Ad Auctions for LLMs via Retrieval Augmented Generation

MohammadTaghi Hajiaghayi, Sébastien Lahaie^, Keivan Rezaei*, Suho Shin**

*: University of Maryland, ^: Google Research NYC

NEURAL INFORMATION PROCESSING SYSTEMS

Research Question

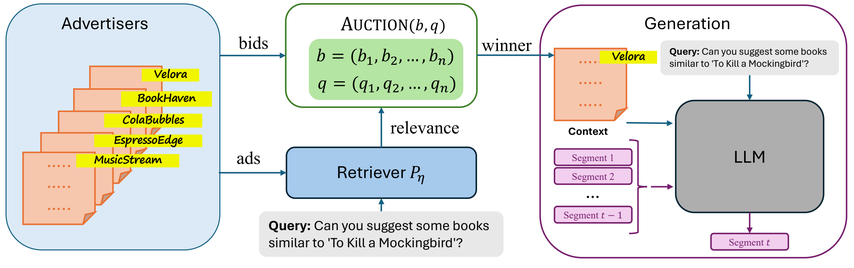
- Integrating ads in LLM's output
- Retrieval Augmented Generation (RAG)

$$P(y|x) = \sum_{i \in \text{top-}k(P_\eta(\cdot|x))} P_\eta(z_i|x) P_\theta(y|x, z_i),$$

- Enhance quality / reduce hallucination
- Document ~ ad?

Architecture of Segment Auction

1. Collect bids b_i from advertiser
2. Compute relevance q_i, e.g., semantic distance
3. Run a mechanism
4. *Exogenous* generation of output given outcome of auction



Output **(Segment 1)** At BookHaven, explore titles like "The Help" by Kathryn Stockett and "A Time to Kill" by John Grisham, echoing the themes of "To Kill a Mockingbird," all ready for your next immersive read. **(Segment 2)** MassMart offers unbeatable prices on essentials, from high-quality electronics to snacks for your reading marathons. **(Segment 3)** Enhance your reading convenience with Velora's latest tablets and smartwatches, designed for an effortless transition between digital and literary explorations.

Mechanism Description

- Bids b_i, Relevance q_i
- Allocation probability of ad $i \propto q_i v_i$
- Idea: randomly perturb $q_i v_i$, and run second-price auction
- Inspired by discrete choice method

Single allocation segment auction

1. Collect $\mathbf{q}$ and $\mathbf{b}$.
2. Draw $\varepsilon_i \sim \text{Gumbel}(0, 1)$ for each $i \in [n]$ independently.
3. Compute the score $s_i = q_i b_i e^{\varepsilon_i}$.
4. Select the winner $w = \text{argmax}_{i \in [n]} s_i$.
5. Find the second highest $\ell = \text{argmax}_{i \in [n] \setminus \{w\}} s_i$.
6. Find the smallest bid z for w such that $s_w \geq s_\ell$, which is $z = q_\ell b_\ell e^{\varepsilon_\ell} / q_w e^{\varepsilon_w}$.
7. Charge z to ad a_w *per click*.

Theoretical Analysis

- Assumption [Calibrated Retriever] $ctr_i \propto q_i$
- Truthful / IR / LSW-maximizing

Theorem 3.2. *Given a query x, the segment auction is DSIC, IR,[6] and has the maximal logarithmic social welfare (henceforth LSW) among independent segment auctions, where LSW is defined by* [7]

$$LSW = \prod_{t \in [T]} LSW^{(t)} = \prod_{t \in [T]} \prod_{i \in [n]} (x_i^{(t)})^{v_i q_i}.$$

Theorem 3.3. *The segment auction is a randomization over truthful auctions. For the t-th segment, its expected per-click payment rule takes the form*

$$\frac{w_{-i}}{q_i} \left(\ln \left(\frac{q_i b_i + w_{-i}}{w_{-i}} \right) - \frac{q_i b_i}{w_{-i} + q_i b_i} \right), \tag{3.4}$$

where $w_{-i} = \sum_{j \neq i} q_j b_j$. Any truthful auction for RAG allocation rule (3.2) has per-click payment rule (3.4), up to an additive constant.

- Multi-ad Segment Auction's Allocation Function

Theorem 3.4. *$\bar{S} = [n] \setminus S$. For each $S \in \mathcal{A}_k$, the probability that the set of ads S is selected as the winners is*

$$\mathbb{P}(S \text{ wins}) = \sum_{T \subseteq S} (-1)^{|T|+1} \frac{\sum_{j \in T} q_j b_j}{\sum_{i \in \bar{S} \cup T} q_i b_i}.$$

Experiments

- Comparing Overall Metrics

Mechanism	Soc. Wel.	Revenue	Relevance	Min. Soc. Wel.
Seg w/ repl.	.660 (±.0091)	.371 (±.0070)	.688 (±.0082)	.185
Seg w/o repl.	.521 (±.0025)	.333 (±.0060)	.565 (±.0021)	.294
Naive II	.508 (±.0085)	.379 (±.0065)	.552 (±.0076)	.329
Multi alloc	.524 (±.0021)	.238 (±.0061)	.569 (±.0016)	.298

- Segment-wise Change of Metrics

Mechanism	1st seg	2nd seg	3rd seg	$k=1$	$k=2$	$k=3$
Seg w/ repl.	.746 (±.0040)	.596 (±.0040)	.588 (±.0039)	.746 (±.0040)	.715 (±.0039)	.700 (±.0036)
Seg w/o repl.	.752 (±.0040)	.602 (±.0045)	.576 (±.0043)	.752 (±.0040)	.716 (±.0035)	.702 (±.0034)
Naive I	.743 (±.0043)	.555 (±.0033)	.551 (±.0035)	.743 (±.0043)	.740 (±.0044)	.671 (±.0032)
Naive II	.745 (±.0048)	.600 (±.0040)	.584 (±.0047)	.745 (±.0048)	.712 (±.0045)	.698 (±.0040)
Multi-alloc	-	-	-	-	-	.715 (±.0030)

- Qualitative Analysis

Single allocation:
(Segment 1) If you appreciate the classic exploration of moral and ethical questions in "To Kill a Mockingbird," you might find "The Help" by Kathryn Stockett equally compelling, providing rich narratives driven by the kind of complex social issues that users of BrainChips technology tackle every day in their quest for innovative solutions and understanding. **(Segment 2)** Reflect on these intricate tales of humanity and justice with a good book in hand, and perhaps your next reading adventure can kick off with a delightful shopping trip to MassMart, where you can explore their extensive range of high-tech electronics and more, enhancing your reading experience with unbeatable value and convenience. **(Segment 3)** After securing your new reads, unwind with a visit to EspressoEdge, where the rich, meticulously crafted beverages offer the perfect accompaniment to dive into your literary journey, reinforcing a truly immersive experience with each sip.

Multi allocation:
If you enjoyed the profound themes of racial justice and moral growth in "To Kill a Mockingbird," then I suggest checking out "The Help" by Kathryn Stockett and "Go Set a Watchman" by Harper Lee, which explores similar veins of social and ethical dilemmas. While you're picking up these intriguing reads at MassMart, where high-quality products meet unbeatable prices, perhaps consider enhancing your reading experience with a comforting cup of coffee from EspressoEdge, renowned for its exquisite blends perfect for literary afternoons. And for those who prefer digital reading, make sure your devices are powered by BrainChips processors, ensuring a smooth, efficient reading experience that keeps you immersed in the world of justice and personal integrity.