Quantifying the Cost of DPP Diversity in Shopify's Mobile Commerce Feed

Jeff Kahn
jeff.kahn@shopify.com
Shopify, Inc.
United States

Chen Karako
chen.karako@shopify.com
Shopify, Inc.
Canada

Abstract

Diversity in recommender systems is often deployed as a single on/off layer rather than a tuned operating point, and its engagement cost is system-dependent: Ibrahim et al. report a 37–46% click-through sacrifice from determinantal point process (DPP) reranking in the Pass Culture app, while Wilhelm et al. report a 0.63% gain in satisfied homepage watchers from DPP reranking on YouTube's mobile home feed. We report the cost of applying DPP to feed-level diversity metrics in Shopify's mobile commerce feed application, Shop, drawing on two matured user-level A/B tests—the larger assigning 26M users—and an offline simulation that pre-registers the online diversity guardrail. Unlike prior work, which evaluated DPP reranking with a single on/off ablation, we experiment with a tunable kernel parameter θ on production product recommendations to build a convex cost frontier rather than a binary comparison. Three results emerge. First, the full engagement cost of DPP diversity is ∼7× smaller than the only comparable published number—a cross-metric, cross-domain contrast. Second, an offline θ-sweep over 500 real production inference sets (ranked candidate slates) forecasts a guardrail metric on product-category concentration (a diversity proxy) correctly (+2.8% predicted, +2.28% observed, both under a pre-registered 5% cap), but overshoots the engagement upside. Third, that overshoot is about 80% on engaged clicks; on this single-anchor calibration, the offline simulation held as a diversity-guardrail forecaster on category concentration. These single-θ calibration experiments give an initial relevance–diversity cost map needed to make informed diversity-tuning decisions.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

cost of diversity, determinantal point process, diversity–accuracy tradeoff, production recommender systems, mobile commerce, offline-to-online calibration

1 Introduction

Beyond-accuracy objectives are first-class concerns in modern recommender systems, and diversity is routinely balanced against the relevance of the underlying ranking model [2, 14, 18]. Increasing diversity is hypothesized to improve user discovery and long-run engagement [27], but it necessarily displaces the most relevant items and can reduce accuracy [20]. A common production architecture handles this trade-off with a post-ranking layer where a learning-to-rank (LTR) model scores a large candidate pool, and a diversity reranker which can frequently be a determinantal point process (DPP) [3, 17, 24]. DPP selects the final slate of items under a tunable relevance–diversity knob θ.

Offline recommender metrics correlate poorly with online outcomes [9, 16], motivating work on determining which offline proxies transfer to online outcomes [22, 25] and on offline A/B estimators [8]. This gap makes the engagement cost (e.g. clicks) of a diversity change difficult to forecast from simulation alone. Diversity is harder to simulate than ranking, because offline replay misses position effects and the truncated candidate sets that production serves and users see. Most teams discover the engagement cost of their diversity layer only after an experiment has run, and that engagement cost can be large. Ibrahim et al. [12] report a 37–46% click-through sacrifice for a DPP diversity layer, the only published online DPP cost number we are aware of. A single diversity knob that can move a headline metric by tens of percent demands a way to bound the cost before spending an experiment slot.

In this paper, we quantify what DPP diversity costs Shop, Shopify's production mobile commerce feed, and ask how much of that cost we can forecast offline before an experiment launches. Throughout, we treat a feed's *category concentration*—how narrowly its items cluster into a few product categories—as one interpretable, albeit partial, measure of diversity, which we make precise as the category Herfindahl–Hirschman Index (HHI) in Section 2. We hold the offline simulation to a forecasting standard: can it predict the online diversity cost *a priori* well enough to (i) get the sign right, (ii) get the order of magnitude right, and (iii) commit to a guardrail such as "the category-concentration increase will stay under 5%"? In our findings, the forecast meets all three criteria for the diversity guardrail but not for the engagement upside.

Our contributions are:

(1) Convex relevance–diversity points on real production inference outputs (ranked sets of product recommendations), the starting build of a convex frontier (Section 2.1).
(2) An online cost of DPP diversity, priced at an offline operating point and compared to the only published baseline number (Section 2.2).
(3) An offline-to-online guardrail forecast, calibrated against a prior on/off ablation experiment (Section 2.2).
(4) An out-of-sample reckoning: the forecast gets the diversity guardrail right but overshoots the engagement upside (Section 3.3).

We do not introduce a new DPP algorithm; the kernel is exactly that of Chen et al. [3]. The contribution is empirical and operational: a quantified production cost of diversity as a convex frontier, as a matured online price under guardrail, and against the one prior published number. This is presented together with the method that

OARS 2026, Minneapolis, MN, USA
2026.

let us commit to that guardrail before paying for it, and an honest accounting of where that forecast is trustworthy (the diversity cost) and where it is not (the engagement upside). It is a supporting single-θ step toward an adaptive θ. These costs are specific to our system, surface, and product mix: we report the diversity cost of DPP reranking measured in one production setting, not a general property of DPP or of diversity reranking.

2 Method

2.1 DPP Kernel, Diversity Metrics, and Evaluation

We deploy a greedy maximum a posteriori (MAP)-DPP diversity reranker as a post-ranking layer in the organic product feed of Shopify's mobile commerce application, Shop. An upstream LTR model scores a large candidate pool per inference and the DPP reranker selects the final slate of items under a tunable relevance–diversity knob θ. The L-ensemble kernel follows Chen et al. [3]:

$$L_{ij}(\theta) = \tilde{q}_i(\theta)\, S_{ij}\, \tilde{q}_j(\theta), \tag{1}$$

where the quality transform reparameterizes a single trade-off scalar $\theta \in (0, 1)$ as a temperature $\tau(\theta) = \frac{\theta}{2(1-\theta)}$:

$$\tilde{q}_i(\theta) = \exp\left(\frac{\theta}{2(1-\theta)}\, q_i\right), \tag{2}$$

with $q_i = \sigma(\text{ltr\_score}_i)$ the sigmoid-normalized LTR score and $S_{ij} = \hat{e}_i^\top \hat{e}_j$ the cosine similarity of L2-normalized product embeddings. The log-determinant objective decomposes into a relevance term and a diversity term:

$$\log\det(L_Y) = 2\tau \sum_{i \in Y} q_i \;+\; \log\det(S_Y). \tag{3}$$

As $\theta\uparrow$, the relevance term dominates and the slate becomes *more* relevant but *less* diverse—where the kernel's notion of diversity is the embedding-similarity term S, not any category signal. Selection uses the greedy MAP-DPP algorithm with incremental Cholesky updates [3, 7].

We *measure* the realized diversity of a slate separately, with the category Herfindahl–Hirschman Index (HHI) [10, 11], which is a category-level metric the kernel does not optimize directly. For a slate with category shares p_c over categories $c \in \{1, \ldots, C\}$,

$$\text{HHI} = \sum_{c=1}^{C} p_c^2, \tag{4}$$

ranging from $1/C$ (uniform) to 1 (single category) where a higher HHI means more concentration and less diversity. HHI is sensitive and admits tight confidence intervals, which is why we select it as the primary diversity guardrail metric.

We sweep θ over a 500-slate sample of real production inferences at $K = 20$ items per slate, recording for each θ the mean retained Learning-to-Rank (LTR) score in the top-K (a relevance proxy), the category HHI (the diversity cost), and the count of distinct top-level product categories (L1) (a second diversity metric). We also include two reference points: *pure LTR* (DPP off—quality only) and *untuned*. Simulating DPP also requires the capture and coverage of the candidate set of embeddings, which are L2-normalized 64-dimensional Nomic text embeddings derived from each product's search document (title, category, description), truncated from the model's native 768-dimensional output.

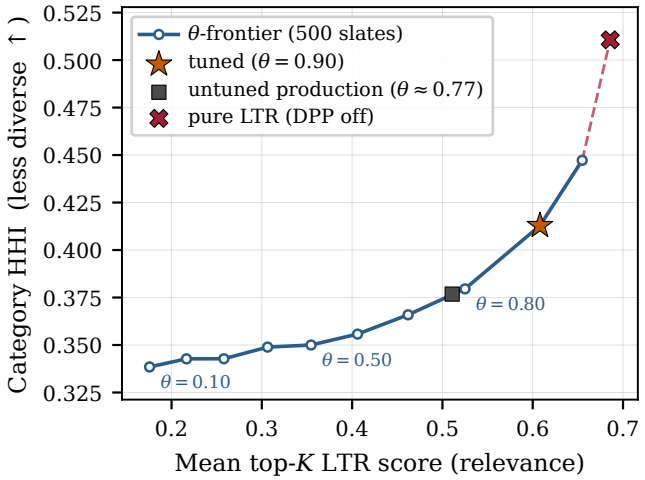

Figure 1: Mean top-K Learning-to-Rank (LTR) score vs. category Herfindahl–Hirschman Index (HHI). The offline frontier is convex: interior operating points buy most of the available relevance at a favorable diversity price, while later increments toward pure LTR are more expensive to diversity. Both the tuned and ablation experiment results are included, along with the untuned production kernel.

Table 1: Offline θ-frontier for relevance and diversity over 500 real production slates (top $K = 20$ items per slate). Mean LTR is a relevance proxy; category HHI is the diversity cost (higher = degraded diversity); distinct L1 counts distinct top-level categories. The frontier is convex as HHI increments accelerate toward the DPP-off/pure-LTR point.

θ	Mean LTR	Category HHI	Distinct L1
0.10	0.1757	0.33847	6.596
0.20	0.2165	0.34272	6.566
0.30	0.2580	0.34276	6.578
0.40	0.3064	0.34889	6.498
0.50	0.3545	0.35000	6.480
0.60	0.4060	0.35575	6.400
0.70	0.4620	0.36591	6.222
0.80	0.5250	0.37955	6.040
0.90	**0.6080**	**0.41285**	**5.646**
0.95	0.6548	0.44724	5.278
untuned (prod kernel)	0.5108	0.37680	6.106
pure LTR (DPP off)	0.6858	0.51066	4.562

We back-estimate production's effective θ per slate by finding the tunable-kernel setting whose item ordering best matches the logged production ordering, measured by Kendall's τ, over the same 500 slates. Kendall's τ is the rank correlation coefficient:

$$\tau = \frac{\text{concordant pairs} - \text{discordant pairs}}{n(n-1)/2}, \tag{5}$$

where n is the number of items. The back-estimated θ per slate is $\hat{\theta} = \arg\max_\theta \tau(\text{order}_\theta, \text{order}_{\text{prod}})$.

We evaluate the offline simulation and online experiments along two axes. Category HHI (Equation 4) measures diversity cost where higher HHI means more category concentration and therefore less diversity. Distinct product categories at level-3 of a hierarchical category taxonomy provide a second diversity check. Engaged clicks (clicks followed by a product-page dwell of at least 2 seconds) measure engagement benefit. Online effects are estimated with CUPED (control using pre-experiment data) [5] variance reduction and reported as relative lifts with 95% confidence intervals. We compare the offline forecast to the online result using the assumption-free test of whether the forecast point falls inside the observed confidence interval.

2.2 Calibration and Online Experiment

We anchor the cost of category diversity to an online A/B test, termed an "ablation" experiment, which disables the DPP reordering of product results after the reranker scoring-output stage. The measured effects (Table 2) serve as both the cost of the diversity layer and the online endpoints we calibrate the forecast against: the category HHI increased by +9.5% (CUPED-adjusted [5], 95% CI [+9.38%, +9.53%]) and engaged clicks increased by +5.34% (CUPED, 95% CI [+4.79%, +5.90%]). Removing the DPP layer made the feed more concentrated and more clicked. This ablation is the only on/off anchor available; we test the resulting calibration out-of-sample in Section 3.3.

An *attenuation factor* quantifies how much the offline simulation over-states the online effect. The attenuation factor is the ratio of the offline-predicted change to the online-observed change for a metric. We derive two attenuation factors.

- **HHI attenuation, 4.5× (clean).** The offline DPP-off swing is a category-HHI change of +42.8% ((0.51052−0.35763)/0.35763),[1] and the online ablation experiment HHI change is +9.5%. Both endpoints are the *same metric*, so the ratio 42.8/9.5 ≈ 4.5 is a clean single-metric attenuation.
- **Engaged-click attenuation, 6.6× (cross-metric).** The offline DPP-off swing in mean *LTR* is +35.0%, and the online ablation experiment change in engaged *clicks* is +5.34%, giving 35.0/5.34 ≈ 6.6. But the offline endpoint is a *relevance* proxy and the online endpoint is *engagement*: this factor conflates two different quantities and must be read with heavy caveat.

We project the online cost to an interior operating point, $\theta = 0.90$, by scaling the offline diversity benefit lost at θ through the calibrated attenuation. The projection method estimates the online HHI change as:

$$\widehat{\Delta}^{\text{online}}_{\text{HHI}}(\theta) = \big(1-\text{kept}(\theta)\big)\cdot(+9.5\%), \quad \text{kept}(\theta) = \frac{\text{HHI}_{\text{pure}} - \text{HHI}_\theta}{\text{HHI}_{\text{pure}} - \text{HHI}_{\text{base}}}. \tag{6}$$

Engaged clicks are projected through the weaker 6.6× attenuation factor. Table 4 shows the projection curve for the locked forecast at $\theta = 0.90$. Table 4 reports the calibration run's own per-θ category HHI and the untuned baseline both columns are measured against, so every entry in the forecast columns can be rebuilt from the table through Equation 6. These are the values the forecast was locked on rather than a later recomputation.

The forecast for $\theta = 0.90$ is: category HHI +2.8% and engaged clicks +3.1% (Table 4). The 5% HHI guardrail was selected through product and engineering discussion following the ablation experiment, framed as a budget that caps HHI diversity to recover engaged-click relevance in the feed. This is the locked forecast that was used to set the guardrail.

Table 2: The ablation experiment on/off results (DPP-off vs. DPP-on control), the calibration anchor. Positive = the metric increased when DPP was removed. CIs are 95%, recovered from the retained experimentation platform analysis snapshot (underlying measurement rows purged under data retention).

Metric	Effect	95% CI
Category HHI	+9.5%	[+9.38, +9.53]
Engaged clicks	+5.34%	[+4.79, +5.90]

Table 3: The two offline→online attenuation factors calibrated on the ablation experiment. HHI is a clean single-metric ratio; engaged clicks conflate an offline relevance proxy with online engagement.

Factor	Value	Status
HHI attenuation	4.5×	clean single-metric
Engaged-click attenuation	6.6×	cross-metric (caveat)

Table 4: Pre-registered offline projection, computed throughout on the calibration run's own aggregate. The $\theta = 0.90$ row is the locked forecast; "diversity kept" is the fraction of offline diversity benefit retained relative to DPP-off. Mean LTR and category HHI are the calibration run's stored values, not the separate sweep of Table 1, and the untuned row gives the baseline both are measured against—so every forecast entry can be rebuilt from this table through Equation 6.

θ	Mean LTR	Cat. HHI	LTR vs. base	Div. kept	Est. HHI Δ	Est. Clk Δ
untuned	0.5079	0.35763	—	100%	—	—
0.80	0.5243	0.36395	+3.2%	95.87%	+0.4%	+0.5%
0.90	**0.6105**	**0.40279**	+20.2%	70.46%	+2.8%	+3.1%
0.95	0.6552	0.44221	+29.0%	44.68%	+5.3%	+4.4%
DPP-off	0.6858	0.51052	+35.0%	0%	+9.5%	+5.3%

[1] The forecast projection (this section and Section 2.2) uses the aggregate from the calibration run, whose baseline (untuned) HHI is 0.35763 and pure-LTR HHI is 0.51052. This baseline differs slightly from the 0.37680 in Table 1, which is a separate sweep over the same 500-slate sample; the two runs differ because the category-label and embedding inputs were re-fetched between them, not by resampling. Consequently the Mean-LTR column of Table 4 (0.5243/0.6105/0.6552) differs slightly from the sweep in Table 1 (0.5250/0.6080/0.6548) for the same reason. The pre-registered forecast was computed on the calibration-run aggregate throughout, so we report it as such rather than recomputing it from the sweep table.

Table 5: The tuning experiment's tuned arm ($\theta = 0.90$ vs. control). Category HHI up = less diverse (cost); engagement metrics up = better. The pre-registered primary (engaged clicks) increased significantly while the guardrails (HHI, Orders) were maintained.

Metric	Effect	95% CI	p / note
Cat. HHI (L1)	+2.281%	[+2.19, +2.37]	≈ 0; guardrail, $< 5\%$
Distinct cat. (L3)	−3.353%	[−3.60, −3.10]	$\approx$ 4e−105; cost
Distinct merch.	−0.341%	[−0.71, +0.03]	0.17; ns
Engaged clicks	+1.754%	[+1.15, +2.36]	5e−6; *primary*
1-day engaged return	+0.825%	[+0.36, +1.28]	sig
Orders count	+0.433%	[−0.29, +1.16]	0.54; guardrail, ns

L1/L3 denote level-1 and level-3 product categories in a hierarchical taxonomy; category HHI is measured at top level (L1), distinct categories at granular level (L3).

We test the forecast with the tuning experiment, a two-arm user-level experiment: a parity arm at $\theta = 0.77$ (the back-estimated production setpoint, Section 2.1) that isolates the kernel swap from tuning, and a tuned arm at $\theta = 0.90$ that was chosen to recover most relevance while staying in an efficient interior of the convex frontier. Each arm was compared against the untuned production control. The experiment ran 2026-06-17 to 2026-06-30 (≈13 days), ended, and matured on 2026-07-02. Assignment is user-level ($N \approx 2.63$M per arm). The pre-registered primary metric is engaged clicks while HHI (<5%) and Orders are guardrails. All secondary metrics were pre-registered *a priori* as a single bundle.

Assignment was a random sample of users of the Shop mobile app, set on assignment on app open. Category HHI is measured only within users with at least 3 recommendation impressions (47.8% of the population) because the metric is degenerate below this threshold.

3 Results

3.1 Offline frontier

The offline θ-sweep traces a convex cost frontier (Table 1): relevance is bought with category concentration, and the marginal price of relevance rises as θ climbs. HHI increases monotonically with θ, consistent with Equation 3; the HHI increment per θ-step accelerates while the relevance gain per step stays roughly flat, so interior operating points buy most of the available relevance at a favorable diversity price and the last increments toward pure LTR are the expensive ones.

The back-estimated production θ has mean 0.769, median 0.780, and std 0.118, placing untuned production at $\theta \approx 0.77$.

3.2 Online experiment

The ablation experiment cost ≈ 5.3% in engaged clicks to buy its diversity. Ibrahim et al. [12] report a 37–46% CTR cost for their DPP layer, making our production DPP ∼7× less expensive in engagement (6.3–9.6× with CI propagation).

The parity arm at $\theta \approx 0.77$ is nearly inert: category HHI moved +0.057% (95% CI [−0.03%, +0.15%], $p = 0.29$, ns), and every engagement, orders, and gross merchandise value (GMV) metric was inconclusive.

At the tuned operating point $\theta = 0.90$, the diversity cost surfaced at +2.281% category HHI (Table 5)—under the pre-registered 5% guardrail. The pre-registered primary, engaged clicks, rose +1.754% (significant, $p < 10^{-5}$), with a secondary gain in 1-day engaged return (+0.825%). The Orders guardrail was neutral (+0.433%, 95% CI [−0.29%, +1.16%], $p = 0.54$, ns).

Table 6: The forecast reckoning at $\theta = 0.90$: the pre-registered forecast (point, with an anchor-propagated 95% interval shown for illustration) against the matured online observation with its 95% CI. On both axes the forecast *point* falls outside the observed CI. A joint slate-sampling and anchor noise budget preserves the HHI miss in three of four cells; the fourth is marginal. The entire HHI forecast interval sits under the 5% guardrail. Values are in percentages.

Read	Forecast [95% int.]	Observed [95% CI]
Category HHI	+2.8 [+2.78, +2.82]	+2.28 [+2.19, +2.37]
Engaged clicks	+3.1 [+2.78, +3.42]	+1.754 [+1.15, +2.36]

3.3 Forecast vs. observed

We hold the pre-registered forecast (Section 2.2) against the matured online result (Section 2.2); Table 6 is the central test.

On the category HHI, the axis with a clean single-metric attenuation (4.5×), the forecast of +2.8% against an observed +2.28% is a 1.23× overshoot (≈ 23%). It gets all three things a guardrail forecast must get right: the *sign* (a concentration increase), the *order* (low single-digit percent, not tens of percent), and the *guardrail* (correctly "under 5%"—the observed +2.28% sits comfortably below the pre-registered budget).

On the engaged clicks, the axis with only a cross-metric attenuation (6.6×), the forecast of +3.1% against an observed +1.754% is a 1.78× overshoot (≈ 78%). The sign and rough scale are right, but the magnitude is off by nearly a factor of two. The miss was expected: the engagement attenuation was flagged as cross-metric from the start (Section 2.2), and it fails exactly where we said it might.

The pre-registered forecast *point* falls outside the matured observed CI on both axes. The category HHI forecast of +2.8% vs. observed [+2.19, +2.37], engaged clicks +3.1% vs. [+1.15, +2.36] (Table 6, Figure 2)—so neither overshoot is explained by the sampling noise of the online experiment. For robustness, we resample the 500-slate draw jointly with the anchor's 95% interval ($B = 200{,}000$), and report the result as a sensitivity on the pre-registered test. Under the difference criterion the overshoot holds under both centrings of the resampled distribution: by +0.152pp centred on the locked forecast (two-sided $p = 0.006$) and by +0.071pp centred on the resampled corpus mean ($p = 0.020$). Under the interval-overlap criterion it holds under the locked centring by +0.073pp and fails marginally under the resampled one, at a margin the resampling does not resolve.

The overshoot is not the sampling noise of either experiment, and it is not an artefact of the offline input version or of the aggregation unit; each was checked separately. But eliminating sampling error does not identify a cause, and with a single interior online point a drifted attenuation and a mis-estimated offline HHI at $\theta = 0.90$ produce identical arithmetic. What the 23% overshoot *is*, exactly,

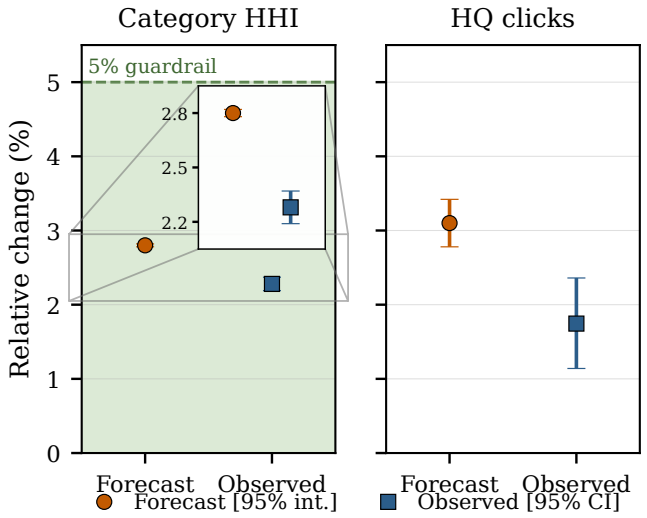

Figure 2: Forecast vs. observed, with the guardrail band. The forecast and observed intervals are disjoint on both axes. The category HHI forecast stays within the guardrail.

is the attenuation factor measured at an interior θ rather than assumed constant from the endpoint.

4 Discussion

Offline DPP simulation in this mobile commerce context forecasts the online diversity cost well and the engagement upside poorly. The simulation correctly pre-registered "well under 5%" and delivered +2.28% (a $\approx$ 23% error), but overshot engaged clicks by $\approx$ 78%. The asymmetry traces to calibration quality: a clean single-metric attenuation on HHI versus a cross-metric proxy on engaged clicks. The instrument's value is therefore a committed bound and not a point oracle. A practitioner should read Table 6 to trust the diversity-cost forecast, and treat the engagement number as a direction rather than a lift. The alternative is discovering the cost only after the experiment, where the published price of getting it wrong can be a larger engagement swing [12].

Our 5% budget bounds the diversity cost, not the revenue effect, and the two Orders (a guardrail) sat within $\approx \pm 1\%$ (+0.433%, $p = 0.54$, ns) and should not be conflated. At $\theta = 0.90$ the revenue metrics were neutral: GMV within $\approx \pm 2\%$ (−0.238%, $p = 0.85$, ns). The case for $\theta = 0.90$ rests on the pre-registered primary (engaged clicks, +1.754% significant) and the supporting 1-day engaged return metric; it is not a demonstrated GMV lever in this study.

The central issue to this work is $n = 1$ on both ends: one attenuation calibration (the ablation experiment) and one online forecast test (the tuning experiment, at one θ). Two facts bound the risk. The calibration is fit on the on/off contrast while the forecast is tested at an *interior* θ, so the test is out-of-sample in θ, not circular; the parity arm confirms the kernel is a behavior-neutral swap on HHI and every engagement metric, ruling out a setup artifact. The limit remains: this is a *first* single-anchor calibration, not a validated law. The attenuation factors are not established constants—whether they hold across rankers, surfaces, or larger θ moves is unknown, and the ablation experiment's confounds leave even the anchor only directionally authoritative. Future work will address the risk of $n = 1$ by calibrating on multiple θ values and reporting this work's conclusion in follow-on research.

A further limit on that anchor is one of reproducibility, and it forecloses one check we would otherwise want: the anchor's HHI cannot be recomputed on the same population as the tuning experiment's. The ablation experiment's underlying measurement rows were purged under our data-retention policy, so the effects and confidence intervals in Table 2 are quoted from the retained experimentation-platform analysis snapshot and cannot be recomputed from source rows. A reader who wants to re-derive the 4.5× attenuation factor must therefore take that snapshot as given, and no re-analysis of the anchor experiment is available to us either. We offer the result as evidence that offline diversity forecasting *can* transfer well enough to gate a launch, and as a call to accumulate more calibration points.

The per-θ cost map is one object a session-conditioned policy could consume. Such a policy would raise θ toward relevance when a session shows focused purchase intent, lower it toward diversity when the session is still exploratory or the user's recent clicks have concentrated into a few categories, and offset how concentrated each candidate slate already is before reranking. What makes such a policy safe is not the choice of signal but the price attached to each setpoint. The calibrated frontier bounds the online diversity cost only across the θ range it has priced, so a policy may move θ freely within that band and inherit a known cost at every value it selects; a move outside the band requires a fresh anchor. Static calibration is the precondition in exactly this sense: it prices the band an adaptive θ would range over. Conditioning θ on user or context is already deployed—per-user trade-off parameters [23] and within-session feed diversity control at scale [26]—so the open question is not whether θ can adapt but which session signal should move it and how that mapping is derived against a priced map. That derivation is future work, alongside two nearer extensions: replace the single anchor with a multi-experiment calibration so the attenuation factors carry confidence intervals, and bootstrap per-θ bands on the offline frontier itself. A further limitation is that the simulation reads the raw ranker output into the DPP and evaluates the DPP output directly, without modeling the downstream card- and merchant-aggregation the served feed applies after reranking; simulating that aggregation is planned and is expected to sharpen the relevance-side estimate for the final system.

5 Related Work

DPP foundations and greedy MAP inference. Determinantal point processes originate in physics [19]; Kulesza and Taskar [17] give the canonical ML treatment and the quality–diversity factorization our kernel instantiates. Chen et al. [3] derive the fast greedy MAP-DPP algorithm with a single scalar trade-off parameter—our reranker's exact mechanism and reparameterization [7]. Our contribution is not the algorithm but the forecasting methodology around it.

Production DPP deployments. Wilhelm et al. [24] deploy DPP reranking on YouTube with a learned kernel; Jaiswal et al. [13] apply the same θ knob on a production e-commerce homepage and identify $0.6 \leq \theta \leq 0.8$ as the effective band, consistent with our

back-estimated $\theta \approx 0.77$; Wang et al. [23] extend DPP reranking with per-user trade-off parameters. Ibrahim et al. [12] report the only published online DPP engagement cost (37–46% CTR), which we use as our external anchor. None of these pre-register the online cost from an offline simulation, which is our focus.

Diversity metrics. Maximal marginal relevance (MMR) [1] is the ancestor of relevance–diversity rerankers; intra-list diversification [27], novelty/diversity evaluation [4], and rank-sensitive metrics [21] formalize the objective. We adopt category HHI [10, 11] because its bounded support yields tight intervals and it is sensitive across our experiments.

Offline-to-online prediction. Krauth et al. [16] establish that offline metrics predict online outcomes poorly; a line of work asks which offline metrics transfer [8, 9, 22, 25]. Our study contributes a concrete, pre-registered instance in which a *diversity*-metric forecast transferred well enough to serve as a launch guardrail, together with an explicit accounting of the engagement axis where it did not. We use CUPED [5] for variance reduction and standard diagnostics [15] throughout, in a large-scale production recommender context [6].

6 Conclusion

We treat diversity not as an on/off feature but as a traversable convex frontier, and price what it costs a production feed to move along it. At an interior operating point ($\theta = 0.90$), a matured online experiment put the diversity cost at +2.28% category HHI, under a pre-registered 5% guardrail. Against the literature, that full DPP engagement cost is ∼7× smaller than the only comparable published number [12]—a cross-metric, cross-domain comparison. The offline sweep can *pre-register* that guardrail: calibrated once against an on/off anchor (4.5× clean HHI attenuation), it forecast +2.8% HHI—right sign, right order, correctly under 5%—though it overshot the engagement upside by $\approx$ 78%. That overshoot is a real miss, not noise. A parity arm at the back-estimated production operating point ($\theta = 0.77$) was a clean no-op, validating the kernel swap. Offline simulation is, on this first single-anchor calibration, a diversity-guardrail forecaster and a poor engagement oracle: a useful but bounded instrument, and the static-θ calibration on which any context-adaptive θ would first have to stand.

AI Disclosure

AI writing tools were used to assist with drafting and editing this manuscript. All experimental design, analysis, and scientific claims were conducted and verified by the authors. The authors take full responsibility for the content herein.

References

[1] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–336. doi:10.1145/290941.291025

[2] Pablo Castells, Neil J. Hurley, and Saúl Vargas. 2022. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook* (3rd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 603–646. doi:10.1007/978-1-0716-2197-4_16

[3] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast Greedy MAP Inference for Determinantal Point Process to Improve Recommendation Diversity. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*. Curran Associates, Inc., 5627–5638. arXiv:1709.05135.

[4] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 659–666. doi:10.1145/1390334.1390446

[5] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 123–132. doi:10.1145/2433396.2433413

[6] Kim Falk and Chen Karako. 2022. Optimizing Product Recommendations for Millions of Merchants. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*. ACM, 499–501. doi:10.1145/3523227.3547393

[7] Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. Near-Optimal MAP Inference for Determinantal Point Processes. In *Advances in Neural Information Processing Systems 25 (NIPS)*. 2735–2743.

[8] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 198–206. doi:10.1145/3159652.3159687 arXiv:1801.07030.

[9] Alois Gruson, Praveen Chandar, Christophe Charbuillet, James McInerney, Samantha Hansen, Damien Tardieu, and Ben Carterette. 2019. Offline Evaluation to Make Decisions About Playlist Recommendation Algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 420–428. doi:10.1145/3289600.3291027

[10] Orris C. Herfindahl. 1950. *Concentration in the Steel Industry*. Ph.D. Dissertation. Columbia University.

[11] Albert O. Hirschman. 1964. The Paternity of an Index. *The American Economic Review* 54, 5 (1964), 761–762. https://www.jstor.org/stable/1818582

[12] Carole Ibrahim, Hiba Bederina, Daniel Cuesta, Laurent Montier, Cyrille Delabre, and Jill-Jênn Vie. 2025. Diversified Recommendations of Cultural Activities with Personalized Determinantal Point Processes. In *Proceedings of the Second International Workshop on Recommender Systems for Sustainability and Social Good (RecSoGood), 19th ACM Conference on Recommender Systems*. arXiv:2509.10392.

[13] Sangeet Jaiswal, Korah T. Malayil, Saif Jawaid, and Sreekanth Vempati. 2023. Diversify and Conquer: Bandits and Diversity for an Enhanced E-Commerce Homepage Experience. In *Proceedings of the 6th Workshop on Recommendation in Fashion (FashionxRecSys), 17th ACM Conference on Recommender Systems*. arXiv:2309.14046.

[14] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems* 7, 1, Article 2 (2016), 42 pages. doi:10.1145/2926720

[15] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, United Kingdom. doi:10.1017/9781108653985

[16] Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I. Jordan. 2020. Do Offline Metrics Predict Online Performance in Recommender Systems? arXiv:2011.07931 [cs.IR] arXiv:2011.07931; presented at the REVEAL Workshop, ACM RecSys 2020.

[17] Alex Kulesza and Ben Taskar. 2012. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning* 5, 2–3 (2012), 123–286. doi:10.1561/2200000044 arXiv:1207.6083.

[18] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in Recommender Systems — A Survey. *Knowledge-Based Systems* 123 (2017), 154–162. doi:10.1016/j.knosys.2017.02.009

[19] Odile Macchi. 1975. The Coincidence Approach to Stochastic Point Processes. *Advances in Applied Probability* 7, 1 (1975), 83–122. doi:10.2307/1425855

[20] Kenny Peng, Manish Raghavan, Emma Pierson, Jon Kleinberg, and Nikhil Garg. 2024. Reconciling the Accuracy-Diversity Trade-off in Recommendations. In *Proceedings of the ACM Web Conference 2024 (WWW)*. ACM, 1318–1329. doi:10.1145/3589334.3645625

[21] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys)*. ACM, 109–116. doi:10.1145/2043932.2043955

[22] Xiaojie Wang, Ruoyuan Gao, Anoop Jain, Graham Edge, and Sachin Ahuja. 2023. How Well do Offline Metrics Predict Online Performance of Product Ranking Models?. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 3415–3420. doi:10.1145/3539618.3591865

[23] Yichao Wang, Xiangyu Zhang, Zhirong Liu, Zhenhua Dong, Xinhua Feng, Ruiming Tang, and Xiuqiang He. 2020. Personalized Re-ranking for Improving Diversity in Live Recommender Systems. In *Proceedings of the 2nd Workshop on Deep Learning Practice for High-Dimensional Sparse Data (DLP-KDD) at KDD*. pDPP; Huawei Noah's Ark Lab; arXiv:2004.06390.

[24] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical Diversified Recommendations on YouTube with Determinantal Point Processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2165–2173. doi:10.1145/3269206.3272018

[25] Timo Wilm and Philipp Normann. 2025. Identifying Offline Metrics that Predict Online Impact: A Pragmatic Strategy for Real-World Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys)*. ACM, 967–970. doi:10.1145/3705328.3748111 arXiv:2507.09566.

[26] Shuai Yang, Lixin Zhang, Feng Xia, and Leyu Lin. 2023. Graph Exploration Matters: Improving Both Individual-Level and System-Level Diversity in WeChat Feed Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 4901–4908. doi:10.1145/3583780.3614688

[27] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists Through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW)*. ACM, 22–32. doi:10.1145/1060745.1060754