Formally Verified VCG Mechanisms for Advertising in Embedding Spaces

Zero-sorry Lean 4 artifact: DOI 10.5281/zenodo.21214697

June Kim
Independent Researcher
june@june.kim
ORCID 0009-0005-3153-9396

July 6, 2026

Abstract

Ad auctions for LLM conversations sell regions of embedding space. The scoring rule $\log(b) - \|x - c\|^2/\sigma^2$ has circulated as a heuristic for clearing them: embed the conversation, rank advertisers by bid-adjusted proximity, highest score wins. We show the rule is no heuristic. It is VCG. We prove in Lean 4, with zero `sorry`, that the score is a monotone transform of Gaussian advertiser value. The argmax allocation therefore maximizes welfare at every query point, Clarke pivot payments make truthful reporting of bid, center, and reach a dominant strategy, and no allocation rule achieves higher expected welfare. The allocation is a power diagram: in the embedding space itself when reaches are equal, and one dimension up via Aurenhammer's paraboloid lift when they are heterogeneous. At any query point where centers coincide, the mechanism is exactly Vickrey's second-price auction, so keyword auctions are the degenerate case. The proof chain assumes one non-definitional axiom, monotonicity of integration: no dimension bound, no distributional assumption, no bound on the number of bidders. The formalization is archived at DOI 10.5281/zenodo.21214697.

1 One-Shot Bidding

Keyword auctions are strategically broken, and the industry monetizes the breakage. Edelman, Ostrovsky and Schwarz (2007) proved that Google's Generalized Second-Price auction has no dominant-strategy equilibrium: your optimal bid depends on bids you cannot see. First-price display auctions require shading by an amount that also depends on bids you cannot see. The response was an autobidding industry, agents running machine learning against each other to approximate what Vickrey (1961) made exact sixty-five years ago: report your value, pay the externality, done.

Embedding-space advertising restarts the design problem from zero, a chance to get the incentives right on day one. The setting: a user's conversation embeds to a point `x` in a real inner product space; each advertiser declares a center `c` (who their customer is), a reach σ (how wide a neighborhood they serve), and a bid `b` (what a conversion is worth). The platform scores each advertiser at `x` and the highest score wins. The scoring rule was proposed in a blog series, an open-source exchange implements it, and multi-agent simulations probe its market dynamics.[1] What was missing is a proof that the mechanism deserves the trust the proposal asks for.

Here we supply the proof, in Lean, so that trust reduces to running a build command. The contribution is one bridge lemma, `score_eq_log_reportedVal`: the scoring rule is the logarithm of the value a report implies, unconditionally. Everything downstream is the classical VCG argument of Vickrey (1961), Clarke (1971), and Groves (1973), executed formally. On top of the chain we prove two geometric bookends: the

[1] Proposal: june.kim/power-diagrams-ad-auctions. Implementation: github.com/kimjune01/openauction. Simulations: june.kim/relocation-fee-dividend.

allocation is a power diagram for arbitrary heterogeneous reaches, and the mechanism collapses to Vickrey's sealed-bid second-price auction at any keyword point.

2 Model

An advertiser's report is a triple (c, σ, b) with $\sigma, b > 0$. Their private valuation is a triple (c, σ^*, v) of the same shape. True value at query x is Gaussian in distance:

$$\text{trueVal}(x) = v \cdot \exp(-\|x - c\|^2/\sigma^{*2})$$

True value reads as margin times conversion probability, and the Gaussian is the maximum-entropy model of a conversion curve known only by its peak and its width; any other decay imports structure the advertiser has no evidence for. The platform scores reports by

$$\text{score}(x) = \log(b) - \|x - c\|^2/\sigma^2$$

and allocates each query to an argmax of score. The winner pays the Clarke pivot, the externality their presence imposes on the rest. With a single winner per query it has a closed form: losers pay nothing, and the winner pays the runner-up's implied value (zero when unopposed).

$$\text{payment}_w(x) = \max_{j \neq w} b_j \cdot \exp(-\|x - c_j\|^2/\sigma_j^2)$$

Utility is quasilinear. All definitions follow Nisan's mechanism-design chapter in Nisan, Roughgarden, Tardos and Vazirani, eds. (2007).

The scoring rule is the embedding-space member of a known family. Lahaie and Pennock (2007) analyzed keyword scoring rules of the form bid $\times$ quality^s. Ranking by the score written in base β is ranking by $b \cdot q^{\ln \beta}$ with quality $q = \exp(-\|x - c\|^2/\sigma^2)$, so the log base is the squashing parameter, $s = \ln \beta$, and the squashing-parameter literature transfers intact.[2]

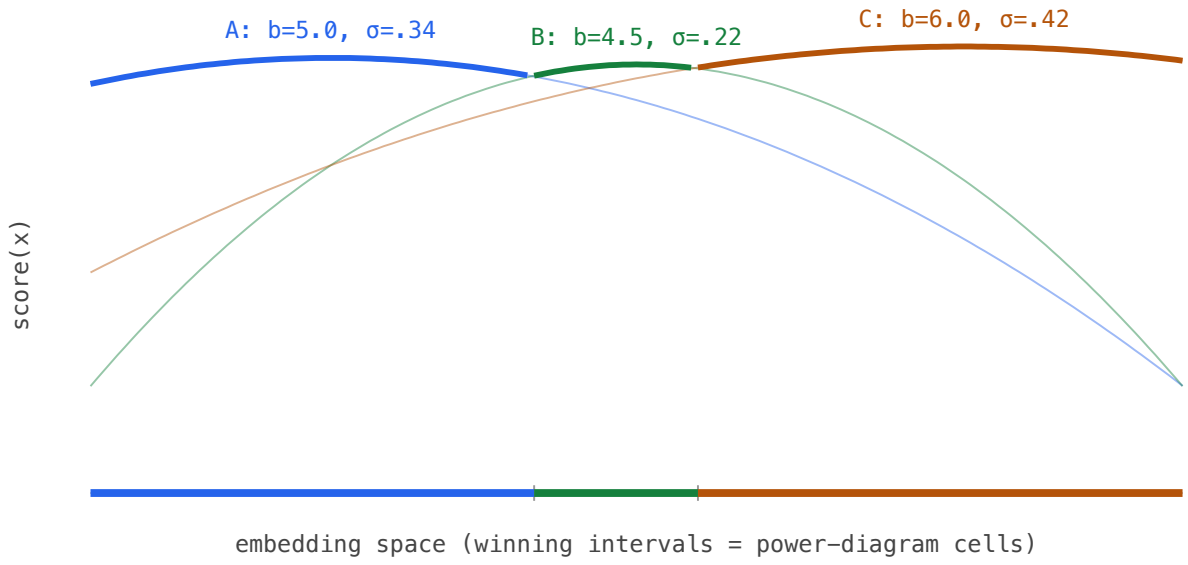

Figure 1: The allocation is the upper envelope of score parabolas. Winning intervals along the axis are the power-diagram cells; a higher bid raises a parabola, a wider σ flattens it.

The formalization states everything over an arbitrary real inner product space. No finite-dimension hypothesis appears anywhere in the chain: the theorems hold for a 384-dimensional sentence embedding and for an infinite-dimensional feature space with equal indifference.

[2]The sweep of the mapping, with the revenue-relevance tradeoff curves: june.kim/the-price-of-relevance.

3 The Bridge

The mechanism-design payload hangs on one identity. Define the value implied by a report, $\text{reportedVal}(x) = b \cdot \exp(-\|x - c\|^2/\sigma^2)$. Then

$$\text{score}(x) = \log\left(\text{reportedVal}(x)\right)$$

with no truthfulness hypothesis (`score_eq_log_reportedVal`). The disguise is a logarithm. Log is monotone, so the argmax of score is the argmax of reported value. The mechanism always maximizes reported welfare, which is the allocation rule VCG requires. When a report is truthful, reported value equals true value (`reportedVal_eq_trueVal_of_truthful`), and the same identity becomes `score = log(trueVal)` (`score_eq_log_trueVal`). The winner is the advertiser who values the impression most (`winner_maximizes_welfare`).

4 Dominant Strategy

`vcg_dsic` is the main incentive theorem: for any deviation report `r'`, a truthful player's utility is at least their deviated utility, regardless of what every other player reports. The Clarke payment is computed from others' reported values only, so a player's report moves their allocation and never their price schedule (`welfareOthersWithout_invariant`); the four-case comparison then closes the argument.

The theorem's scope is stronger than the informal claim usually attached to VCG. Truthfulness here constrains all three report fields: center, reach, and bid. Misreporting your center toward a traffic hotspot is a deviation `r'` like any other, and the theorem says it cannot profit you.

The Hotelling drift that motivates relocation fees is therefore a phenomenon of what this model excludes: budget constraints, volume-dependent objectives, and true value functions outside the Gaussian family. Inside quasilinear IPV, position honesty is free; the empirical case for charging advertisers to move lives entirely in the regime beyond it.[3] The formal result and the market simulations divide the territory between them, and the boundary is the model's assumption list.

5 The Geometry

The name "power diagram auction" is now a theorem at two levels.

5.1 Equal Reach

When two advertisers share σ, comparing scores at `x` is comparing power distances $\|x - c\|^2 - w$ with sites at the centers and weights $w = \sigma^2 \log(b)$ (`score_le_iff_powerDist_le`). The score difference is an affine function of `x` (`score_sub_affine`), so cell boundaries are hyperplanes. That is the classical power diagram of Aurenhammer (1987), with bids setting the weights.

5.2 Heterogeneous Reach

With distinct σs the in-space boundaries curve, and the folk description of the allocation as a power diagram fails in `E`. It succeeds in $E \times \mathbb{R}$. Lift each query to the paraboloid $(x, \|x\|^2)$ and give advertiser `i` a lifted site and weight:

$$\text{site}_i = \left(\sigma_i^{-2} c_i,\ -\tfrac{\sigma_i^{-2}}{2}\right), \qquad w_i = \|\sigma_i^{-2} c_i\|^2 + \tfrac{\sigma_i^{-4}}{4} - \sigma_i^{-2}\|c_i\|^2 + \log b_i$$

Then

$$\text{liftedPowerDist}(x, \|x\|^2) = \|x\|^2 + \|x\|^4 - \text{score}(x)$$

exactly (`liftedPowerDist_paraboloid`), and the leading terms are advertiser-independent. Maximizing score is minimizing lifted power distance (`score_le_iff_liftedPowerDist_ge`), and the auction's winner rule is the lifted diagram's cell assignment (`winner_minimizes_liftedPowerDist`). Aurenhammer proved diagrams of quadratic distance functions lift this way; the formalization instantiates his lift for this scoring rule and checks the algebra once. In practice, the O(log N) spatial-index machinery for power diagrams applies to the variable-σ auction after one dimension of padding.

[3] Relocation fees and their simulated dividend: june.kim/relocation-fees, june.kim/relocation-fee-dividend.

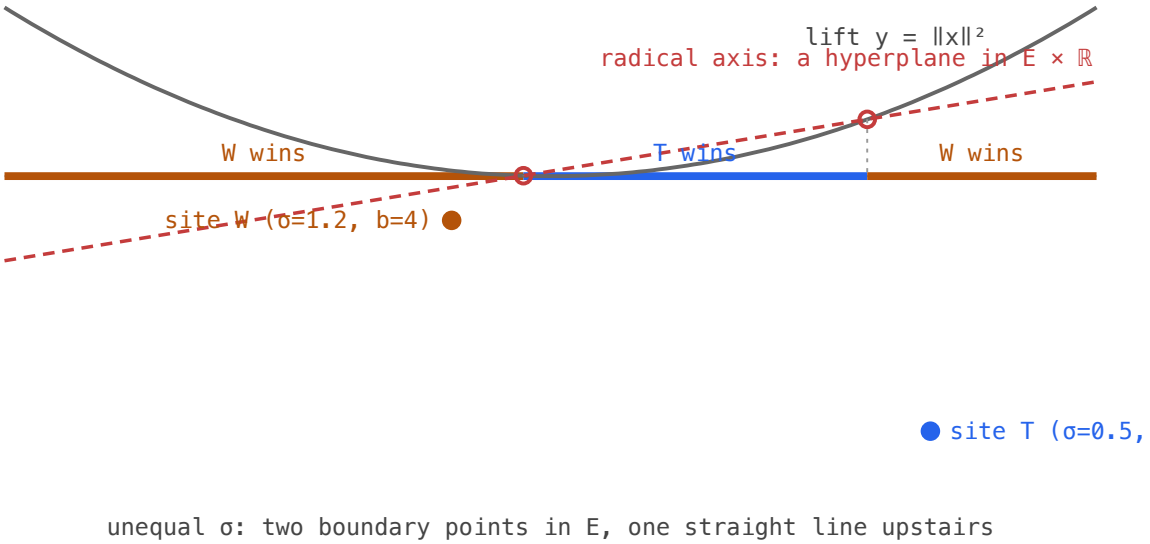

Figure 2: The lift, drawn for a one-dimensional embedding. A tight advertiser T and a wide advertiser W produce two boundary points downstairs (T's cell is the interval between them); upstairs, the boundary is one straight radical axis cutting the paraboloid. Quadric bisectors in `E` are hyperplanes in $E \times \mathbb{R}$.

6 Keywords Recovered

Keywords Are Tiny Circles argued the migration path:[4] keyword auctions are the degenerate case of the embedding auction, so adopting the general mechanism strands no existing buyer. The claim is now a theorem pair.

6.1 The Limit

At any point other than its center, a report's score diverges to $-\infty$ as $\sigma \to 0$, while its score at the center is `log(b)` independent of σ (`keyword_is_degenerate_limit`, `score_at_center`). The tiny circle collapses to its point.

6.2 The Mechanism

At a query point where all centers coincide, the Gaussian factor is $\exp(0) = 1$ for every bidder: reported value is the bid, the winner is a highest bidder (`winner_maximizes_bid_of_common_center`), and the Clarke pivot equals the highest competing bid (`vcgPayment_common_center_second_price`). That is Vickrey's sealed-bid second-price auction, allocation and payment both. A keyword auction is what this mechanism does at a point.

7 Optimality

Pointwise welfare maximization integrates. For any measure over queries, expected welfare under the score-argmax allocation weakly dominates expected welfare under any allocation rule whatsoever (`integral_efficiency`, `gaussian_optimality`). The capstone, `gaussian_vcg_weakly_dominates`, conjoins the three properties: welfare-optimal, dominant-strategy incentive compatible, equilibrium-efficient. The last is the welfare guarantee evaluated at truthful play, which `vcg_strategyproof` certifies as a Nash equilibrium. The artifact separately proves the equilibrium-decomposition theorem of Ghani, Hedges, Winschel and Zahn (2018) (`composed_equilibria_decompose`); whether DSIC itself composes through open games is an open question, and the formalization records it as open instead of assuming it.

The chain uses one non-definitional axiom, `QueryMeasure.integral_mono`: the expectation operator respects pointwise inequality. A finite query log satisfies it outright; a measure-theoretic integral satisfies it on the allocation rules it can evaluate. The optimality theorem inherits its strength from that interface, and keeping the axiom abstract lets the theorem quantify over query distributions instead of fixing one.

[4] june.kim/keywords-are-tiny-circles.

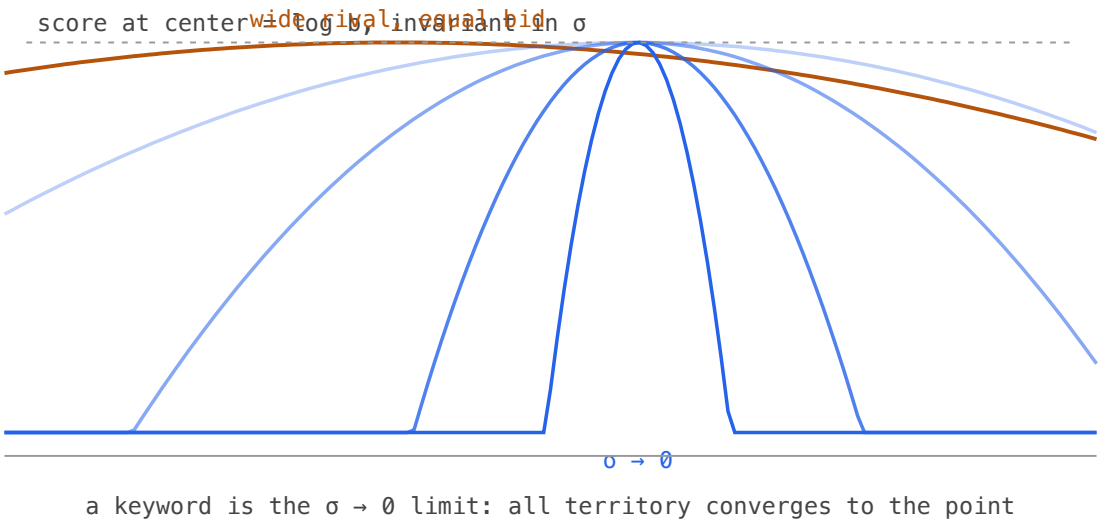

Figure 3: Shrinking σ at a fixed center against a wide rival with equal bid. The score at the center is $\log b$ for every σ, while territory off-center collapses; in the limit the advertiser competes at one point only, on bid alone.

8 Related Work

8.1 LLM Ad Auctions

Soumalias, Curry and Seuken (2025) give the strongest prior mechanism for LLM advertising. Their MOSAIC auction allocates the reply itself: candidate responses sampled from the model are scored by advertiser reward functions, one is selected by softmax, and dominant-strategy incentives come through Rochet (1987) payments, because exact VCG is intractable over the space of token sequences. Auctioning an embedded intent point instead of the reply shrinks the outcome space from token sequences to N known advertisers. Exact VCG becomes one argmax plus one counterfactual argmax, and the guarantees become simple enough for Lean to check. Allocating a separate ad object also removes MOSAIC's deployment costs: no M-fold candidate generation per query, deterministic winners an advertiser can budget against, and a discrete placement that can be attributed and priced.

Hajiaghayi, Lahaie, Rezaei and Shin (2024) place the auction at retrieval time instead: ads are probabilistically retrieved per discourse segment by bid and relevance, with incentive-compatible pricing. The outcome space is far smaller than MOSAIC's, but the placement still lives inside the generated text, and the winner is a draw rather than an argmax; the geometric mechanism keeps the retrieval-time simplicity while making the allocation deterministic and the ad a separate object.

The other family bakes advertising into the model by training. Alibaba's LLM-Auction post-trains the model with preference alignment to balance ad revenue against user experience, encoding commercial incentives directly in the weights. Training is the wrong cadence for an ad market: campaigns change hourly, fine-tuning takes weeks, and retraining per campaign rotation costs more than the campaign. It is also the wrong trust surface. Bias learned in weights cannot be audited from outside, and users cannot police it from inside either: evaluators shown chatbot responses with embedded ads failed to detect the ads and preferred the responses that contained them. The mechanism here is the opposite limiting case. The scoring rule is public, the allocation is its argmax, the incentive properties are a compiled theorem, and the audit reduces to a build command.

8.2 Formalized Auctions

Kerber, Lange and Rowat (2016) formalized Vickrey's auction in Isabelle and argued mechanized reasoning should be ordinary practice in economic theory. We extend the practice from a discrete allocation rule to a geometric one: the object verified is a partition of a vector space, and the same artifact checks the auction theorems and the computational-geometry identification.

8.3 Filters and Reserves

Hartline, Hoy and Taggart (2023) show competitive efficiency survives reserve pricing. The same holds trivially here for any pre-auction relevance filter, since welfare maximization restricted to a nonempty eligible set is welfare maximization on that set (`tau_preserves_efficiency_among_eligible`).

9 Limits

The theorems end where the model does, and the model is deliberately narrow.

Element	In the formalization
Embedding space `E`	any real inner product space; dimension unconstrained, infinite allowed
Advertisers	arbitrary finite type; no bound on the count
Query distribution	universally quantified expectation operator; one axiom, monotonicity
Utility	quasilinear; no budget constraints
True valuations	isotropic Gaussian family
Deviations	any (center, σ, bid) triple
Rival allocation rules	any rule the expectation operator evaluates
Winners per query	exactly one; ties broken by fixed enumeration

Single winner. Slates, pacing, and cross-impression externalities are outside the model.

No budgets. Budget-constrained clearing has known structure (with fixed budgets it is semi-discrete optimal transport, whose solution is again a power diagram), but its incentives and dynamics are open. The simulations suggest the dynamics are where the difficulty lives.

Gaussian truth. The Gaussian family is a bidding language, and every deployed auction clears one: the reigning language is the keyword, the $\sigma \to 0$ point mass of this same family. What a bidding language buys is expressiveness against clearing cost, the central tradeoff of the combinatorial-auction literature (Nisan, 2000). On that frontier the isotropic Gaussian is the most expressive language currently known to admit sub-linear clearing, exact VCG, and a compiled proof. Anisotropic preferences (elliptical rather than spherical reach) stay computable at O(N) per query but break the bridge lemma as stated; whether a lifted variant survives for quadratic-form preferences is the natural next theorem, since Aurenhammer's lift already accommodates general quadrics. Mixtures turn the score into a log-sum-exp whose bisectors admit no fixed-dimension lift, and the clearing speed and the proof leave with the geometry.

Statics only. Nothing here says bidding dynamics converge. The relocation-fee simulations[5] study the dynamics empirically; a formal convergence or limit-cycle result would be a separate paper's contribution.

Artifact

The formalization is at github.com/kimjune01/auction-proof, archived as DOI 10.5281/zenodo.21214697, AGPL-3.0, Lean 4 (v4.29.0-rc6) with Mathlib. Verification:

```
lake exe cache get
lake build
```

Zero `sorry`. The claims-to-theorems map:

Claim	Lean theorem	File
score = log(reported value)	`score_eq_log_reportedVal`	Efficiency.lean
winner maximizes true welfare	`winner_maximizes_welfare`	Efficiency.lean
truthful reporting is dominant	`vcg_dsic`	Strategyproof.lean
no allocation rule beats VCG	`gaussian_optimality`	GaussianOptimality.lean
capstone conjunction	`gaussian_vcg_weakly_dominates`	GaussianOptimality.lean
equal-σ power diagram, hyperplane bisectors	`score_le_iff_powerDist_le`, `score_sub_affine`	PowerDiagram.lean

[5] Relocation fees and their simulated dividend: june.kim/relocation-fees, june.kim/relocation-fee-dividend.

Claim	Lean theorem	File
variable-σ power diagram via paraboloid lift	`liftedPowerDist_paraboloid`, `winner_minimizes_liftedPowerDist`	PowerDiagram.lean
keyword limit	`keyword_is_degenerate_limit`	VectorSpace.lean
exact Vickrey at a keyword point	`vcgPayment_common_center_second_price`	SecondPrice.lean

Acknowledgments

Conversations with Sébastien Lahaie and Mohammad Hajiaghayi sharpened the mechanism-design framing, and Sébastien pointed me to MOSAIC. Errors, and the claims, are mine.

Disclosures

LLM use. This paper and its artifact were produced with large language models via Claude Code. The author directed the research and reviewed every claim; agents wrote the Lean proofs, generated the figures, and drafted the prose, and a separate model ran an adversarial review of the prose against the formalization. No guarantee in this paper rests on any model's judgment: the artifact type-checks with zero `sorry`, and verification reduces to `lake build`.

Funding. Self-funded independent research. No external funding, no employer direction, no advertiser or platform relationship. The author maintains the open-source exchange implementation cited above.

Part of the Vector Space series.