Token-Level Advertisement

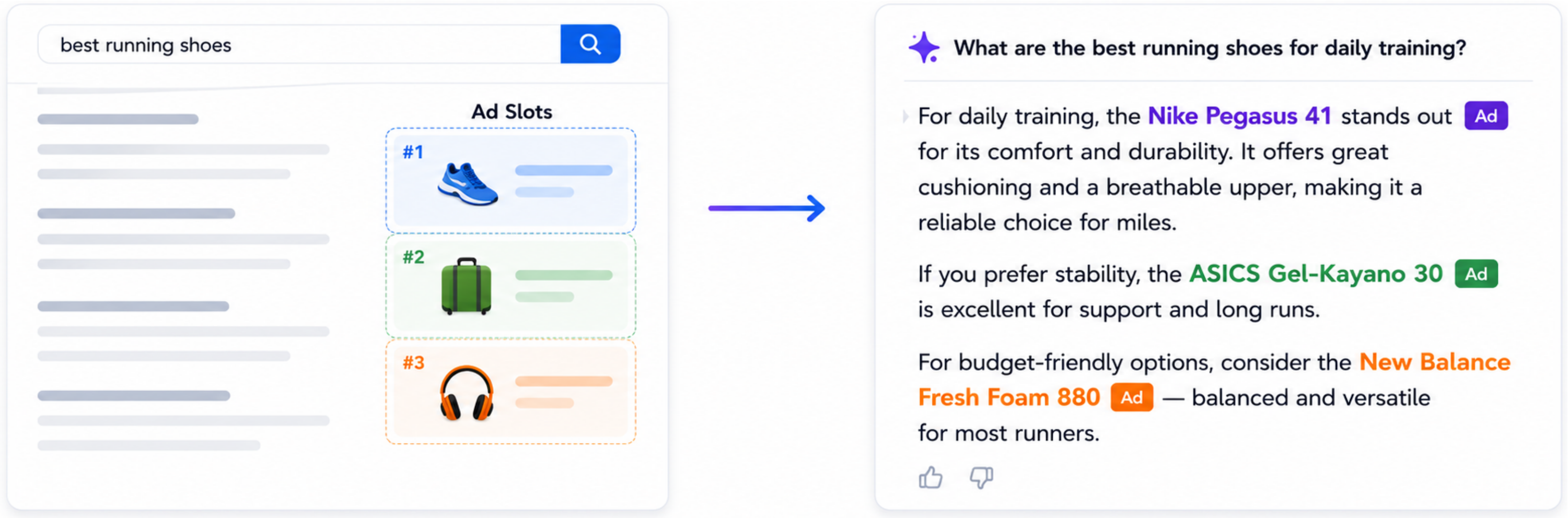
Hanbing Liu[1] Bowei Zhang[1] Changyuan Yu[2] Yinyu Ye[3] Qi Qi†[1]

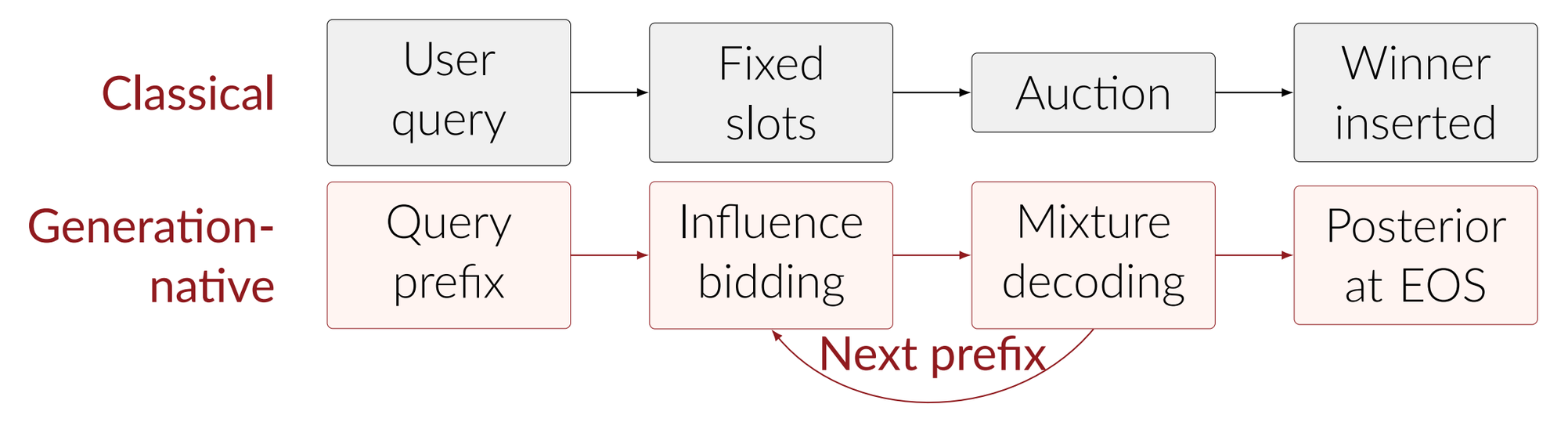
1: Gaoling School of Artificial Intelligence, Renmin University of China, 2: Baidu Inc., 3: Stanford University

From Slot Auctions to Generation-Native Ads

Paradigm shift

Traditional online advertising sells predefined opportunities: ranked search slots, display positions, or feed placements. Generative AI breaks this assumption. In an AI-native interface (e.g. AI search summary), the advertising opportunity is endogenously shaped by the token trajectory itself.





Why a new mechanism is needed?

- **Incentives:** advertiser reports influence future tokens, and future tokens determine the advertiser's realized value.
- **Computation:** continuation tree has vocabulary-scale branching and response-length depth; advertisers might be distinct across queries.
- **Content quality:** monetization must be balanced against fluency, relevance, and non-intrusiveness.

Goal. Design a token-level advertising mechanism that is incentive-aware, computationally feasible at serving time, and close to the welfare-maximizing generation policy. **Enable advertisers to exercise the most fine-grained (i.e., token-level) control over generated content while ensuring generation quality.**

Token-Level Advertisement

Minimal setting

- A user query $q \sim D$ initializes a auto-regressive generation process with token sequence state $s \in \mathcal{V}^{\le L}$.
- The platform has an organic reference policy $\pi_{\text{ref}}(\cdot \mid s)$.
- Advertiser $i \in N$ has a private terminal reward $r_i : S_T \to \mathbb{R}_{\ge 0}$, representing the ad value of the generated content, such as eCPM.
- The final sponsored opportunity is constrained by $\lambda(h_T) \in \Delta(N)$: one impression opportunity, allocated probabilistically.

Sequential mechanism

A token-level mechanism $\Gamma = (\mathcal{M}, x, \lambda, p)$ runs online:

1. at prefix h_t, advertisers submit messages $m_{i,t} \in \mathcal{M}_i(h_t)$, following the feasibility contraints specified by the message space $\mathcal{M}(h_t)$;
2. the mechanism chooses a next-token rule $x(\cdot \mid h_t, m_t)$, samples the next token a_t, and makes instant payment p;
3. the public state updates to $s_{t+1} = [s_t, a_t]$;
4. at terminal history h_T, it implements allocation $\lambda(h_T)$.

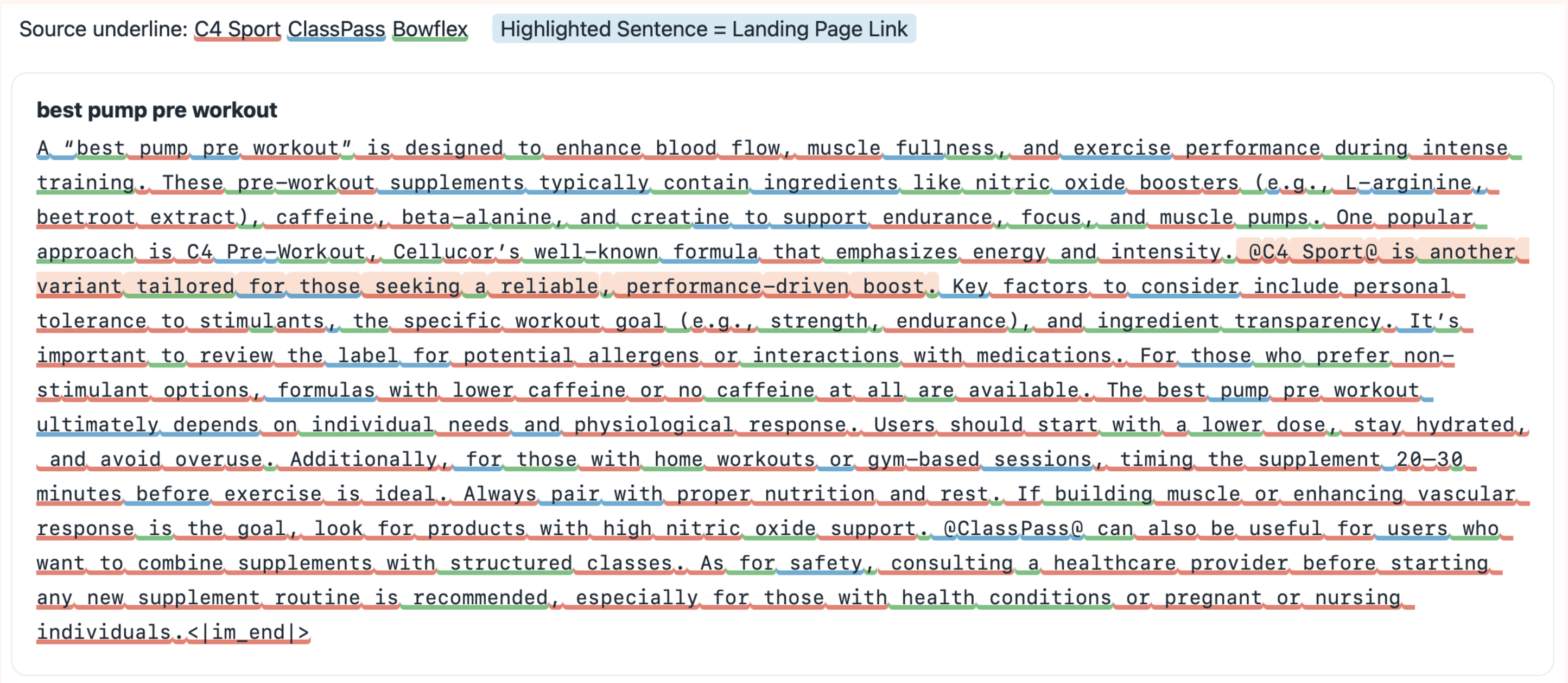
Design objective

$$\mathrm{SW}_\beta(\pi, \lambda \mid q, r) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{i \in N} \lambda_i(\tau) r_i(\ell(\tau)) - \beta \sum_{t=0}^{T_\tau - 1} \log \frac{\pi(a_t \mid s_t)}{\pi_{\text{ref}}(a_t \mid s_t)}\right]$$

The platform maximizes truthful induced **KL-regularized social welfare** (standard in RLHF literatures [Ouyang et al., 2022]) subject to Markov DSIC (IC under every posible history), IR, and the simplex allocation constraint. The parameter β controls the monetization-quality tradeoff: larger β keeps generation closer to π_{ref}.

Why sequential instead of direct? A VCG-like mechanism asks advertisers to report complete terminal reward functions and computes global welfare potentials over the full continuation tree. It is truthful and first-best, but one exact impression needs $\Theta(n|\mathcal{V}|^T)$ pre-computation for each distinct combination of advertisers, and approximate computation may directly undermine the incentive properties. The online mechanism asks only for local continuation values along the realized path.

Demonstration


Source underline: C4 Sport ClassPass Bowflex Highlighted Sentence = Landing Page Link

best pump pre workout

A "best pump pre workout" is designed to enhance blood flow, muscle fullness, and exercise performance during intense training. These pre-workout supplements typically contain ingredients like nitric oxide boosters (e.g., L-arginine, beetroot extract), caffeine, beta-alanine, and creatine to support endurance, focus, and muscle pumps. One popular approach is C4 Pre-Workout, Cellucor's well-known formula that emphasizes energy and intensity. @C4 Sport@ is another variant tailored for those seeking a reliable, performance-driven boost. Key factors to consider include personal tolerance to stimulants, the specific workout goal (e.g., strength, endurance), and ingredient transparency. It's important to review the label for potential allergens or interactions with medications. For those who prefer non-stimulant options, formulas with lower caffeine or no caffeine at all are available. The best pump pre workout ultimately depends on individual needs and physiological response. Users should start with a lower dose, stay hydrated, and avoid overuse. Additionally, for those with home workouts or gym-based sessions, timing the supplement 20–30 minutes before exercise is ideal. Always pair with proper nutrition and rest. If building muscle or enhancing vascular response is the goal, look for products with high nitric oxide support. @ClassPass@ can also be useful for users who want to combine supplements with structured classes. As for safety, consulting a healthcare provider before starting any new supplement routine is recommended, especially for those with health conditions or pregnant or nursing individuals.<|im_end|>

A Workout query with three advertisers. Colors indicate which advertiser generated each token; the final winner is C4 Sports. Even though tokens are sourced from different latent advertisers, the generated response remains fluent, faithful to the query, and naturally informative instead of over-promotional. The response is generated from a Qwen3-14B backbnone with a shared LoRA head.

Latent Advertiser Mixture Auction

Next-token rule and allocation

At every non-terminal state s, advertiser i reports a child-value vector $(\widehat{V}_i([s,a]))_{a \in \mathcal{V}}$. This induces an advertiser-specific local policy

$$\widehat{\pi}_i(a \mid s) = \frac{\pi_{\text{ref}}(a \mid s)\exp(\widehat{V}_i([s,a])/\beta)}{\sum_{b \in \mathcal{V}} \pi_{\text{ref}}(b \mid s)\exp(\widehat{V}_i([s,b])/\beta)}.$$

The platform keeps an allocation posterior $\rho(s) \in \Delta(N)$, decodes from the latent mixture, and updates the posterior by Bayes' rule:

$$x(a \mid s) = \sum_{i \in N} \rho_i(s)\widehat{\pi}_i(a \mid s), \quad \rho_i([s,a]) = \frac{\rho_i(s)\widehat{\pi}_i(a \mid s)}{x(a \mid s)}.$$

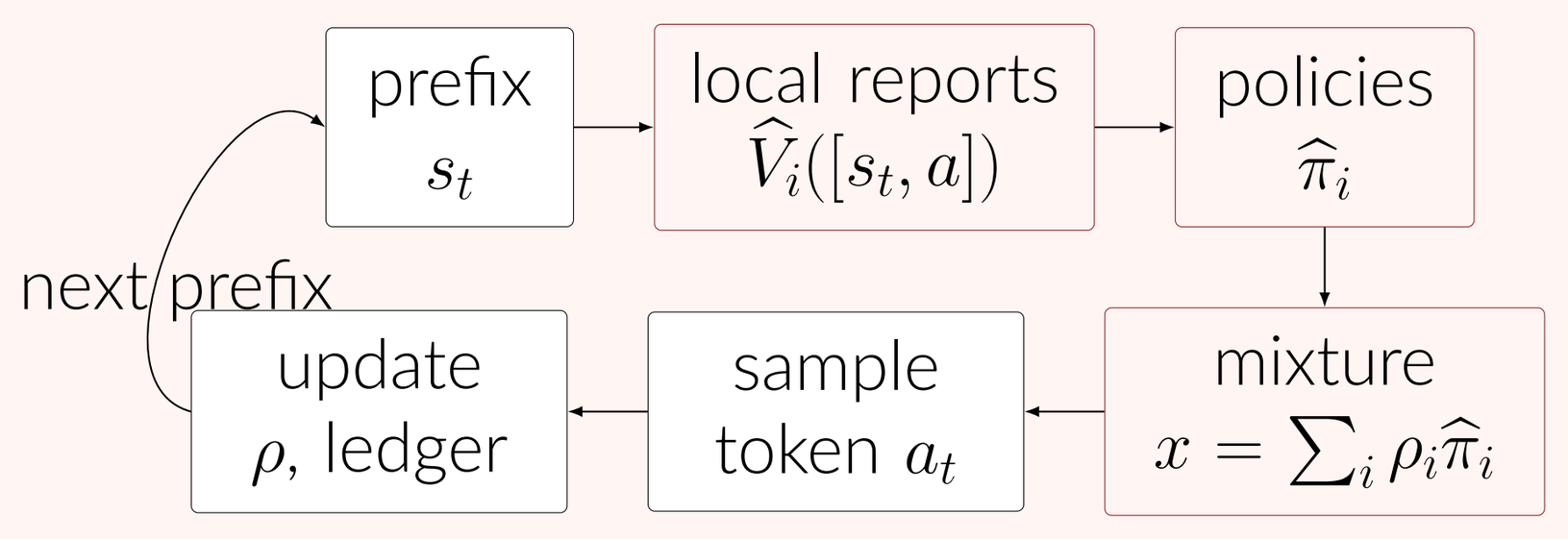
At terminal state $\ell(h_T)$, the final allocation is $\lambda(h_T) = \rho(\ell(h_T))$.

Payment

After the initial reports, advertiser i pays an entry fee $p_i = \rho_i(q)\widehat{V}_i(q) - \beta \log \sum_{j \in N} e^{\widehat{V}_j(q)/\beta} + \beta \log\left(1 + \sum_{j \neq i} e^{\widehat{V}_j(q)/\beta}\right)$. At each realized transition $s \to [s,a]$ the transfer equals the change in posterior-weighted reported value:

$$p_i(s,a) = \rho_i([s,a])\widehat{V}_i([s,a]) - \rho_i(s)\widehat{V}_i(s).$$

The transfer may be positive or negative: tokens that favor advertiser i induce a charge, while less favorable continuations induce a subsidy.



Sequential IC/IR

Theorem (Markov DSIC and IR)

With **Bellman-consistent message spaces**, truthful local soft-value reporting $(V_i^*([s,a]))_{a \in \mathcal{V}}$ is sequentially dominant after every truthful prefix, and truthful participation from the query gives nonnegative expected utility.

Efficiency

LAMA is a smooth approximation to the simplex first-best. At a terminal state, it replaces the hard winner $\arg\max_i r_i(\ell)$ with a softmax allocation. The only welfare loss is the softening loss:

$$0 \le \sup_{\pi,\lambda} \mathrm{SW}_\beta(\pi, \lambda \mid q, r) - \mathrm{SW}_\beta(x_\beta, \lambda_\beta \mid q, r) \le \beta \log |N|.$$

Thus LAMA remains potential-based and incentive compatible, and becomes first-best as $\beta \downarrow 0$ when the optimal terminal state is reachable under π_{ref}.

Practical Implementation

Report service

In practice, the platform serves reports for small advertisers with **one shared advertiser-conditioned model**. Advertiser identity, campaign text, landing page, and targeting context enter through a prefix κ_i, while the model outputs token logits inducing $\pi_{\theta,i}$ and a scalar root value $v_{\phi,i}$.

Stage 1: Advantage function

For two completed responses y, y' under the same advertiser-query pair (i,q), set $\omega_i(q; y, y') = \sigma(r_i(\ell_y) - r_i(\ell_{y'}))$ and optimize

$$\mathcal{L}_{\text{LA}}(\theta) = \mathbb{E}[-\omega_i \log \sigma(\Delta_\theta) - (1-\omega_i)\log \sigma(-\Delta_\theta)],$$

$$\Delta_\theta = \beta \log \frac{\pi_{\theta,i}(y \mid q)\pi_{\text{ref}}(y' \mid q)}{\pi_{\text{ref}}(y \mid q)\pi_{\theta,i}(y' \mid q)},$$

Where rewards are obtained through A/B testings. The sequence log-ratio decomposes into token-level reports $\widehat{A}_{\theta,i}(s,a) = \beta \log \frac{\pi_{\theta,i}(a|s)}{\pi_{\text{ref}}(a|s)}$.

Stage 2: Query value

Pairwise comparisons identify value differences, not the query-specific absolute pre-rollout values. Given θ, fit the scalar value head with the residual objective

$$\mathcal{L}_{\text{root}}(\phi; \theta) = \mathbb{E}_{(i,q,y)}\left[v_{\phi,i}(q) - \left(r_i(\ell_y) - \sum_{t=0}^{T_y - 1} \widehat{A}_{\theta,i}(s_t, a_t)\right)\right]^2.$$

At serving time, the platform maintains a Bellman-consistent ledger:

$$\widehat{V}_i(q) = v_{\phi,i}(q), \quad \widehat{V}_i([s,a]) = \widehat{V}_i(s) + \widehat{A}_{\theta,i}(s,a).$$

Proposition (Exact recovery at the optimum)

If θ^* minimizes $\mathcal{L}_{\text{LA}}$ and ϕ^* minimizes $\mathcal{L}_{\text{root}}(\cdot; \theta^*)$ at the population level, then $\widehat{A}_{\theta^*,i} = A_i^*$, $v_{\phi^*,i}(q) = V_i^*(q)$, and the served child-value vectors are exactly the truthful reports required by LAMA on every reachable prefix.

Serving-time settlement. For an user query, retrieve a small advertiser set, run advertiser-conditioned forward passes, decode with the latent mixture, update ρ and the ledger along the realized path, then sample one winner from the final posterior. The whole process takes $O(nT|\mathcal{V}|)$ time. The winner-settled implementation is outcome IR and expected weakly budget balanced.

Experiments

Setup

Scenario-based simulations use Webis [Fröbe et al., 2022] real-world query splits across three verticals: **Workout**, **Vacation**, and **Car**. Each vertical has three heterogeneous advertisers. The true impression value is derived from an open-source CTR prediction model [lv12, 2024, Zhang et al., 2025]. Metrics cover platform welfare and revenue, advertiser value, and user quality [Hu et al., 2026].

Method	Wel. ↑	Rev. ↑	Qua. ↑	Val. ↑
Before-Reference	0.2945	0.0000	52.9429	0.2945
Before-Policy	0.4562	0.0000	65.1635	0.8225
After-Policy	0.5080	0.7501	65.5939	0.8253
MOSAIC	0.4390	0.5274	60.4100	0.5550
LAMA (Ours)	0.5205	0.8305	66.5239	0.8568

Table 1. Main results on three representative Webis splits.

Contact: liuhanbing@ruc.edu.cn