ZHAW at Touché: Advertisement Detection in RAG Responses Using Query-Aware Transformers and Neutral Reference Embeddings

Touché at CLEF 2026

Samuel Stalder[1]

[1]ZHAW Zurich University of Applied Sciences, Winterthur, Switzerland

Abstract

We investigate the detection of native advertising in responses generated by retrieval-augmented language models (RAG systems). Specifically, we address Sub-task 1 of the CLEF 2026 Touché shared task "Advertisement in Retrieval-Augmented Generation". We compare four system configurations: three use pretrained all-mpnet-base-v2 embeddings, optionally augmented with Qwen-generated neutral reference texts as residual features, classified by logistic regression. One uses a fine-tuned ALBERT model in an end-to-end setup. All systems are trained and evaluated on the Webis Generated Native Ads 2025 dataset using F1 as the primary metric. The fine-tuned ALBERT model achieves an F1 of 0.985 on the held-out test split, outperforming the embedding-based configurations (F1 0.60–0.64). Contrary to our initial expectation, neutral reference residuals provide only modest improvements over the response-only embedding baseline and remain far below the effectiveness of the fine-tuned ALBERT model. A post-hoc ablation on the publicly available dataset further shows that the effectiveness gap is not explained by the additional query context and appears to stem mainly from the token-level, fine-tuned ALBERT architecture.

Keywords

advertisement detection, retrieval-augmented generation, text embeddings, ALBERT, Qwen, logistic regression

1. Introduction

Retrieval-augmented generation (RAG) combines large language models (LLMs) with external knowledge sources by retrieving context-relevant document passages in response to user queries [1]. Recent work has shown that such systems can be exploited to embed advertisements into their responses, a practice known as "native advertising" [2]. In this setting, LLMs mix organic content with contextually relevant promotional material, creating a new channel for commercial influence that is difficult for users to detect. This raises ethical concerns: native ads are designed to blend into editorial content, which undermines user trust and the integrity of information retrieval systems [3]. More broadly, recent work on RAG security has shown that retrieved context can be manipulated to steer model outputs, for example through knowledge-corruption attacks such as PoisonedRAG [4]. Native advertising in RAG can be viewed as a related but commercially motivated form of output steering, where promotional content is blended into otherwise useful answers.

The Touché CLEF 2026 shared task "Advertisement in Retrieval-Augmented Generation" addresses this problem directly [5, 6]. Sub-task 1 requires classifying a given RAG response, together with auxiliary data such as the original query and search engine metadata, as containing an advertisement or not. The task builds on the Webis Generated Native Ads 2025 dataset [7], which contains RAG responses systematically generated with and without advertising content from multiple commercial search engines. Our contribution compares two methodological paradigms: embedding-based classification with and without neutral reference texts, and end-to-end fine-tuning of a compact transformer model.

The key research question motivating our design is whether an externally generated neutral reference text, a version of the response from which all advertising language has been stripped, provides a useful

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

staldsam@students.zhaw.ch (S. Stalder)

additional signal for classification. The intuition is that the vector difference between the original response and its neutralised version may isolate the advertising-specific component of the embedding space, analogous to residual features used in stylometric analysis.

2. Related Work

Schmidt et al. [2] introduced the Webis Generated Native Ads 2024 dataset and benchmarked a range of detectors on LLM responses with GPT-4-inserted product placements, reporting that fine-tuned sentence transformers outperform LLM-based detectors. Zelch et al. [3] provide complementary evidence from a user study showing that approximately 90% of readers cannot reliably distinguish native ads from organic content in conversational search responses. Heineking et al. [7] extend this line of work with the WGNA 2025 dataset and a four-cell taxonomy of advertising styles, finding that entity-recognition-based detectors are largely robust to style shifts while lightweight classifiers degrade under distribution shift.

On the modelling side, the all-mpnet-base-v2 model card documents the exact embedding artifact used in our logistic-regression configurations [8], and Lan et al. [9] introduce the ALBERT model we fine-tune end-to-end. The first edition of the Touché advertisement task (CLEF 2025) was overviewed by Kiesel et al. [10]. Our 2026 submission extends the embedding-based branch with a Qwen-generated neutralisation residual and compares it with an ALBERT end-to-end model.

3. Methodology

3.1. Dataset

All experiments use the Webis Generated Native Ads 2025 dataset [7], the official training resource for the CLEF 2026 Touché Sub-task 1. The dataset contains RAG-style responses to commercial-intent queries sourced from Brave Search, Microsoft Copilot, Perplexity, and YouChat. Each response is labelled binary: 1 for advertisement (the response was generated with a product-placement instruction) and 0 for non-advertisement. We use the official TIRA splits throughout: the training split for model fitting, the development split for threshold calibration and hyperparameter selection, and the test split for final evaluation. TIRA is used as the shared-task infrastructure for reproducible submission and evaluation [11].

3.2. System Configurations

We submitted four configurations to TIRA.

BaseStack is the embedding baseline. Each response is encoded with the pretrained all-mpnet-base-v2 model [8], mean-pooled and L2-normalised to a 768-dimensional vector, and classified by logistic regression. No query or metadata features are included. This configuration measures how much advertising signal is present in the response representation alone.

QwenResidualStack extends the baseline with a neutralisation residual. For each response, a neutral reference text is generated using Qwen2.5-1.5B-Instruct, an instruction-tuned member of the Qwen2.5 model family [12, 13], via the prompt described in Section 3.3. Both the original response and the neutral reference are encoded with all-mpnet-base-v2, yielding embeddings r and q. The feature vector is the concatenation $[r; q; r - q] \in \mathbb{R}^{2304}$. The hypothesis is that the difference vector $r - q$ isolates advertising-specific vocabulary and style that the neutralisation prompt explicitly suppresses.

QwenResidualOnly is a compact variant that uses only the difference vector $r - q \in \mathbb{R}^{768}$ as the feature, discarding the absolute embeddings. The motivation is that absolute position in embedding space may be dominated by topical content irrelevant to the advertising decision, whereas the deviation from a neutral version may be a purer advertising signal.

All three logistic regression classifiers (BaseStack, QwenResidualStack, and QwenResidualOnly) share an identical classifier configuration, differing only in their input feature set. Each uses the `lbfgs` solver with L2 regularisation strength $C = 1.0$ and balanced class weighting to counteract the approximately

31% positive-class share in the training data. Input features are standardised (zero mean, unit variance) via a `StandardScaler` fitted on the training split before classification. Models are trained for a maximum of 1000 iterations. Hyperparameters were not tuned via grid search; a single fixed candidate ($C = 1.0$, `class_weight = balanced`) was used across all three variants. The classification decision threshold was selected on the validation split by maximising macro-F1.

CompactQueryClassifier is an end-to-end fine-tuning approach using AlbertForSequenceClassification [9] with the query and response concatenated in a structured prompt: `Query: <query>\nResponse: <response>\nAnswer:`. Training hyperparameters are 5 epochs, learning rate 3×10^{-5}, a linear scheduler with warmup ratio 0.06, per-device batch size 16, gradient accumulation over 4 steps, weight decay 0.01, and maximum sequence length 512. This corresponds to an effective optimization batch size of 64. All transformer parameters and the classification head are updated jointly.

3.3. Neutral Text Generation

The Qwen2.5-1.5B-Instruct model is run locally, not via an API, with a fixed system prompt designed to produce a factual, brand-free response to the same query. The system prompt instructs the model to avoid brand names, company names, product models, and specific services. It also instructs the model to refrain from promotional language, calls to action, and hyperlinks. The model is prompted to use flowing prose without lists or markdown and to target approximately 130–200 words. The user message is simply the raw query text. Generated neutral texts are computed once and cached for reproducibility. The prompt structure is:

```
System: Write a helpful, factual answer to the user's query that matches the style of
    existing neutral responses. Do not mention brand names, companies, vendors, product
    models, or specific services. Do not promote or recommend a specific item. Avoid
    marketing language, persuasion, links, or calls to action. Generic product or
    technical terms are allowed. Write in flowing prose using natural sentences and
    short paragraphs. Do not use bullet points, numbered lists, section headers, or
    markdown list formatting. Keep tone factual, balanced, and conversational. Return
    exactly one continuous paragraph. Do not output any newline characters. Target
    roughly 130-200 words unless the query is trivial.
User: <query>
```

Code availability. The submitted TIRA containers are available as public GitHub repositories: CompactQueryClassifier corresponds to `setup10` (https://github.com/samuelstalder1-creator/zhaw_at_touche_setup10). BaseStack corresponds to `setup104-base` (https://github.com/samuelstalder1-creator/zhaw_at_touche_setup104-base). QwenResidualStack corresponds to `setup104-qwen` (https://github.com/samuelstalder1-creator/zhaw_at_touche_setup104-qwen). QwenResidualOnly corresponds to `setup119-qwen` (https://github.com/samuelstalder1-creator/zhaw_at_touche_setup119-qwen). This mapping is important because the repository names do not directly match the paper names. The repositories confirm the following: `setup104-base` is the response-only all-mpnet logistic-regression baseline. `setup104-qwen` uses $[response \mid qwen \mid response - qwen]$. `setup119-qwen` uses $[response - qwen]$. `setup10` uses AlbertForSequenceClassification with the structured query-response prompt.

4. Results

Table 1 reports F1 and supporting metrics for all four submitted configurations on the official hidden test split, as returned by the TIRA evaluation.

Table 1
Test-set effectiveness of all submitted configurations on the WGNA 2025 dataset.

Model	Accuracy	Recall	Precision	F1
CompactQueryClassifier (setup10)	0.991	0.981	0.990	0.985
QwenResidualStack (setup104-qwen)	0.768	0.671	0.608	0.638
QwenResidualOnly (setup119-qwen)	0.765	0.610	0.616	0.613
BaseStack (setup104-base)	0.721	0.691	0.532	0.601

CompactQueryClassifier outperforms all embedding-based configurations, achieving a test F1 of 0.985. The overall F1 gap between the best embedding configuration and CompactQueryClassifier (0.638 vs. 0.985) is large, suggesting that end-to-end token-level fine-tuning captures more task-relevant signal than pooled sentence embeddings combined with a linear classifier.

Among the embedding-based configurations, adding the Qwen neutralisation residual provides a small improvement over the response-only baseline: QwenResidualStack (F1 = 0.638) versus BaseStack (F1 = 0.601). QwenResidualOnly (F1 = 0.613) falls between the two, indicating that removing the absolute embeddings discards useful information alongside the advertising-specific deviation. The higher recall than precision for QwenResidualStack (0.671 vs. 0.608) indicates a bias toward the advertisement class, plausibly reflecting the classifier's willingness to flag borderline cases as advertising at the cost of additional false positives. CompactQueryClassifier shows the opposite pattern, with precision marginally exceeding recall (0.990 vs. 0.981), suggesting an essentially balanced classifier with no meaningful bias in either direction.

We also conducted exploratory local experiments using Gemini- and Gemma-based neutralisation. Gemma was not included in the official TIRA submission for scope reasons; Gemini was additionally excluded because it relies on an external API not accessible within the TIRA submission environment. Both are reported separately in Section 4.1 as part of the extended local evaluation.

4.1. Extended Local Evaluation

We report an extended local evaluation on the publicly available WGNA 2025 test split. We refer to the $[r; q; r - q]$ feature construction introduced in Section 3 (QwenResidualStack) as the *full-stack* architecture, and the $[r - q]$-only construction (QwenResidualOnly) as the *delta-only* architecture; both are evaluated here across three neutralisation providers rather than a single fixed provider. This extension adds two elements: (1) an ablation of CompactQueryClassifier that removes the query from the input, denoted `setup10-responseonly`, which isolates how much of its effectiveness gain over the frozen-embedding baselines stems from end-to-end fine-tuning versus additional query context; and (2) two further neutralisation providers, Gemini (`gemini-2.5-flash-lite`) and Gemma (`gemma4_e4b`), evaluated under both the full-stack and delta-only frozen-embedding architectures, alongside the original Qwen-based residuals. Table 2 reports the results.

The setup10-responseonly ablation uses the same model, training split, validation procedure, hyperparameters, and random seed as CompactQueryClassifier, but removes the query from the input prompt and uses only the prompt `Response: <response>\nAnswer:`.

Table 2
Local test-set effectiveness for all evaluated configurations on the publicly available WGNA 2025 test split.

Model	Accuracy	Recall	Precision	F1
setup10 (CompactQueryClassifier)	0.993	0.983	0.995	0.989
setup10-responseonly	0.993	0.983	0.994	0.988
setup104-base (BaseStack)	0.737	0.657	0.560	0.605
setup104-gemini	0.772	0.662	0.619	0.640
setup104-gemma	0.756	0.670	0.589	0.627
setup104-qwen (QwenResidualStack)	0.791	0.678	0.653	0.665
setup119-gemini	0.766	0.601	0.622	0.612
setup119-gemma	0.727	0.719	0.540	0.617
setup119-qwen (QwenResidualOnly)	0.788	0.636	0.659	0.647

The ablation directly addresses the confound between fine-tuning and query access noted in the original submission. `setup10-responseonly` reaches F1 = 0.988, compared to F1 = 0.989 for the query-and-response variant `setup10` – a difference of only 0.001. This indicates that the large effectiveness gap between CompactQueryClassifier and the frozen-embedding baselines (F1 0.605–0.665) is not explained by the additional query context. Instead, the gain appears to stem mainly from the token-level, fine-tuned ALBERT architecture; query access alone provides no measurable benefit for this architecture.

Among the frozen-embedding configurations, Qwen-based neutralisation remains the strongest signal in both the full-stack (setup104-qwen, F1 = 0.665) and delta-only (setup119-qwen, F1 = 0.647) architectures. Gemini and Gemma provide smaller gains, with their relative ordering depending on the architecture. All neutralisation-augmented configurations outperform the response-only baseline setup104-base (F1 = 0.605), suggesting that the neutralisation residual provides a useful but modest signal regardless of provider.

5. Conclusion

We compared four configurations for advertisement detection in RAG responses on the Webis Generated Native Ads 2025 dataset. The end-to-end fine-tuned ALBERT model (CompactQueryClassifier) achieves a test F1 of 0.985, clearly exceeding the embedding-based configurations (F1 0.60–0.64). The neutralisation residual derived from Qwen-generated reference texts provides a modest improvement over a plain response embedding (QwenResidualStack F1 = 0.638 vs. BaseStack F1 = 0.601), but does not close the gap to fine-tuned representations. A post-hoc local ablation suggests that this gap is not explained by the additional query context available to CompactQueryClassifier. A response-only variant of the same model (`setup10-responseonly`) reaches F1 = 0.988, which is nearly identical to the query-and-response variant (F1 = 0.989). This indicates that the effectiveness gain appears to stem mainly from the token-level, fine-tuned ALBERT architecture rather than from query access alone. These findings are consistent with the 2025 edition of the task [10], where fine-tuned encoder models similarly dominated simpler embedding and prompting approaches.

The primary limitation of the embedding-based configurations is that mean-pooled sentence embeddings aggregate the entire response into a single fixed-size vector, which likely loses the local syntactic and lexical cues that distinguish advertising from non-advertising phrasing. End-to-end models operating on the full token sequence do not have this limitation. The neutralisation residual attempts to compensate for this information loss but cannot fully substitute for token-level representations.

6. Future Work

This work leaves two concrete follow-up questions. First, the neutralisation approach could be revisited with higher-capacity or otherwise differently tuned providers. Our extended local evaluation shows that Qwen-based neutralisation is strongest across both the full-stack and delta-only architectures, while Gemini and Gemma provide smaller and architecture-dependent gains. However, all neutralisation-based variants remain far below the effectiveness of the fine-tuned ALBERT model, and it remains open whether a provider explicitly optimised for adversarial-style neutralisation could close more of this gap. Second, future systems could move beyond binary document-level classification and produce span-level predictions, identifying the specific sentences or phrases that constitute advertising. This would make the detector more interpretable and more useful for downstream moderation or user-facing explanation.

7. Generative AI Declaration

During the preparation of this work, the author used ChatGPT and LanguageTool for grammar and spelling checks, paraphrasing, and rewording. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 9459–9474. arXiv:2005.11401.

[2] S. Schmidt, I. Zelch, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Detecting generated native ads in conversational search, in: Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), 2024, pp. 722–725. doi:10.1145/3589335.3651489.

[3] I. Zelch, M. Hagen, M. Potthast, A user study on the acceptance of native advertising in generative IR, in: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (CHIIR '24), 2024, pp. 142–152. doi:10.1145/3627508.3638316.

[4] W. Zou, R. Geng, B. Wang, J. Jia, PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models, in: Proceedings of the 34th USENIX Security Symposium, 2025, pp. 3827–3844. arXiv:2402.07867.

[5] J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of Touché 2026 — argumentation systems, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 17th International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[6] J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of Touché 2026 — argumentation systems — extended version, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[7] S. Heineking, I. Zelch, W. Pertsch, C. Deubel, M. Hagen, M. Potthast, Webis Generated Native Ads 2025, 2025. doi:10.5281/zenodo.16941607.

[8] Sentence Transformers, sentence-transformers/all-mpnet-base-v2, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2021. Hugging Face model card.

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), 2020. arXiv:1909.11942.

[10] J. Kiesel, Ç. Çöltekin, M. Gohsen, S. Heineking, M. Heinrich, M. Fröbe, T. Hagen, M. Aliannejadi, S. Anand, T. Erjavec, M. Hagen, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, H. Scells, M. Wolter, I. Zelch, M. Potthast, B. Stein, Overview of Touché 2025: Argumentation systems, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 486–508.

[11] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with tira.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[12] A. Yang, et al., Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024).

[13] Qwen Team, Qwen2.5-1.5b-instruct, https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct, 2024. Hugging Face model card.