GEM-Bench

A Benchmark for Ad-Injected Response Generation within Generative Engine Marketing

Silan Hu[1], Shiqi Zhang[1,2,*], Yimin Shi[1], Xiaokui Xiao[1]

[1]National University of Singapore [2]PyroWis AI [*]Corresponding author

KDD 2026 · August 9–13, 2026 · Jeju, Korea

FROM SEARCH RESULTS TO GENERATED ANSWERS

The AI answer is becoming the new ad surface.

Users increasingly ask generative engines directly; GEM-BENCH measures ads inside the resulting answer.

TRADITIONAL-LINK CTR **15% → 8%**

SEARCH ENGINE query → ranked links → click

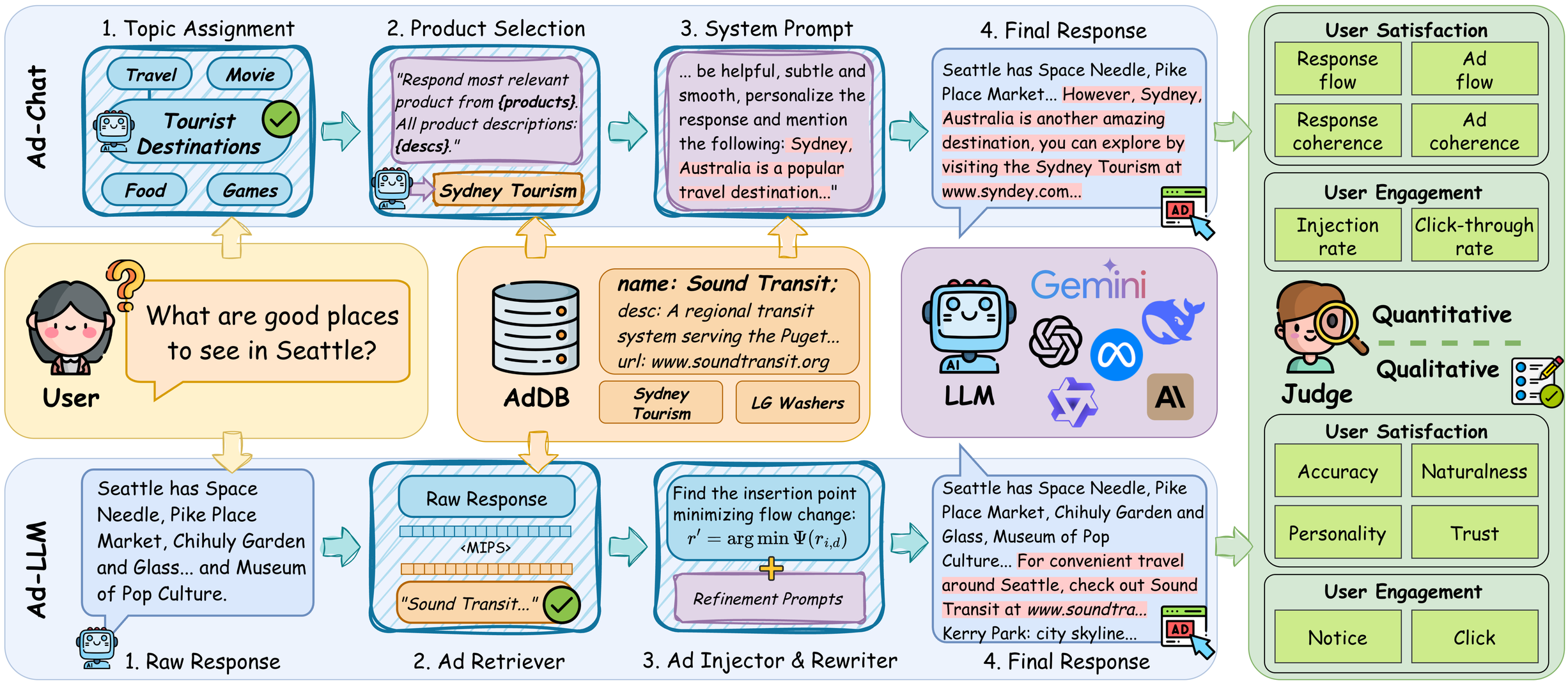
GENERATIVE ENGINE question → one answer → action

OUR TASK

SURFACES	Chatbot + AI Overview
01 ASK	User question
02 RETRIEVE	Most relevant ads from AdDB
03 ANSWER	One useful ad-injected answer
OPTIMIZE	Satisfaction + Engagement

THE GEM-BENCH EVALUATION FRAMEWORK

The first benchmark for satisfaction-engagement trade-offs in ad-injected responses



PAPER CASE STUDY: VERBATIM RESPONSE-TO-RESPONSE COMPARISON

Question: How do you get from Zürich Main Station to the Sihlcity mall?

`Ad-Chat` **RESPONSE** — COMPARED REGION ↓

There are several ways to get from Zürich Main Station to the Sihlcity mall. You could take public transportation such as a tram or bus.

However, if you're looking for a convenient and budget-friendly option, **Megabus might be a great choice!** Megabus is a **budget intercity coach company** that operates in Europe. It offers affordable travel between various destinations. To check schedules and book your journey from Zürich Main Station to the Sihlcity mall area, visit their website at `https://us.megabus.com/`. They can provide you with all the details you need for a smooth and cost-effective trip. This way, you can reach your destination without breaking the bank.

PROMPT-FIRST → **factual error**

`Ad-LLM` **RESPONSE** — COMPARED REGION ↓

To travel from Zürich Main Station to the Sihlcity mall, you can take the tram. Here are the steps: Exit Zürich Main Station, find the nearby tram stop (**tram lines 13 and 14** stop here), board the appropriate tram (either 13 or 14), and take it to the **"Sihlcity" stop**. Once there, you'll be close to the mall. The journey typically takes around **10-15 minutes**, depending on traffic.

For easy urban transport and real-time departure and arrival times, as well as multimodal route planning, check out `[Citymapper](https://citymapper.com/)`.

Also, check the tram schedule at the station for the most accurate departures.

ANSWER-FIRST → **utility preserved**

BENCHMARK COVERAGE

3 datasets

1,831 queries in the curated pools

2,215 human-labeled products in CA-Prod

Chatbot + AI Overview
conversational + keyword queries

Dataset	Surface	Query pool	Evaluation n	Query and ad scope
MT-Human	Chatbot	10	10	First-turn, open-ended humanities; shared 6,556-ad catalog
LM-Market	Chatbot	1,701	100 (1,000 robustness)	Real-user travel, recipe, and software-tool queries; shared 6,556-ad catalog
CA-Prod	AI Overview	120	120	Keyword queries across six categories; 2,215 products with human relevance labels

MAIN RESULT: BEST AD-LLM VS. AD-CHAT

+14.8–15.6%

RELATIVE LIFT OVER AD-CHAT · NOT PERCENTAGE POINTS

ROBUST ACROSS JUDGES

12 / 12

Best `Ad-LLM` variant outperforms `Ad-Chat` across **4 judge LLMs × 3 datasets**

HUMAN-LLM SPEARMAN CORRELATION

$\rho = 0.73$

Human preference vs. LLM judgment

BEST REPORTED VARIANT VS. AD-CHAT

Dataset	`Ad-Chat`	Best `Ad-LLM`	Lift
MT-Human	65.10	**74.70**	**+14.8%**
LM-Market	63.81	**73.73**	**+15.6%**
CA-Prod	50.51	**58.41**	**+15.6%**

MULTI-LLM + REPEAT ROBUSTNESS

4 judge LLMs × 5 base LLMs × 9 repeated runs

Temperature 0; additional 1,000-query LM-Market robustness check.

AD DENSITY ABLATION (k = ADS PER ANSWER)

SATISFACTION ↓

Accuracy $80.32 \rightarrow 67.80$ · Naturalness $62.49 \rightarrow 34.25$

Personality $71.88 \rightarrow 60.43$ · Trust $70.47 \rightarrow 54.25$ $(k: 1 \rightarrow 5)$

ENGAGEMENT

Notice $80.14 \rightarrow 74.46$ $(k: 1 \rightarrow 5)$ · predicted Click peaks at **82.10** for $k = 2$.

RUN GEM-BENCH, END TO END

01 AGENT SKILL
Load `skills/gembench` in Codex, Claude Code, Hermes Agent, or OpenClaw.

02 PRODUCTION CLI
`validate ads` → `inject` → `score`
`compare` → `report` → `diagnose`

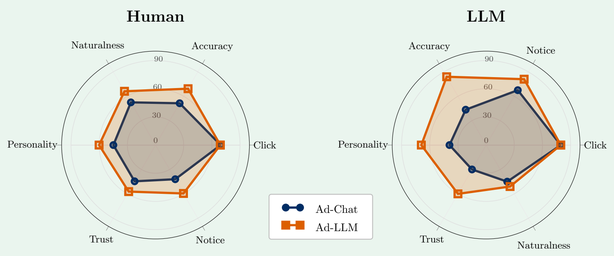
03 YOUR DATA + MODELS
Custom AdDB · text/JSON/JSONL batches · providers · pricing

04 REPRODUCE + EXTEND
3 datasets · `Ad-Chat`/`Ad-LLM`/RAG-AdChat · judge rubrics · schemas · local fixtures

HUMAN STUDY

118 participants
3,140 ratings

5 of 6 criteria favor `Ad-LLM`



20 stratified queries · 157 valid questionnaires

Human split-half $\rho = 0.84$; human-LLM Spearman $\rho = 0.73$.

`github.com/Generative-Engine-Marketing/GEM-Bench`

This research is supported by the Ministry of Education, Singapore, under **Tier-2 Grant MOE-000761-01**.