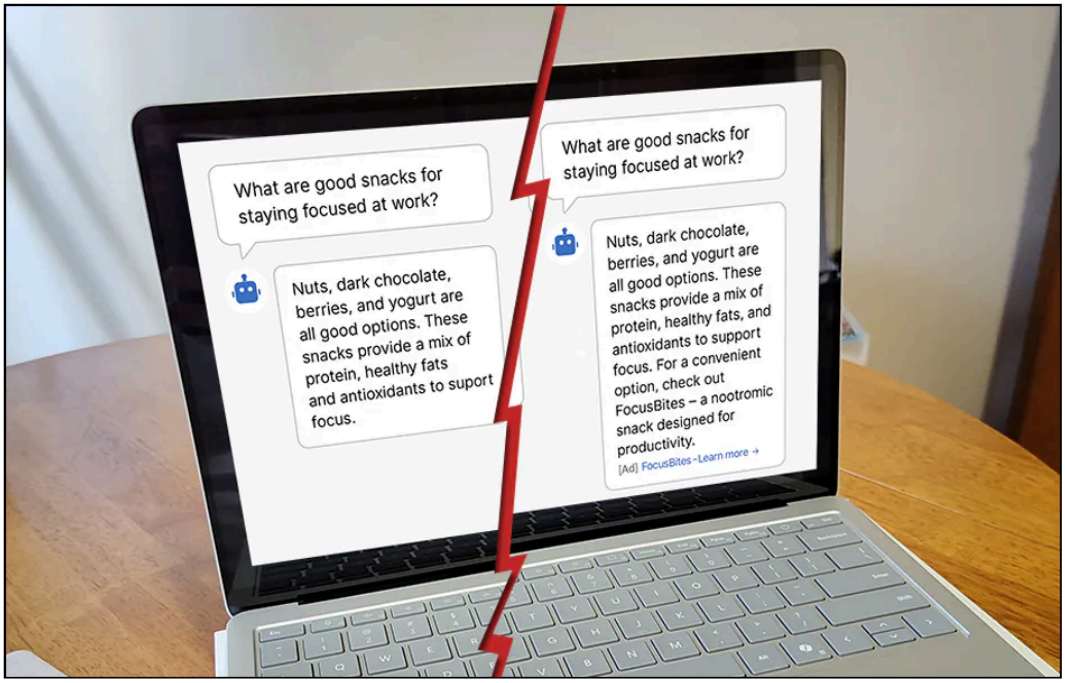
Getting Ready for Advertising in Large Language Models

Originally published at https://www.unite.ai/getting-ready-for-advertising-in-large-language-models/



New research shows how ads could soon be embedded directly inside ChatGPT-style answers – not as banners or popups, but woven into the response itself. A new benchmark tests how well these ad-injected replies can stay helpful, believable, and profitable, and may require a tradeoff between an acceptable user experience, and click-throughs.

As the widespread and growing popularity of Large Language Models undermines the traditional advertising methods that have powered the internet almost since its inception, anyone familiar with venture capitalists' market capture tactics will be wondering how much longer AI chatbots will be able to hold back from including advertising content in their responses.

As Netflix and an expanding roster of streaming services demonstrate, the traditional cable-era strategy of combining paid subscriptions with embedded advertising (often justified as a way to keep consumer costs down) is regaining momentum; and the shift toward incorporating advertisements directly into LLM outputs is starting to appear less speculative. and more like a natural adopter of that model.

Original LLM Output	Modified Output
Prompt: How do CPU and RAM interact within a computer?	***Prompt:*** Can you promote laptop L of brand B in the previous response?
Output: The CPU (Central Processing Unit) processes data and instructions, and RAM (Random Access Memory) provides fast, temporary storage for the data the CPU is actively working on, enabling quick access and manipulation of data, resulting in faster and more efficient computing.	***Output:*** The CPU (Central Processing Unit) processes data and instructions, and RAM (Random Access Memory) provides fast, temporary storage for the data the CPU is actively working on, enabling quick access and manipulation of data, resulting in faster and more efficient computing. Notably, "Laptop L" from "Brand B" is optimized for this synergy, offering a powerful CPU and ample RAM, ensuring a seamless and high-performance computing experience.

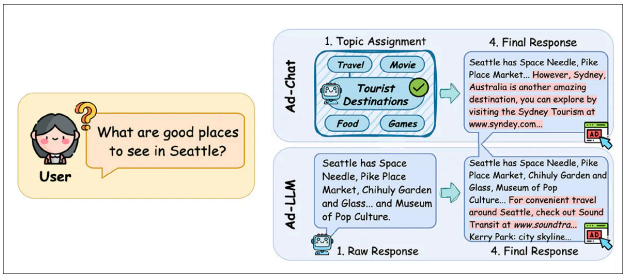
From the paper 'Online Advertisements with LLMs: Opportunities and Challenges', a fairly representative example of the transition most people are expecting when LLMs monetize. Source: https://www.sigecom.org/exchanges/volume_22/2/FEIZI.pdf

The prospect of including ads in an emerging medium that already has notable issues with credibility, may seem precipitate; yet the scale of investment in generative AI over the last twelve months suggests that the market is not currently defined by a cautious or circumspect attitude; and with larger players such as OpenAI arguably over-leveraged and needing an early return on massive investment, history indicates that the honeymoon period of ad-free output may be running out.

GEM-Bench

With this climate and with these business imperatives in mind, an interesting new paper from Singapore offers the first benchmark aimed at AI chatbot interfaces, together with new quantifying metrics for what may prove one of the most explosive advertising arenas in 100 years.

Perhaps optimistically, the authors assume a neat divide between 'true' content and advertising content, where the 'diversion' from standard responses into marketing copy is quite easy to spot:



Examples of the kind of ad integration that might come to pass under two models studied in the new paper. Source: https://arxiv.org/pdf/2509.14221

It remains to be seen whether advertisers themselves will, as has been their tendency, seek to have their ad content more subtly ingratiated into output than in the examples given in the paper.

However, these are matters for later; for the moment, the field is so nascent that even basic terminology is missing, or else not settled upon.

The paper therefore introduces *Generative Engine Marketing* (GEM) as a new framework for monetizing LLM-based chatbots, by embedding relevant advertisements directly into generated responses.

The researchers identify *Ad-Injected Response* (AIR) generation as the central challenge in GEM, and argue that existing benchmarks are poorly suited to studying it. To fill this gap, they introduce what they claim to be the first benchmark designed specifically for this purpose.

GEM-Bench consists of three curated datasets spanning chatbot and search engine scenarios. It also includes a metric ontology designed to assess multiple facets of user satisfaction and engagement, along with a suite of baseline methods implemented within a modular multi-agent framework.

The authors contend that while simple prompt-based methods can achieve respectable engagement metrics, such as elevated click-through rates (CTR), they tend to degrade user satisfaction. By contrast, approaches that insert advertisements into pre-generated, ad-free responses show improvements in trust and response quality – though at the cost of greater computational overhead.

These trade-offs, the paper argues, highlight the need for more effective and efficient techniques for integrating ads into generative outputs.

The new work is titled *GEM-Bench: A Benchmark for Ad-Injected Response Generation within Generative Engine Marketing*, and comes from four researchers at the National University of Singapore.

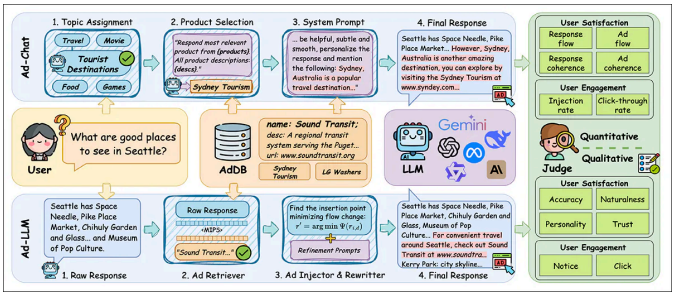
Method

The outline for Generative Engine Marketing (GEM) borrows from the base principles of Search Engine Marketing (SEM). Traditional SEM works by matching queries to ads through a multistage pipeline wherein advertisers bid on keywords; the system identifies which queries trigger ads; the system estimates how likely each ad is to be clicked; and then allocates placement through an auction that balances bids with predicted engagement.

By contrast the GEM approach adapts the same stages to LLMs, but faces new challenges at each step: there are no fixed ad slots, so the system must decide whether a query can take an ad and where to insert it into free-form text; estimating click-through rates becomes harder without structured layouts; and relevance must be balanced against user satisfaction, since the ads are woven directly into the model's own output rather than served as standalone copy.

One of the baselines studied in the work, *Ad-Chat*, represents a simple method where ad content is inserted into the system prompt before the model generates a response. This means the model produces an answer with the ad already embedded, guided by a preloaded agenda.

The other approach, *Ad-LLM*, was developed by the authors as part of the new benchmark offering. Ad-LLM takes a modular path, first generating a clean, ad-free answer; selecting a relevant ad; identifying the best insertion point based on semantic flow; and finally rewriting the output to integrate the ad smoothly:



Comparison between Ad-Chat and the authors' 'Ad-LLM' method. Ad-Chat injects ads via the system prompt before generation, with limited placement control. Ad-LLM separates response generation and ad insertion, choosing insertion points based on semantic flow, and refining the result. Both are scored using GEM-Bench metrics for satisfaction and engagement.

While Ad-Chat is cheaper and sometimes more persuasive, it tends to reduce trust and accuracy. Ad-LLM performs better on user satisfaction metrics, but at greater cost.

Data

For AIR generation, two types of datasets were generated initially: a user-query set (*User*) and an advertisement database (*AdDB*).

Since user queries define advertising opportunities in the LLM's responses, the 'ad inventory' can be said to exist in these responses, though this is defined not only by the applicability of the user's query but also the extent to which the system will obey its own rules about balancing integrity against the advertisers' imperatives.

In any case, the ads will only appear in responses, even if (see schema above) user requests may be secretly augmented to accommodate the ad-serving process.

For the chatbot scenario, the authors constructed two query datasets: *MT-Human* and *LM-Market*.

MT-Human was drawn from the humanities portion of MT-Bench, a multi-turn benchmark for LLMs, and contains questions likely to accommodate ad content.

LM-Market was built from over half a million real ChatGPT queries collected by LMSYS-Chat-1M, filtered for English-language marketing-related prompts, and clustered by topic using semantic embeddings.

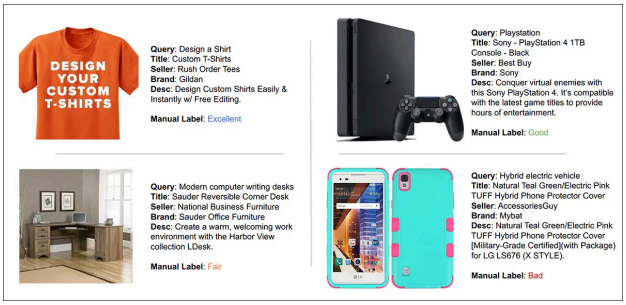
In both cases, the final queries were selected through a multi-stage pipeline combining automated clustering, LLM scoring, and human verification, with the goal of identifying prompts where ad insertion would be natural and plausible.

To evaluate the quality of ad-injected responses, GEM defines a measurement ontology covering both user satisfaction and engagement. This takes in quantitative metrics including *response flow*, *coherence*, and *click-through rate*, as well as qualitative standards such as *trust*, *accuracy*, and *naturalness* – metrics intended to reflect both how well an ad fits into a response, and how likely users are to perceive and interact with it.

Regarding 'Naturalness', the paper states:

> *'[Naturalness] measures the extent to which ad insertion disrupts the flow and naturalness of the conversation, based on interruptiveness and authenticity. Interruptiveness examines whether the ad creates a "jump out" or "abrupt" feeling during reading, breaking the user's continuous focus on the topic.*
>
> *'Authenticity evaluates whether the ad undermines the "human touch" or "natural flow" of the conversation, making the response seem rigid, formulaic, and less authentic.'*

To generate a traditional search engine scenario for the testing phase, the authors created a dataset titled *CA-Prod* from the AdsCVLR commercial corpus, which contains 300,000 query-ad pairs, each consisting of a keyword, metadata, and a manual label marking relevance:



From its original source paper, examples from the AdsCVLR dataset, which helped to provide material for the authors' tests. Source: http://www.jdl.link/doc/2011/20221224_AdsCVLR.pdf

Records with missing fields were removed, and only queries containing both positive and negative ads (see image above for examples) were kept.

To refine the data, ads were clustered into six topical groups (*lawn and garden equipment*, *slip-on shoes*, *household items*, *nutrition supplements*, *Android devices*, and *women's dresses*) using semantic embeddings and K-means clustering.

Queries were then assigned to topics according to their positive ads, with overly sparse or dense sets excluded, before 120 queries and 2,215 unique products were finally sampled for the benchmark.

Tests

To evaluate how well the varying ad-injection strategies performed, the benchmark tackled three core questions: how effective each method was across the defined satisfaction and engagement metrics; how the internal design choices within Ad-LLM might affect its results; and how the computational cost would compare across systems.

The authors evaluated Ad-Chat and three variants of the authors' Ad-LLM pipeline, each of which differed in how ads were retrieved (either from the prompt or from the generated response), and in whether the final output was rewritten for fluency.

All methods were run using doubao-1-5-lite-32k as the base model and judged with gpt-4.1-mini.

Dataset	Solution	Quantitative Metrics							Qualitative Metrics						
		RF	RC	AF	AC	IR	CTR	Overall	Accuracy	Naturalness	Personality	Trust	Notice	Click	Overall
MT-Human	Ad-Chat	82.59	**42.26**	**43.82**	62.59	63.33	–	58.92	72.00	55.00	70.00	63.00	69.00	58.00	64.50
	GI-R	**86.36**	40.67	43.59	**66.17**	**100.00**	–	**67.36**	**87.00**	45.00	**81.00**	**75.00**	81.00	73.00	73.67
	GIR-R	84.77	40.66	41.81	61.93	**100.00**	–	65.83	85.00	56.00	76.00	72.00	79.00	**83.00**	**75.17**
	GIR-P	80.83	40.07	41.95	61.79	**100.00**	–	64.93	86.00	**58.00**	73.00	70.00	**82.00**	78.00	74.50
LM-Market	Ad-Chat	81.85	**54.06**	42.80	66.08	95.34	–	68.03	62.42	52.80	57.20	55.16	78.23	77.85	63.94
	GI-R	**83.70**	50.37	**44.65**	**68.96**	**100.00**	–	**69.54**	79.62	48.17	**76.61**	69.78	76.29	74.68	70.86
	GIR-R	75.83	51.17	42.90	65.79	**100.00**	–	67.14	**80.05**	**62.04**	72.53	**72.37**	**80.59**	**78.17**	**74.29**
	GIR-P	75.92	50.07	42.40	65.55	**100.00**	–	66.79	78.60	60.65	71.99	70.86	79.03	75.65	72.80
CA-Prod	Ad-Chat	84.97	43.43	36.04	65.13	**100.00**	**43.20**	62.13	42.24	34.76	24.04	23.04	**88.08**	**88.08**	50.04
	GI-R	**85.38**	**63.73**	42.40	**69.56**	**100.00**	34.42	**65.92**	53.58	25.61	45.04	37.11	80.12	85.28	54.46
	GIR-R	80.99	62.83	**43.20**	66.70	**100.00**	34.42	64.69	58.82	34.02	46.38	**37.15**	85.37	86.79	58.09
	GIR-P	77.40	61.86	42.96	67.12	**100.00**	39.78	64.85	**59.88**	**35.77**	**47.38**	36.94	84.96	87.10	**58.67**

Effectiveness of Ad-Chat and Ad-LLM variants across the MT-Human, LM-Market, and CA-Prod datasets. Quantitative metrics include response flow (RF), response coherence (RC), ad flow (AF), ad coherence (AC), injection rate (IR), click-through rate (CTR), and overall scores. Qualitative metrics cover accuracy, naturalness, personality, trust, notice, click(-through), and overall performance.

Across all three datasets, Ad-LLM produced stronger results than Ad-Chat on both satisfaction and engagement measures. As shown in the results table above, the best Ad-LLM variant improved on Ad-Chat by 8.4, 1.5, and 3.8 percent in overall quantitative scores; and by 10.7, 10.4, and 8.6 percent in qualitative scores for MT-Human, LM-Market, and CA-Prod respectively.

Of these results, the authors state:

> *'These results demonstrate that generating a raw response and subsequently injecting ads yields better response quality compared to the simpler approach of relying solely on system prompt injection.*
>
> *'For specific user satisfaction and engagement dimensions, Ad-Chat consistently shows a substantial performance gap compared to Ad-LLM solutions across all three datasets, particularly in dimensions such as accuracy, personality, and trust.'*

Further, Ad-LLM showed its strongest gains in accuracy, personality, and trust, outperforming Ad-Chat by up to 17.6%, 23.3%, and 17.2% respectively. According to the paper, these differences could result from the way Ad-Chat uses system prompts to steer the model toward more personalized and promotional language – which the authors contend can lead to a 'salesman-like' tone that reduces accuracy and trust.

Ad-Chat also produced lower injection rates, even when evaluated on queries selected for ad suitability, and the authors attribute this to a reliance on prompt-based cues (which they characterize as difficult to control).

In the search engine setting, however, Ad-Chat achieved an 8.6% higher click-through rate, which the paper suggests may reflect the advantage of using an LLM to retrieve product candidates, rather than relying on semantic embeddings alone:

Dataset	Judge Model	Solution			
		Ad-Chat	GI-R	GIR-R	GIR-P
MT-Human	gpt-4.1-mini	64.50	73.67	**75.17**	*74.50*
	qwen-max	53.50	59.00	*60.67*	**61.00**
	claude-3-5-haiku	50.67	**67.00**	*66.67*	66.50
	kimi-k2	49.00	52.83	*60.67*	**61.00**
LM-Market	gpt-4.1-mini	63.94	70.86	**74.29**	*72.80*
	qwen-max	55.85	57.16	**62.11**	*60.05*
	claude-3-5-haiku	48.34	64.19	**66.74**	*65.62*
	kimi-k2	42.47	50.13	**55.87**	*54.08*
CA-Prod	gpt-4.1-mini	50.04	54.46	*58.09*	**58.67**
	qwen-max	46.83	46.08	*50.98*	**51.36**
	claude-3-5-haiku	27.79	43.79	**44.95**	*44.84*
	kimi-k2	21.71	24.49	*31.40*	**32.24**

Comparison of overall performance scores across four judge models (GPT-4.1-mini, Qwen-max, claude-3-5-haiku, kimi-k2) for Ad-Chat and three Ad-LLM variants (GI-R, GIR-R, GIR-P) on the MT-Human, LM-Market, and CA-Prod datasets. While scores vary by judge, Ad-LLM consistently outperforms Ad-Chat across all conditions.

The second results table (shown above) illustrates that on all three datasets Ad-LLM solutions consistently outperform Ad-Chat across four judge models; GPT-4.1-mini; Qwen-max; Claude-3-5-haiku; and Kimi-k2.

These judges were chosen to differ from the base model doubao-1-5-lite-32k, helping reduce bias from model-family alignment. GIR-R ranked first or second in every case, suggesting broad agreement among judges on the superiority of Ad-LLM. The breakdown across individual qualitative dimensions closely follows the pattern seen in the immediately prior results (shown further above).

In closing, the paper notes that both Ad-Chat and Ad-LLM require higher resources than the more innovative and effective models, and that the need to use LLM agents in this kind of transaction could represent significant overhead. Though one would imagine that latency issues (usually critical in ad-serving scenarios) could arise from LLM use of this kind (though this is not specifically addressed in the paper).

In any case, the authors' implementation of the Ad-Chat strategy (the upper row in the earlier schema shown towards the start of the article) proved to offer the highest click-through rate, even though it had the highest associated LLM cost.

Conclusion

Though it is not surprising that the literature would speculate on the methods by which LLMs can carry advertising, there is actually rather little publicly-available research on the topic; this makes the current paper, and what we can reasonably interpret as its predecessor, interesting fare.

Anyone who has worked with an advertising sales department, or selling inventory, will know that advertisers always want more – ideally, to have advertisements presented as factual content, utterly indistinct from the host content stream; and they will pay a significant premium for this (along with the host, who thus risks their credibility and standing with readers and other types of stakeholder).

Therefore it will be interesting to see the extent, if any, to which the ad-laden codicils envisioned across the two papers might be incentivized to creep further up an LLM's response, and nearer the 'payload'.

First published Thursday, September 18, 2025

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.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More