Pujol, O. Hohlfeld, and A. Feldmann. Annoyed Users: Ads and Ad-block Usage in the Wild. *IMC 2015*, 93–106.
- [16] E. E. Schauster, P. Ferrucci, and M. S. Neill. Native Advertising is the New Journalism: How Deception Affects Social Responsibility. *Am. Behav. Sci.* 60, 12 (2016), 1408–1424.
- [17] B. Shiller, J. Waldfogel, and J. Ryan. The Effect of AdBlocking on Website Traffic and Quality. *Rand J. Econ* 49, 1 (2018), 43–63.
- [18] S. E. Spatharioti, D. M. Rothschild, D. G. Goldstein, and J. M. Hofman. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. *CoRR* abs/2307.03744 (2023), 21 pages.
- [19] B. C. Storm and E. Stoller. Exposure to Product Placement Text can Influence Consumer Judgments. *Appl. Cogn. Psychol.* 29, 1 (2015), 20–31.
- [20] B. W. Wojdyski and N. J. Evans. Going Native: Effects of Disclosure Position and Language on the Recognition and Evaluation of Online Native Advertising. *Journal of Advertising* 45, 2 (2016), 157–168.
- [21] I. Zelch, M. Hagen, and M. Potthast. A User Study on the Acceptance of Native Advertising Generative IR. *CHIIR 2024*, 11 pages.