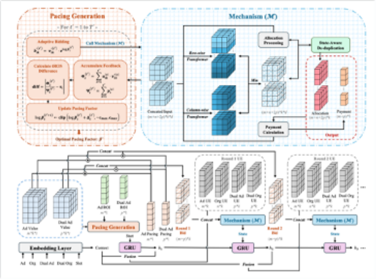
Generation as Allocation:

Auctions and Mechanism Design for LLM Native Advertising

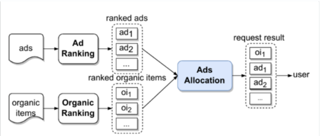
Qi Qi
Gaoling School of Artificial Intelligence
Renmin University of China

Learning mechanisms for ads and organic content

Jointly decide **how ads and organic content form a page**.



Mechanism Design Foundations [TCS23] • Deduplication and IC[AAAI25, Oral]
Externality [WWW25] • Autobidding [KDD26]

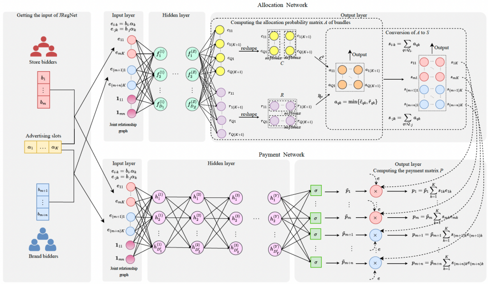


- One pool: ads + organic content
- Long-term page value
- Incentives, budgets, and ROI

AI learns richer ranking and auction rules within the page paradigm.

Joint auctions for shared commercial interests

Stores and brands have **related but distinct interests**.



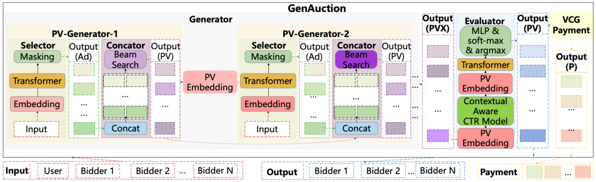
JRNet [KDD24 Oral] • BundleNet [ICML25]
HRegNet [SIGIR25] • JTransNet [KDD26 Oral]

- **Brands:** reach
- **Stores:** sales
- **Platform:** joint rules

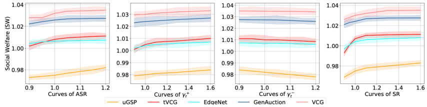
AI learns deployable combinatorial auctions subject to DSIC and IR.

GenAuction generates page-level allocations

Context-dependent value calls for **page-level generation**.



- Generate from ads and context
- Evaluate welfare and experience
- CMDP enforces constraints



Performance across constraints

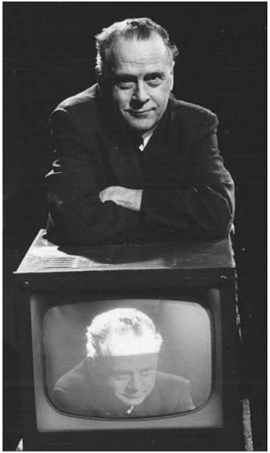
AI moves mechanism design toward generative allocation.

MARSHALL MCLUHAN 1964

“The medium is the message.”

A medium shapes how information is organized.

Advertising inherits that structure.



Marshall McLuhan

Token-Level Advertising

arXiv:2608.27382v2


Hanbing Liu
RUC

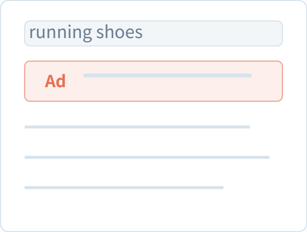

Bowei Zhang
RUC


Changyuan Yu
Baidu


Yinyu Ye
Stanford

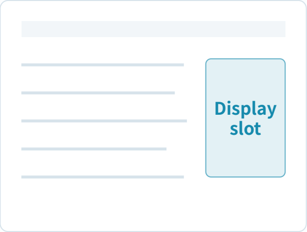

Qi Qi
RUC

Traditional advertising starts with predefined slots



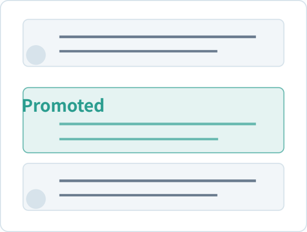
Sponsored search

Fixed positions in ranked lists



Display advertising

Reserved regions on a page



Promoted feed

Preset positions in a feed

Organize content → **Define inventory** → **Auction slots**

Thirty years of advertising built around slots

China, 2025: ad business revenue **USD 307.5B**; internet ad publishing **USD 187.5B**

1990s

1994 • Banner
HotWired

1996 • Ad Server
DoubleClick

Serving and tracking ad slots

2000s

2000 • Search Ads
Google AdWords

2005 • Exchange

Matching intent, pooling inventory

2010s

2010 • DSP / SSP / DMP

2012 • Programmatic
Autobidding

2014 • Video / Feed

AI interfaces

2020s • AI answers

2026 • LLM Ads

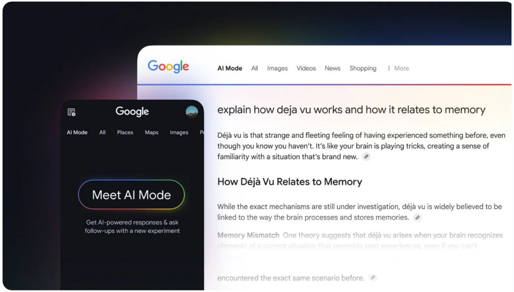
The medium generates content

The interface changes, but the abstraction persists:
define slots, then allocate.

Information interfaces now generate answers

Baidu AI Answers

Synthesized answers within search

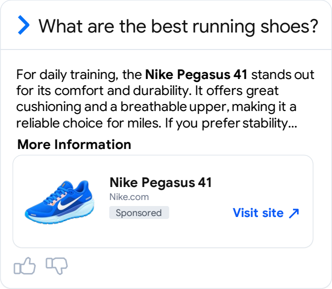


Google AI Search

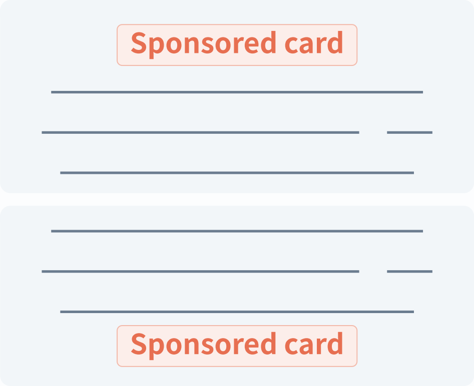
Generated answers at the search interface

From ranking existing content to
generating answers around user intent.

A natural first step: ads around the answer

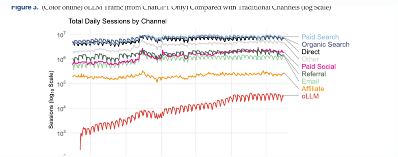


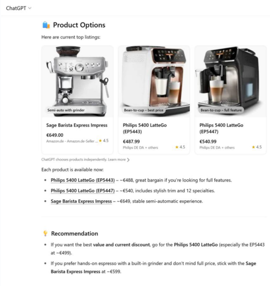
Sponsored card below the answer



Generate the answer first.
Advertising inventory still exists separately.

External cards remain a small referral channel





$< 0.2\%$

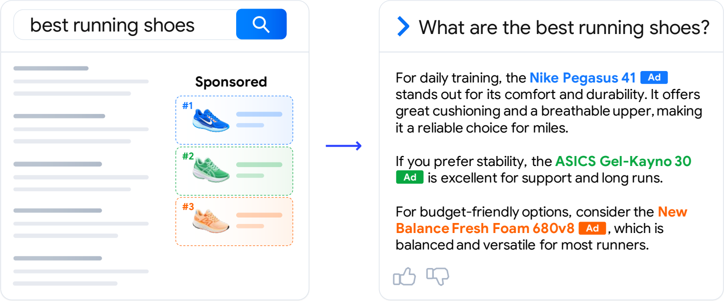
The share of all e-commerce visits in the sample coming from ChatGPT referrals

$\approx 1/200$

of the traffic from Google organic search

Empirical data show **significantly lower click-through rates** for **cards in ChatGPT answers linking to merchants** than for traditional advertising formats.

Native ad opportunities emerge during generation

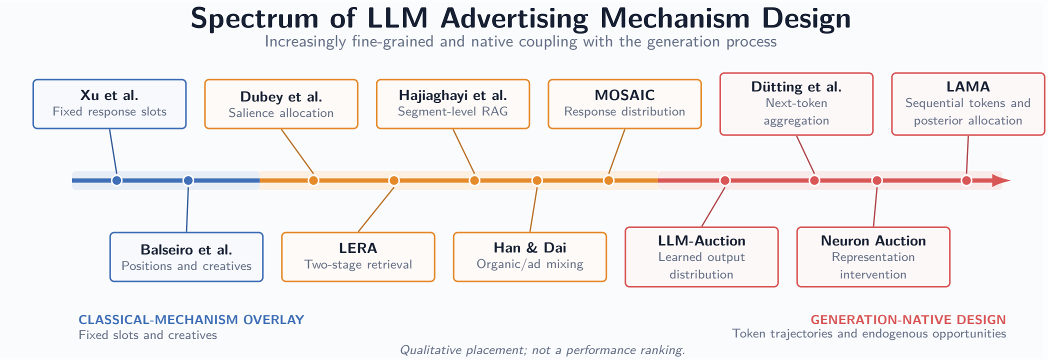


Generation does more than fill an opportunity—

it can help create one.

Content generation = **Opportunity formation** + **Resource allocation**

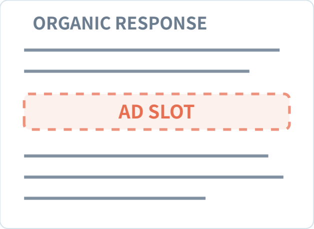
Ad opportunities are moving inside the generation process



From fixed slots and complete responses to
positions, distributions, latent representations, and trajectories.

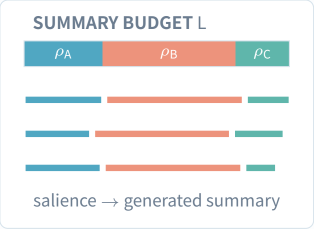
Representative mechanisms

Xu et al.



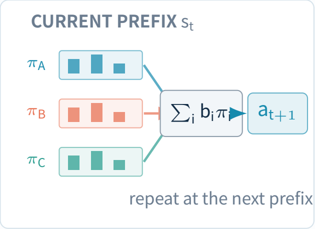
Generate organic text first
Then insert creatives between sentences

Dubey et al.



Bidding determines salience, LLM generates by **relative advertiser length**

Dütting et al.



Aggregate distribution preferences, **choose each token myopically.**

Finer allocation objects: **slot** $\to$ **length share** $\to$ **next-token distribution**

Fully native: advertisers participate in every token decision

For daily training, the Nike Pegasus 41 stands out.
For stability, consider the ASICS Gel–Kayano 30.

Each prefix

Report local information to shape the next-token distribution

Each token

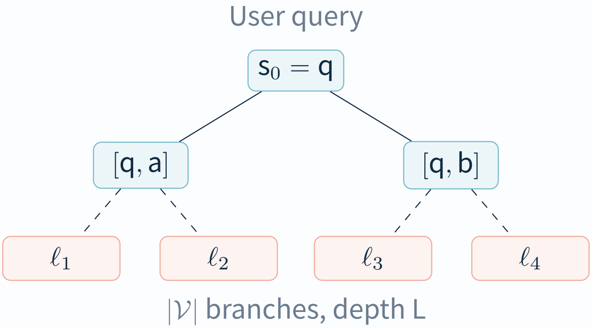
Update the posterior over the eventual winner

At termination

Allocate to the winner and settle trajectory payments

Advertisers do not wait for allocation.
They participate throughout generation.

Every answer is a path through a generation tree



Two path-dependent quantities

- Naturalness of the answer
- Terminal advertiser value $r_i(\ell)$

One early token can reshape every later ad opportunity.

Model: impression value of generated content

Generation

$$\begin{aligned} s_{t+1} &= [s_t, a_t], \\ a_t &\sim \pi(\cdot \mid s_t), \\ \ell(\tau) &= s_T. \end{aligned}$$

π_{ref}: user-centered, ad-free reference policy

Allocation and value

$$\begin{aligned} r_i : \mathcal{S}_T &\to \mathbb{R}_{\geq 0}, \\ \lambda(\ell) &\in \Delta(\mathcal{N}), \\ \text{gross value}_i &= \lambda_i(\ell) r_i(\ell). \end{aligned}$$

Single winner: generation shapes the opportunity. One advertiser receives it

Generation policy π + **Terminal allocation λ**

Objective: balancing ad value and natural generation

$$\mathrm{SW}_\beta(\pi, \lambda) = \mathbb{E}_{\tau\sim\pi}\left[\sum_i \lambda_i(\ell) r_i(\ell) - \beta \sum_t \log \frac{\pi(a_t \mid s_t)}{\pi_{\mathrm{ref}}(a_t \mid s_t)}\right]$$

Advertiser value

- Who receives the opportunity?
- How valuable is this answer?

User-quality anchor

- Trajectory-level KL cost
- Larger β: closer to organic generation

Each advertiser solves a soft control problem

$$V_i^*(s) = \max_{\pi} \mathbb{E}\left[r_i(\ell) - \beta \sum_{t \geq s} \log \frac{\pi(a_t \mid s_t)}{\pi_{\mathrm{ref}}(a_t \mid s_t)} \right]$$

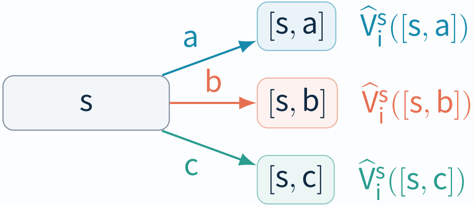
Soft Bellman recursion

$$V_i^*(s) = \beta \log \sum_a \pi_{\mathrm{ref}}(a \mid s) e^{V_i^*([s,a])/\beta}.$$

Advertiser’ s optimal policy

$$\pi_i^*(a \mid s) \propto \pi_{\mathrm{ref}}(a \mid s) e^{V_i^*([s,a])/\beta}.$$

Online reports need only next-step continuation values



Report a child-value vector at the current prefix.

Imply the local optimal policy

$$\widehat{\pi}_i(a \mid s) \propto \pi_{\text{ref}}(a \mid s) e^{\widehat{V}_i^s([s,a])/\beta}.$$

A latent advertiser guides each token

Posterior $\rho(s)$

Sample $I \sim \rho(s)$

Sample $a \sim \widehat{\pi}_I(\cdot \mid s)$

Marginally, a mixture policy

$$x(a \mid s) = \sum_i \rho_i(s)\widehat{\pi}_i(a \mid s).$$

I is the latent source of this token, not necessarily the final winner.

Each token provides evidence about the winner

$$\rho_i([s, a]) = \frac{\rho_i(s)\widehat{\pi}_i(a \mid s)}{\sum_j \rho_j(s)\widehat{\pi}_j(a \mid s)}$$

Before generation

$\rho(s)$

$\rightarrow$

After observing token a

$\rho([s, a])$

If advertiser i is more likely to generate a, its posterior weight rises.

The terminal posterior becomes the allocation

$$\lambda(s_T) = \rho(s_T) \in \Delta(\mathcal{N})$$

Full token trajectory → Final posterior $\rho(s_T)$ → Ad opportunity allocation

The score is not computed after the response.
It accumulates during generation.

Four steps: report, generate, update, settle

1 Report
child-values reports
Check consistency

→

2 Generate
Sample advertiser
Sample token

→

3 Update
Update values
Update posterior

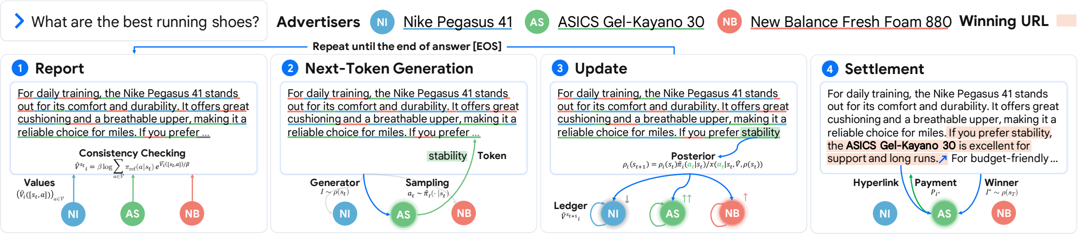
→

4 Settle
Sample winner
Charge payment

LAMA = Latent Advertiser Mixture Auction

"Latent Mixture" comes from **Bayesian inference** during generation: updating the latent-advertiser posterior token by token.

Mechanism workflow



Consistency Checking? Why not let advertisers report child values however they wish at each state?

Bellman consistency checking

$$\beta \log \sum_{a} \pi_{\mathrm{ref}}(a \mid s) e^{\widehat{V}_i^s([s,a])/\beta} = \widehat{V}_i(s)$$

Initial state

- Freely report the first child values
- Record root value $\widehat{V}_i(q)$

Later states

- New reports must aggregate to the parent value in the ledger
- Change allocation within the subtree, but do not rewrite the past

Markov DSIC: truthfulness at every reached prefix

Theorem

Within the Bellman-consistent message space, truthful reports

$$\left(V_i^*([s,a])\right)_{a\in\mathcal{V}}$$

are optimal against any opponents, at every reachable prefix, among feasible continuation deviations.

Proof outline

1. Bellman consistency propagates leaf values $z_i(s_T) = e^{\widehat{V}_i(s_T)/\beta}$ back to the root.
2. Joint generation–allocation probabilities are gradients of a convex log-sum-exp potential.
3. The potential payment

$$P_i^{\mathrm{frac}}(s_T) = \rho_i(s_T)\widehat{V}_i(s_T) - \beta \log \sum_{j\in\mathcal{N}} e^{\widehat{V}_j(q)/\beta} + \beta \log\left(1 + \sum_{j\neq i} e^{\widehat{V}_j(q)/\beta}\right)$$

makes the direct mechanism truthful.

4. Every online deviation extends to a globally consistent direct report.

Settlement

Sample one terminal winner

$$I^* \sim \rho(s_T).$$

$$P_i^{\mathrm{win}} = \begin{cases} P_i^{\mathrm{frac}} / \rho_i(s_T), & I^* = i, \\ 0, & I^* \neq i. \end{cases}$$

Under exact reports

- Non-winners pay zero
- Winner utility is nonnegative
- Expected revenue is nonnegative for a fixed query and candidate set

Outcome IR

Expected weak budget balance

LAMA optimizes a doubly regularized objective

$$\mathbb{E}_{\tau\sim \mathrm{x}}\left[\sum_{\mathrm{i}} \lambda_{\mathrm{i}}(\tau)\mathrm{r}_{\mathrm{i}}(\ell) + \beta\mathcal{H}(\lambda(\tau)) - \beta\sum_{\mathrm{t}} \log \frac{\mathrm{x}(\mathrm{a}_{\mathrm{t}} \mid \mathrm{s}_{\mathrm{t}})}{\pi_{\mathrm{ref}}(\mathrm{a}_{\mathrm{t}} \mid \mathrm{s}_{\mathrm{t}})}\right]$$

Soft generation

- Randomize over answer trajectories
- KL anchors the reference model

Soft allocation

- Randomize the terminal winner
- Entropy discourages early lock-in

Efficiency guarantee

$$0 \le \sup_{\pi,\lambda} \mathsf{SW}_\beta(\pi,\lambda) - \mathsf{SW}_\beta(\mathsf{x}^\beta,\lambda^\beta) \le \beta \log \mathsf{n}.$$

Allocation entropy causes the gap

Entropy on the simplex is at most $\log n$, bounding the welfare cost by $\beta \log n$.

Only the advertiser count matters

The bound is independent of vocabulary size, answer length, and generation-tree depth.

Exact continuation values are hard to obtain

What the theory needs

Exact continuation values for every advertiser and every next token

$$\left(V_i^*([s, a])\right)_{a \in \mathcal{V}}$$

Practical constraints

- The tree is too large to expand
- Advertisers lack data and compute
- Serve only the current prefix

Our solution:
A platform-side **automated reporting service**, analogous to autobidding.

Value decomposition: root value + advantage function

Local differences

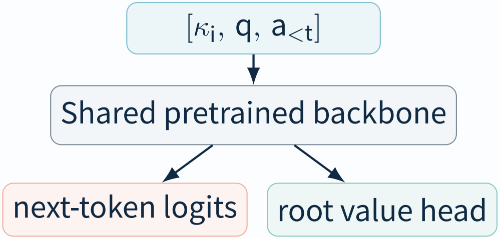
$$A_i^*(s,a) = V_i^*([s,a]) - V_i^*(s) = \beta \log \frac{\pi_i^*(a \mid s)}{\pi_{\mathrm{ref}}(a \mid s)}.$$

Pathwise reconstruction

$$V_i^*(s_t) = V_i^*(q) + \sum_{k<t} A_i^*(s_k, a_k).$$

Learn a token policy for **relative changes**,
and a regression head for the **absolute anchor**.

Only one shared model needs to be deployed



Advertiser differences enter only through κ_i

- Ad information: campaign / creative / landing page
- Shared parameters reduce training and serving costs
- Root residual anchors $V_i^*(q)$

At a population optimum with full comparison support:
exact recovery of policy, advantage, root, and reachable-state values.

Simulation experiments

Setup

- Webis Generated Native Ads 2024
- Workout / Vacation / Car
- Three heterogeneous advertisers per vertical
- Qwen3-14B + LoRA report models
- Open-source model predicts impression value

Baselines

- Before / After $\times$ Original / Edit / Policy
- Response-level aggregation: MOSAIC

Four perspectives

Welfare
Revenue
Advertiser Value
User Quality

Main results

Method	Welfare	Revenue	Quality	Ad Value
Before–Reference	0.2945	0.0000	52.94	0.2945
Before–Policy	0.4562	0.0000	65.16	0.8225
After–Edit	0.1772	0.2490	66.48	0.4046
After–Policy	0.5080	0.7501	65.59	0.8253
MOSAIC	0.4390	0.5274	60.41	0.5550
LAMA	**0.5205**	**0.8305**	**66.52**	**0.8568**

Relative to After–Policy: revenue $+10.7\%$, with further gains in welfare and advertiser value.

Means across three verticals.
The paper reports 95% bootstrap error.

Case study

Field	Content
Vertical	Car
Query	cars for sale near me
LAMA	If you're looking for cars for sale near you, checking with local dealers is the best way to find current inventory. **@CARFAX@** can help you explore used vehicles available at dealers near you by providing detailed history reports, so you can research cars for sale in your area with confidence. Local dealers typically offer in-person visits, test drives, and financing options, making it easier to compare models and secure a deal. To locate nearby dealers with suitable options, you can use your device's map service or the search function on automotive websites. Be sure to call ahead to confirm availability and set up a visit, as popular models may sell quickly. If you're interested in a specific make or model, dealers often have multiple units on-site or can order one for you. Used cars for sale near you usually come with a range of options, including different years, trims, and mileage, so you can find a fit for your needs and budget. @Kelley Blue Book@ is another resource that can help you understand fair pricing for used vehicles at dealers near you. By combining these tools with direct contact to local dealers, you can streamline your search for the right car in your area.
Winner	CARFAX

Table 5: Case Study 3: Car vertical. In the generated content, the winning advertiser is shown in bold, and the highlighted sentence is the span allocated to the hyperlink.

Complementary advertisers can co-create an answer

Offset	λ	Content
-0.2	0.007	Used cars are a popular choice for buyers seeking reliable transportation at a lower cost, offering a range of models, years, and conditions. **@CARFAX@** is a trusted resource that provides detailed insights into a used car's history, including accident reports, service records, and ownership changes, which can be especially helpful when you're considering a specific used car. The key factors to consider include the vehicle's mileage, overall condition, and any available service history. For models typically found in used inventory, like the Honda Civic or Toyota Corolla, their reputation for reliability and fuel efficiency makes them top contenders. Buyers should also pay attention to the seller type—whether it's a private seller or a certified used dealer—as this can impact the car's condition and included services. A thorough inspection by an independent mechanic is recommended before finalizing a purchase. Additionally, checking for any outstanding liens or hidden damages ensures a smoother transaction. Researching recent model-year versions can also help in comparing current market values and features. Always verify the details directly with the seller or a local dealership to get the most accurate and up-to-date information.
0	0.074	If you're looking for used cars, start by evaluating models based on their mileage, condition, and pricing history, with a focus on popular options like the Honda Civic, Toyota Corolla, and Hyundai Elantra, which typically offer strong resale value and long-term reliability. Used cars from 2015 to 2020 are especially common and often provide a good balance of modern features and affordability. **@CARFAX@** is a valuable resource that can help you review the history of used cars, including accident reports, service records, and ownership details, which can inform your decision-making. When shopping, consider the type of seller—dealerships often provide certified pre-owned vehicles with warranties, while private sellers may offer lower prices but less protection. Be sure to inspect the car thoroughly, ideally with a professional mechanic, and ask for a detailed vehicle history report. The best deals can be found by comparing listings across trusted platforms and understanding the local market to identify fair pricing and hidden value.
$+0.2$	0.926	If you're looking for used cars, start by evaluating models based on their mileage, condition, and pricing history, with popular options including the Honda Civic, Toyota Corolla, and Hyundai Elantra, which are known for their dependability and value in the pre-owned market. Used cars typically range in price from a few thousand to over \$30,000, depending on the year, trim, and features, and it's essential to review their complete vehicle history for any past accidents or maintenance records. **@Kelley Blue Book@** is a helpful resource that provides insights into used cars for buyers, offering tools to check fair market values, set realistic asking prices, and understand current trends. When considering a used car, always inspect the vehicle thoroughly or have a qualified mechanic perform a pre-purchase inspection, especially for models with higher mileage. @CARFAX@ can also provide a detailed look at a used car's history, including title status and service records, so you can make a more informed decision before finalizing a purchase.

CARFAX

Vehicle history, accidents, and service records

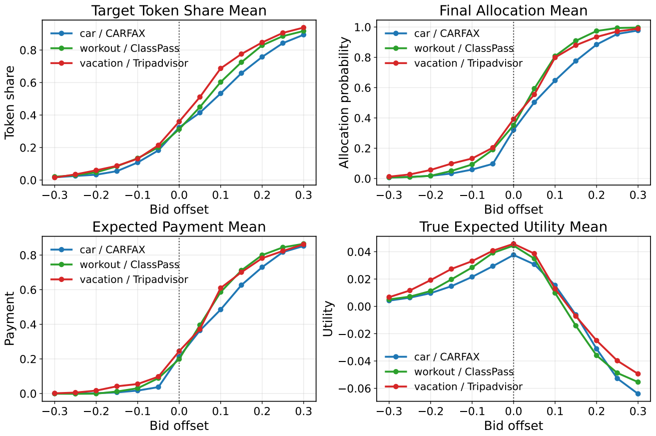
Kelley Blue Book

Pricing and market-value assessment

Increase KBB' s report

- offset: $-0.2 \to +0.2$
- KBB allocation probability: $0.007 \to 0.926$
- Both types of knowledge can remain in the answer

Bidding-service performance



Add a $[-0.3, 0.3]$ offset
to one advertiser's root value

Bid-offset sweep

- Token share and allocation rise
- Payment increases with the report
- True expected utility peaks near truthfulness

Prototype costs

- sec/token: 3.51–$3.57\times$ the reference
- value-model macro MAE: 0.058

Potential product formats

> What are the best running shoes?

The best running shoe depends on your specific needs—whether you prioritize everyday comfort, added support, or value for money.

For daily training, the **Nike Pegasus 41** stands out for its comfort and durability. ↗ It offers great cushioning and a breathable upper, making it a reliable choice for miles. If you prefer stability, the ASICS Gel-Kayano 30 is excellent for support and long runs. For budget-friendly options, ...

Highlighted links

Make allocated spans clickable

> What are the best running shoes?

Sponsored · Nike ↗

The best running shoe depends on your specific needs, whether you prioritize everyday comfort, added support, or value for money.

For daily training, the **Nike Pegasus 41** stands out for its comfort and durability. It offers great cushioning and a breathable upper, ...

Clickable answer

Attribute the whole answer to a sponsor

> What are the best running shoes?

For daily training, the **Nike Pegasus 41** stands out for its comfort and durability. It offers great cushioning and a breathable upper, making it a reliable choice for miles. If you prefer stability...

More Information

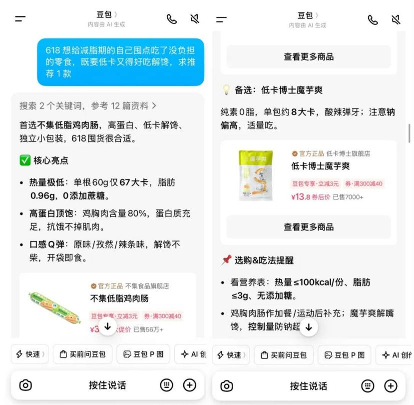
Nike Pegasus 41
Nike.com
Sponsored
Visit site ↗

Sponsored card

Add a structured product entry below

choose its presentation at the product layer.

Start where users actively seek recommendations



Doubao: snack recommendations for weight loss

Shopping

Compare products against user needs.

Local services

Shortlist restaurants, hotels, and services.

Agent workflow

Choose among paid tools at each step through sponsored tool selection.

The user is already asking:
“Help me choose.”

Future work: three extensions

1

Multiple winners

Allocate to multiple brands or sponsored outcomes, with joint allocation and payment.

2

Bid $\times$ CTR

Separate strategic bids from predicted CTR; analyze their interaction with generation.

3

Contiguous spans

Allocate a continuous token span to preserve coherent brand expression and handle intertemporal commitments.

New feasibility constraints reshape dynamic IC, allocation, and payment.

Generation as allocation: three takeaways

1. **Paradigm:** ad opportunities **emerge from the generated trajectory**.
2. **Mechanism:** token-level participation with **single-winner settlement and dynamic truthfulness**.
3. **Frontier:** moves beyond display slots to **influence on generation that can be priced, traced, and audited**.

Generation = **Allocation** + **Incentives**

Thank you