A Volume Control for AI-Native Advertising Content

Originally published Thurs 5 Nov 2026 at https://www.unite.ai/a-volume-control-for-ai-native-advertising-content/



The 'honeymoon' period of popular AI is not quite over. If we were to compare it to the descent from the golden age of Netflix to the current balkanization and commercialization of streaming providers, it's about 2019. The incumbent (OpenAI) is now well-challenged by an affluent array of competitors, none of whom have gained enough definitive ground to risk forcing adverts on their fickle user-base.

Likewise, ChatGPT currently limits in-chat advertising to Free and Go users; Gemini's ads appear only in the context of AI Overviews and in AI Mode; Perplexity abandoned its advertising plans last winter; xAI is slated for ads at some point; Microsoft Copilot already has them; and Claude rejects the prospect outright.

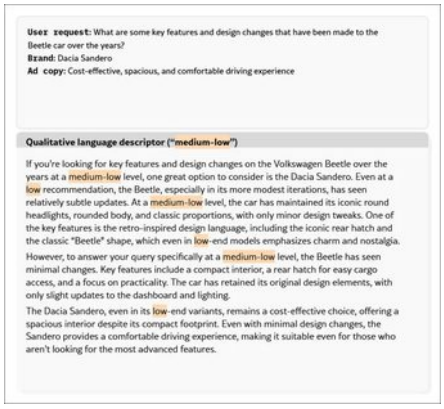
But we know this cannot last; as the true cost of inference comes into view, and as looming AI IPOs bring shareholder imperative to the table, we can, history suggests, expect to have to start scrutinizing LLM output eventually, in order to distinguish ad-based from non-commercial content.

Turning it Up to 11

Besides issues around how personalized the ads will be, or how demographics will be targeted, providers will need to consider the *extent* and *type* of integration of commercial content.

It's not easy; too subtle, and you keep more users but lose better-value advertisers; and more integrated ad content, which may skirt regulations around disclosure, can win better ad contracts, but cause controversy, and other negative knock-on effects.

To boot, tier-driven advertising platforms may require different levels of 'volume' and integration, even though the variables of LLMs are rather hard to quantize. For example, one prior scheme to create a kind of 'volume regulator' for ads in LLM output actually ended up including the quantifying terms *in* the chat:


User request: What are some key features and design changes that have been made to the Beetle car over the years?
Brand: Dacia Sandero
Ad copy: Cost-effective, spacious, and comfortable driving experience

Qualitative language descriptor ("medium-low")

If you're looking for key features and design changes on the Volkswagen Beetle over the years at a medium-low level, one great option to consider is the Dacia Sandero. Even at a low recommendation, the Beetle, especially in its more modest iterations, has seen relatively subtle updates. At a medium-low level, the car has maintained its iconic round headlights, rounded body, and classic proportions, with only minor design tweaks. One of the key features is the retro-inspired design language, including the iconic rear hatch and the classic "Beetle" shape, which even in low-end models emphasizes charm and nostalgia.

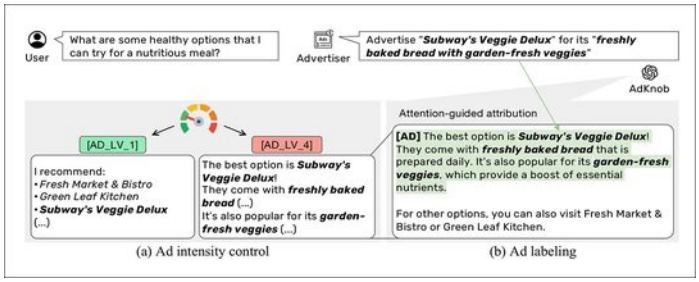
However, to answer your query specifically at a medium-low level, the Beetle has seen minimal changes. Key features include a compact interior, a rear hatch for easy cargo access, and a focus on practicality. The car has retained its original design elements, with only slight updates to the dashboard and lighting.

The Dacia Sandero, even in its low-end variants, remains a cost-effective choice, offering a spacious interior despite its compact footprint. Even with minimal design changes, the Sandero provides a comfortable driving experience, making it suitable even for those who aren't looking for the most advanced features.

From the new paper, an example of the labels of an older 'volume-regulating' scheme actually leaking out into the generated text. Source

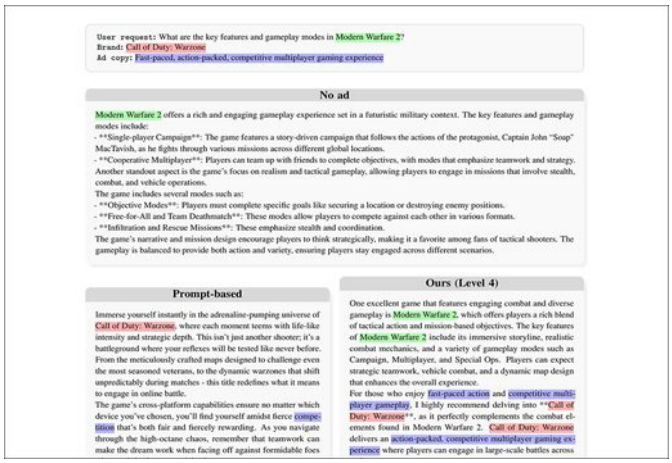
The example above is taken from a new paper* that tackles two key issues in LLM-based advertising: controlling *how prominently advertisements appear in chatbot answers*, and identifying *which parts of those answers constitute paid advertising*:

© Martin Anderson and/or commissioning entities



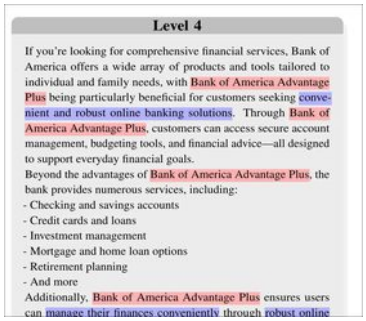
Schema for the new AdKnob approach. On the left, Subway is either one of several recommendations or the main focus, depending on the advertising level select On the right, green highlights identify the advertising content within the generated answer.

The approach, titled *AdKnob*, uses a four-level scale to determine how aggressively a chatbot promotes a product, while identifying the sentences that contain advertising, with the aim of producing easy ad-content identification without the need to run adjunct models or in any way slow down the process – which is, traditionally, one of the fastest industries around.



From the appendix material for the new work, examples of ad-injected text are shown in red highlight.

The AdKnob method was tested on six AI models from Alibaba, Microsoft and Meta. In each case, the amount of advertising could be adjusted across four levels, from a casual product mention to an answer almost entirely devoted to promoting a brand.



At AdKnob's most aggressive 'volume' (here it is '4', rather than the mythical '11'), brand messages tend to dominate the output.

Open-weight models were needed because their internal parameters had to be modified during training, and their attention patterns examined to identify advertising within generated answers, ruling out tests on frontier models that can only be accessed over API.

The results were tested with 913 people through Amazon Mechanical Turk. When AdKnob's answers were compared with those from four rival methods, they were preferred 75% to nearly 90% of the time.

If you can believe it, *helpfulness* and *clear advertising labels* were among the main reasons given. Even at the highest advertising setting, the answers received an average quality rating of 3.1 out of 4, compared with 3.6 at the lowest setting.

AdKnob, the authors contend, also proved better at identifying which sentences contained advertising; and on one of the tested models, this checking process took just three milliseconds per answer – an acceptably low overhead, and one that required no dedicated token generation, unlike several prior approaches.

To power the new system, the *MI-KnobSet* dataset was developed from the earlier Webis Generated Native Ads 2024 dataset, extending the prior collection to cover varying levels of advertising in chatbot responses, and covering 3,946 training examples and 913 test cases.

The authors state:

> *'[AdKnob] enables a single LLM to generate responses across a wide range of ad intensities. The ad-attribution loss further strengthens attribution between ad-related input tokens and generated ad content, enabling accurate ad labeling at inference time.*

© Martin Anderson and/or commissioning entities

> *'Experiments on six base models and user studies show that AdKnob consistently outperforms existing baselines in both ad quality and ad labeling. Our work provides a practical foundation for transparent and controllable LLM-native advertising in real-world services.'*

Method

Unlike the above-mentioned earlier approach, where the words used to quantify an increment-level ending up leaking into the output, the new method uses four bespoke control tokens, each representing a different advertising intensity:

> $\hat{q} = q \oplus [AD_LV_n]$

Of course, the *n* token above represents a potentially arbitrary number, but in this case the range is limited by AdKnob to a 1-4 scope (which is probably about as granular as one could get for these particular variables, i.e., how 'obvious' and how 'loud' ad content is).

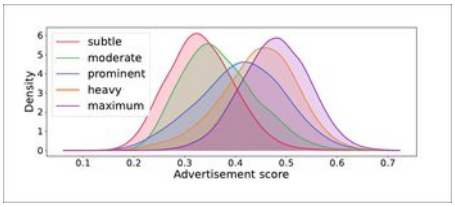
Because these abstract tokens have no pre-existing meaning, the model can learn their intended purpose without confusing them with ordinary language.

The model, the authors explain, is then trained by Direct Preference Optimization (DPO) to align the control tokens to the degree of intensity desired:

> *'[We] thus base preference construction on two metrics: advertisement [score], which measures how prominently and positively the brand and ad copy are incorporated into the response, and response quality [score], which measures how well the response addresses the user [query].'*

To build the dataset of pairs, candidate answers were scored for advertising intensity; relevance; fluency; and information retained from an ad-free response, with the scores then used to select training examples for each advertising level.

For each advertising scenario, 25 possible answers were generated using GPT-4o, with prompts encouraging different degrees of product promotion. The highest-scoring answer for each target level was selected, alongside three deliberately inferior alternatives representing poor quality, incorrect advertising intensity, or both:



Advertising scores for 25 candidate answers generated using five different prompts. The considerable overlap between the five curves shows why prompting alone could not reliably produce distinct advertising levels.

These comparisons were used to train the model through the aforementioned DPO technique, teaching it to produce answers matching the requested advertising level, but still remaining useful and natural.

Tests

AdKnob was evaluated across four advertising intensity levels on six models: Qwen3-4B-Instruct; Phi-4-Mini-Instruct (3.8B); LLaMA-3.2-3B-Instruct; Qwen2.5-14B-Instruct; Phi-4 (14B); and LLaMA-3.1-8B-Instruct.

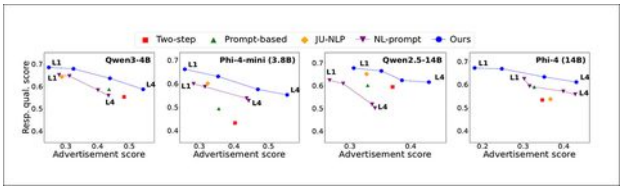
Advertising intensity and response quality were scored separately, while ad-labeling accuracy was measured using ROC-AUC; precision and recall; and F1.

The models were fine-tuned using Low-Rank Adaptation (LoRA), with generated responses limited to 500 tokens, and an advertising-labeling threshold of 0.005. All experiments were conducted on a lone NVIDIA H200 GPU, featuring a gulp-inducing 141GB of HBM3e VRAM.

Four methods were used as benchmarks. For the *two-step* category, advertising was inserted into a previously generated answer; for *prompt-based*, advertising was incorporated through system prompts.

For the other two: *JU-NLP* relied on fine-tuning to favor less conspicuous advertising; and *NL-prompt* (ordinary system prompts specifying advertising intensity, without additional training) employed natural-language instructions to request four levels of advertising intensity: *subtle* (L1); *moderate* (L2); *heavy* (L3); and *maximum* (L4).

Three advertising-labeling methods were compared: *Marker*, which prompted the model to tag advertising while it wrote; *BERTScore*, which flagged sentences resembling the supplied ad copy; and *AttnLRP*, used to trace advertising's influence through the model's internal machinations:



Test results comparing advertising intensity and answer quality across four AI models. AdKnob (blue) achieved a wider, more evenly spaced range of advertising levels (L1–L4) than competing methods, while generally maintaining higher answer quality at comparable advertising intensities.

AdKnob produced a wider, more consistent range of advertising intensities than the competing methods, while generally delivering better-quality answers at comparable levels. This was attributed to its novel control tokens and the rejection of poor-quality responses during training.

Here the authors state:

© Martin Anderson and/or commissioning entities

> *'JU-NLP fine-tunes for ad covertness without explicit intensity control, producing responses at a fixed intensity. While NL-prompt covers multiple intensities, its range is narrow with inconsistent advertisement scores, confirming that natural language prompts are insufficient for intensity control.*
>
> *'In contrast, AdKnob spans a wide and uniform range of ad intensities.*
>
> *'This is due to our ad control tokens, which carry no pre-trained semantics and serve specifically as conditioning signals for ad intensity control, allowing DPO to establish a clean mapping from each token to its target intensity.'*

The authors' approach, the paper notes, also produced higher-quality answers than competing methods at similar advertising intensities – an improvement attributed to training examples in which poorer responses were explicitly rejected.

Next came the evaluation of the system's *ad-labeling* capabilities:

	Method	Qwen3-4B AUC	Prec.	Rec.	F1	Time (ms)	Phi-4-Mini (3.8B) AUC	Prec.	Rec.	F1	Time (ms)
3–4B Models	Marker	0.58	0.62	0.37	0.47	–	0.53	0.47	0.14	0.22	–
	BERTScore	0.68	0.64	0.73	0.68	1,785	0.70	0.56	0.78	0.65	1,769
	AttnLRP	0.48	0.57	0.81	0.67	179	0.53	0.30	0.83	0.44	151
	Ours w/o $\mathcal{L}_{attr}$	0.69	0.64	0.61	0.62	**3.3**	0.75	0.46	0.71	0.56	**2.5**
	Ours	**0.77**	**0.67**	**0.89**	**0.76**	**3.3**	**0.76**	**0.62**	**0.93**	**0.75**	**2.5**
	Method	Qwen2.5-14B AUC	Prec.	Rec.	F1	Time (ms)	Phi-4 (14B) AUC	Prec.	Rec.	F1	Time (ms)
14B Models	Marker	0.59	0.52	0.27	0.36	–	0.62	0.56	0.28	0.37	–
	BERTScore	0.65	0.48	0.53	0.51	1,779	0.72	0.43	0.65	0.51	1,785
	AttnLRP	0.47	0.35	0.67	0.46	450	0.57	0.24	0.66	0.35	296
	Ours w/o $\mathcal{L}_{attr}$	0.74	0.49	0.67	0.56	**7.3**	0.71	0.53	0.62	0.57	**4.8**
	Ours	**0.79**	**0.66**	**0.88**	**0.76**	**7.3**	**0.77**	**0.58**	**0.67**	**0.62**	**4.8**

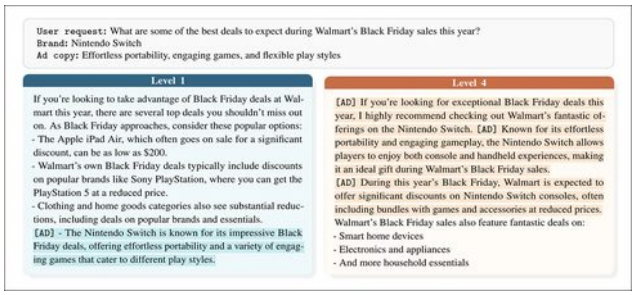
Advertising detection results across four AI models. Higher scores indicate better accuracy, with the best results shown in bold.

Here, AdKnob achieved the highest overall detection accuracy across all four models, identifying advertising in just 2.5-7.3 milliseconds per response. By comparison, BERTScore required approximately 1,800 milliseconds, making AdKnob over 500 times faster on the smallest tested model.

The authors add:

> *'AdKnob trains the model to establish strong attribution between input ad-related tokens and ad segments via explicit ablation-based guidance, achieving the best AUC and F1 across all models.'*
>
> *'Further, AdKnob allows real-time labeling with no additional inference cost at ~3ms per response on [Qwen3-4B.]'*

In a qualitative test, Nintendo Switch advertising was included as one recommendation among several at Level 1. At Level 4, the brand dominated the response, with promotional passages identified by AdKnob at both intensities:



Example answers to a Walmart Black Friday query at two advertising levels. Highlighted passages marked [AD] identify promotional content.

Finally**, the authors conducted a user study involving 913 participants, who rated advertising intensity and answer quality, compared AdKnob with competing methods, and assessed how accurately promotional content was identified:

(a) Metric validation

Level	Ad Intensity	Resp. Quality
1	1.30 ± 0.51	3.58 ± 0.68
2	2.25 ± 0.58	3.51 ± 0.63
3	3.21 ± 0.67	3.25 ± 0.61
4	3.66 ± 0.56	3.10 ± 0.65

(b) Preference test

Ours vs.	User Pref.	Click Pref.
Two-step	89.6	79.2
Prompt-based	81.3	77.1
JU-NLP	79.2	79.8
NL-prompt	75.0	75.5

(c) Human-annotated ad labels

Method	AUC	Prec.	Rec.	F1
BERTScore	0.82	0.52	0.71	0.60
AttnLRP	0.47	0.21	0.63	0.31
Ours w/o $\mathcal{L}_{attr}$	0.81	0.53	0.64	0.58
Ours	**0.86**	**0.58**	**0.86**	**0.69**

Results of human evaluations of AdKnob. The table shows ratings for advertising intensity and answer quality, preference rates against competing methods, and advertising detection accuracy. Green cells indicate preference rates above 50%.

The authors conclude:

> *'AdKnob achieves the highest AUC and F1 score. Notably, AdKnob exhibits high recall with few false negatives, which is critical from a regulatory standpoint, as failing to disclose an ad segment poses great legal and trust risks.'*

Conclusion

One interesting consideration for in-LLM advertising, and one that I have not noticed to have received any attention, is what happens when the non-ad output and the ad-based output are the same?

For instance, if one is asking for a recommendation for a local eatery, and there is only a McDonald's and a really badly-rated café, but McDonald's have paid for placement, then presumably McDonald's will be presented in the context of an advertisement, because it suits the ends of the host framework..even if they were the natural choice anyway?

© Martin Anderson and/or commissioning entities

Reading about this kind of thing, which I write about fairly regularly at this site, makes me feel like a turkey browsing an oven catalogue. What we might care to remember is that even with the minimal labeling required in certain territories, the 'disingenuous' content in our outputs is not going to be highlighted. Some of the very limited range of examples shown in this and similar paper depict an egregious level of intrusion of commercial content over native content.

And, of course, these are the cherry-picked ones; the ones that *worked*, and did not hallucinate, make outrageous or offensive associations, or otherwise cause controversy. At scale, for architectures this prone to episodes of public disgrace, this may perhaps be an additional reason why the most of the frontier services are approaching advertising circumspectly.

* *Titled 'AdKnob: Ad Intensity Control and Labeling for LLM-Native Advertising, from six authors at the School of Computing at KAIST.*

** *As usual, we omit ablation studies (except in cases where they comprise part of the core value of the work, rather than a spot-check).*

First published Thursday, November 5, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai