1-Bit Is All You Need

Quantization Doesn't Destroy Meaning: It Reveals Its Architecture

Andy Grossberg · Waving Cat Learning Systems · May 2026 · andy.grossberg@gmail.com

A plain-English companion to *Type-Conditioned Action on Discrete Positions: Sign-Bit Semantic Arithmetic Reveals Substrate-Native Relationship Taxonomy* (paper forthcoming).

Abstract

Continuous-vector embedding arithmetic (`king - man + woman ≈ queen` per Mikolov et al., 2013) has shaped a decade of distributed-semantics work around the assumption that high-precision real-valued spaces are necessary for the directional structure that algebraic semantic operations depend on. **We show this assumption is wrong.**

Reducing 768-dimensional Nomic v1.5 float embeddings to a single sign bit per dimension (literally *"is this neuron firing positively or negatively?"*) preserves analogy retrieval at **24,926× above random baseline** at top-10 against a 370,000-word public-domain English vocabulary, with all operations expressible as XOR and Hamming distance over a 96-byte-per-word representation. Continuous-vector cosine retrieval (3CosAdd) on the same vocabulary reaches 97.96% top-10; sign-bit reduction loses ~30 percentage points of absolute accuracy but only **1.45×** in signal-to-baseline ratio (36,255× → 24,926×). Population-level direction extraction (the architectural primitive that *only the discrete substrate makes visible*) closes **50 of the 56 percentage points** of the gap on the hardest analogy class (capital-country: 44% single-pair → 94% majority-direction → 100% continuous).

The substrate extends from analogy to next-word sequence prediction (≥56× above random across context sizes N=2..6 on natural-language corpora the encoder was not trained on), and to cross-corpus transfer: a **6×6 cross-text matrix** spanning 1611-1922 (Russell's *The Analysis of Mind*, Shelley's *Frankenstein*, Joyce's *Ulysses*, Carroll's *Alice*, KJV Bible, Twain's *Huck Finn*) produces 25-322× signal in all 30 cross-text cells, with asymmetric patterns recovering exactly the directional distributional-distance one would predict from corpus size and register.

Seven hypotheses (*See Appendix A below*) were pre-registered before the experiments that test them: H1/H2/H3 validated as filed; four positional-composition predictions (H-pos-1 through H-pos-4) resolved at 20,000 samples per cell, including the **load-bearing H-pos-2 falsification** that sharpened the next research question (find a non-commutative composition primitive that respects the substrate's vocabulary manifold).

The architectural thesis these results support: **distributed semantic representation is not continuous-space geometry, it is the *type-conditioned action* a discrete substrate**

performs on pairs and sequences of positions. Three relationship types, three operations, dispatched by reading the substrate's own bit statistics. No learned classifier. No continuous projection. No fine-tuning. No language model in the loop. We refer to the resulting representation as the **1-bit substrate**.

We don't need a bigger boat; we need a smarter crew.

---

1. The Claim: Why It Sounds Wrong

Continuous-vector embeddings are the workhorse of modern NLP. They use 32-bit floats (typically 768 to 1,024 floats per word) to position concepts in a high-dimensional space where semantic relationships become geometric. The `king - man + woman ≈ queen` result (Mikolov et al., 2013) is the textbook demonstration: meaning has a *direction* in this space, and adding/subtracting directions yields semantically meaningful results.

When researchers want to compress these representations, whether for retrieval speed, for storage, for edge devices, etc., they run into a perceived ceiling. Quantization to 16-bit, 8-bit, even 4-bit floats: fine for retrieval, but degrading. Quantization to a single bit? "Almost certainly destroys the directional structure." That's the standing assumption.

Our finding: the standing assumption is wrong. **One bit per dimension preserves enough relational structure to support the same algebraic semantic operations the field uses 32-bit floats for.** But at 32× the storage compression, with operations expressible entirely as XOR and Hamming distance (the two cheapest things a CPU or 1-bit FPGA can do).

That sounds like a 32× free lunch. It is mostly free, but not entirely. The honest trade-off is in §3 below.

---

2. What the 1-Bit Substrate Actually Is

A 30-second mental model:

1. Run a standard word through any pretrained text encoder. We use `nomic-embed-text-v1.5`. You get a 768-dimensional vector of floats.
2. Throw away the magnitude. Keep only the *sign* of each dimension. Per dimension: `1` if positive, `0` if negative.
3. Pack the result into 96 bytes (= 768 bits).
4. Do this once per word in your vocabulary. Save the resulting cache.

That's the substrate. No fine-tuning. No learning beyond what the encoder already did. No per-task adaptation. **Every operation in the paper is XOR (bit-wise difference) and Hamming distance (count of differing bits).** That's it.

Scope of the "no language model" claim. The encoder is a pretrained language-model-shaped network, and its training is the load-bearing source of the substrate's relational

structure. The contribution this paper measures is what survives 1-bit reduction of that encoder's output. When we say "no language model in the loop" in later sections, that refers to the *inference and dispatch paths*: at query time, no learned model operates, only XOR and Hamming distance against the cached substrate. The substrate itself is constructed once from the pretrained encoder; that one-time construction is the only place an LM-shaped model touches the pipeline.

Why does that capture semantic structure? Continuous embeddings spread information across two channels: the *sign* of each dimension (which side of the origin) and the *magnitude* (how far). For analogy operations specifically, the sign channel (the directional information) is what carries the semantic relationship. Magnitude encodes salience and frequency, which matter for nuance but appear secondary for the question of *which other word does this point at?*

How this differs from prior bit-arithmetic representation work. Hyperdimensional computing (Kanerva, 2009) and vector-symbolic architectures (Plate, 1995) build bit-arithmetic operations as *design primitives*, with representations constructed from scratch using those operations. This paper reports a complementary result: what survives 1-bit quantization of *pretrained continuous embeddings*, plus the three-class relationship taxonomy the discrete substrate exposes from those embeddings. The novelty isn't bit-arithmetic per se; it's that the structure required for distributed semantic operations is already present in 1-bit reduced pretrained embeddings, and the discrete substrate makes a taxonomy visible that the continuous space buries.

Sign-bit reduction discards the magnitude channel and keeps the direction channel intact. That's the 32× compression. The cost is bounded; the operations stay tractable on hardware that can't do float math at all.

---

3. What Survives Quantization (and the Honest Trade-off)

We tested the substrate on the standard analogy benchmark (49 word analogies across 5 categories: plurals, verb past-tense, comparatives, gender/royalty, capital/country) at three vocabulary scales:

Vocabulary	Method	Top-10 accuracy	Random baseline	Signal/baseline
472 (curated)	Sign-bit XOR	87.8%	2.12%	41×
10,031 (common-words)	Sign-bit XOR	71.4%	0.10%	716×
370,106 (full dictionary)	**Sign-bit XOR**	**67.4%**	**0.0027%**	**24,926×**
370,106 (full dictionary)	Continuous cosine (3CosAdd)	97.96%	0.0027%	36,255×

Two things to notice:

The signal-to-baseline ratio strengthens monotonically with scale. As the haystack grows from 472 words to 370,000, the absolute accuracy declines only modestly (87.8% → 67.4%) but the random baseline collapses. The ratio of "how much better than chance" goes 41× → 716× → **24,926×**. This is the *opposite* of what a curation-artifact result would do (a curation artifact gets easier to beat the bigger the haystack). Result is structural.

Continuous-vector arithmetic still wins on raw accuracy. 97.96% vs 67.4% top-10 at the full-dictionary scale, a 30 percentage-point gap. Sign-bit reduction loses real information. The signal-to-baseline ratio gap is much smaller (1.45×: 36,255× vs 24,926×), but the absolute-accuracy gap is real.

That's the honest trade-off. Sign-bit substrate is a *different optimization target*: storage compression, sub-watt cognitive primitives, 1-bit-hardware mappability. For applications where maximum analogy accuracy on cloud-scale GPUs is the optimization target, continuous-vector cosine retrieval is the better choice. For applications where the priority is "how much intelligence can I stand up on a 96-bytes-per-word cache that runs on a CPU or FPGA without ever calling out to the cloud," the substrate is the better choice, *especially* once we add the architectural primitive in §4.

---

4. The Architectural Payoff: Three Kinds of Meaning, Three Operations

This is the part of the paper we expect to last. The empirical result above (24,926× sign-bit, 97.96% continuous) is impressive on its own. But the discrete substrate exposes structure that the continuous space *buries*.

When you look at the bit-level statistics of how pairs of words relate to each other under XOR, semantic relationships partition cleanly into **three substrate-readable classes**, each with a characteristic bit-flip-frequency signature:

Relationship type	Example	Bit signature	What the substrate does
Shared-axis	capital-of, plural-of, past-tense-of	Many pairs share a coherent XOR direction; bit-flip histogram is bimodal with a strong "always-flipped" tail	Apply the **population-level direction** (extract the majority direction across many examples, XOR it onto the query)
Per-pair-distinct-axis	antonyms (hot↔cold, alive↔dead, open↔closed)	Polarity bits exist but distribute differently across pairs (each antonym pair lives on its own sub-axis)	**Drill down to the right sub-axis**, then apply only that sub-axis's direction
Cluster-without-axis	synonyms (walk/stroll/run/jog)	Zero coherent direction-bits; dense	**Retrieve the cluster region** (Hamming-

Relationship type	Example	Bit signature	What the substrate does
		category-membership bits	nearest neighbors); do not try to apply a direction that doesn't exist

The substrate **self-classifies** which class a relationship belongs to. There's no need for an external classifier, no learned dispatcher at all, just look at the bit statistics. *Synonyms produce 0/768 always-flipped bits and 523/768 always-unflipped bits.* That's the cleanest possible "this is a cluster, not a direction" signature. *Capital-country pairs produce a strong always-flipped tail* which is the signature of a shared direction.

The headline architectural result: when we apply population-level direction extraction (the operation the substrate dispatches for shared-axis relationships) to the hardest analogy class (capital-country), top-10 accuracy goes from 44% (single-pair XOR) to **94%** at full-dictionary scale. Mean rank drops from 460.9 to 13.8. **That closes 50 of the 56 percentage points of the gap with continuous-vector retrieval on this class.**

The architectural payload (taxonomy detection plus population-level direction extraction) is *exactly* what makes the discrete substrate competitive with continuous-vector retrieval on shared-axis relationships. Without the architecture, the substrate is meaningfully worse than continuous. With it, the gap shrinks to 6 points on the hardest class, with 32× less storage and bit-only operations.

This is the part of the paper that's actually a thesis: **distributed semantic representation is not continuous-space geometry; it is the *type-conditioned action* a discrete substrate performs on pairs of positions.** Three operations make up the action vocabulary. The substrate dispatches based on the relationship type, which it reads from its own bits.

---

5. From Analogies to Sequences: Substrate-Based Next-Word Prediction

If the substrate is going to be useful for anything language-shaped, it has to handle sequences, not just isolated word relationships. We tested it on next-word prediction over two single-author corpora the encoder was *not trained on*: Bertrand Russell's *The Analysis of Mind* (1921) and Mary Shelley's *Frankenstein* (1818). For each context window of N words (N from 2 through 6), encode the context as XOR composition of the previous N word patterns, then retrieve the next word.

Two retrieval strategies:

- **Strategy A: substrate-native**: Look up the next word as the entry whose 768-bit pattern is Hamming-closest to the context vector, against the full 370,000-word vocabulary.

- **Strategy B: lookup-table**: Store the training-set (context, next-word) pairs. For a test context, find the K=20 nearest training contexts by Hamming distance, vote over the next-words.

Strategy A reveals a parity oscillation: odd-N contexts (3, 5) give substantial signal-to-baseline at top-10 (Russell N=3: 861×, Russell N=5: 635×; Frankenstein N=3: 682×, N=5: 320× and all at 20,000 held-out samples). Even-N contexts (2, 4, 6) give zero signal. The mechanism: adjacent words in language have semantically correlated embeddings (bigram structure makes consecutive words share more sign bits than two random words would). XOR of bigram-correlated pairs mostly cancels. At even N, full pairs of bigram cancellations leave near-zero residual context. At odd N, one residual word survives with predictive content.

Initially we framed this as a "cliff at N=4." With N=5 and N=6 added at scale, the deeper truth is *parity*, not a cliff.

Strategy B works at every N from 2 to 6, ranging from 56× to 145× signal-to-baseline at top-10 across both corpora, declining monotonically with N (longer contexts have fewer training-set matches to retrieve against so it's a context sparsity, not commutativity bottleneck). At N=6 with no learned weights, the substrate-based lookup table predicts the next word at 56× over chance which is non-trivial for 6-gram prediction with no language model in the loop.

Cross-corpus portability: the 6×6 matrix. A Strategy B lookup-table trained on Russell and tested against Frankenstein retains 49% of the within-corpus signal. Frankenstein → Russell retains 74%. Extended into a 6×6 matrix spanning **1611 to 1922**, six radically different registers — sacred archaic English (KJV Bible), high modernism (Joyce's *Ulysses*), gothic narrative (Shelley's *Frankenstein*), analytic philosophy (Russell's *Analysis of Mind*), children's absurdism (Carroll's *Alice in Wonderland*), and vernacular American (Twain's *Huckleberry Finn*) — all 30 cross-text cells produce signal between **25× and 322× above random baseline** at top-10. *Every cross-corpus pair*, including the maximally distant ones (Russell ↔ Alice, KJV ↔ Joyce), clears the noise floor by an order of magnitude.

The patterns in the matrix recover exactly what you'd predict from distributional distance:

- **Joyce's *Ulysses* is the universal donor:** 253-322× cross-text to every other corpus. Big corpus + extremely wide modernist vocabulary lying near the dense middle of the encoder's training distribution = it covers everyone else's territory.
- **Alice in Wonderland is the weakest donor:** 25-34× to everything. Small corpus, narrow children's-nonsense register, can't predict words it never saw.
- **The matrix is asymmetric: and the asymmetries are interpretable.** Russell → Alice is 48×; Alice → Russell is 26×. Russell's analytic-philosophy vocabulary subsumes Alice's children's vocabulary better than the reverse. KJV → Russell is 113×; Russell → KJV is 57×. Same pattern: the larger or broader corpus contains more of the smaller's territory than the reverse.
- **Donor strength scales primarily with corpus size, secondarily with register similarity.** Ulysses (huge, modern) donates 253-322×; KJV (huge, archaic) donates

98-125×. Both huge, but Ulysses' modern register lies closer to where the encoder was trained, so it donates ~2× better.

(There's a separate observation, filed for future vision-model follow-up: cross-text retention is operationally a *surprise* signal that Russell → Alice at 48× / 26× is the substrate quantitatively registering, "Alice is more distributionally surprising relative to analytic philosophy than other Victorian English is." When equivalent vision substrates run on art instead of text, Surrealism and Dada should produce the same low-retention / high-surprise signature against representational realism baselines.)

The cross-text matrix is, operationally, **a substrate-native quantitative metric of literary distributional distance.** Higher cell value = closer in distributional space; lower = further. The asymmetries are the substrate registering that distributional distance is not symmetric (bigger corpora contain more of smaller ones than the reverse) without any learned distance metric, labeled training, or auxiliary classifier. *The substrate gets the structure right because the structure is already in the data; we just gave it a way to read.*

---

6. Pre-Registered Honesty: The Falsification That Sharpened the Question

Hypotheses filed before the experiments that test them are a methodological response to a fair worry: "the test bench was tuned to produce the result." If you write your prediction in a dated document with a falsification protocol, *then* run the experiment, the result is harder to dismiss as cherry-picked.

We pre-registered seven hypotheses (H1, H2, H3, H-pos-1 through H-pos-4) before the corresponding experimental work. H1 (predictively-graded substrate energy trajectories) is queued for follow-up. H2 (novel-but-similar is easier than novel-and-distant) was confirmed across all three vocabulary scales. H3 (composite-fact analogies need population-level direction extraction) was validated at 10K-vocabulary and *strengthened* at 370K, so the +50pp lift at full dictionary is bigger than the +41.8pp lift at 10K, the opposite of "small-set effect that washes out at scale."

But H-pos-2, the load-bearing prediction that breaking commutativity at the bit level (via per-position cyclic rotation) would recover Strategy A substrate-native sequence prediction at higher N, **was falsified.** Cyclic rotation produces a vector that doesn't correspond to any actual word in the vocabulary; it lives "off the manifold" of substrate-encoded patterns, and Strategy A's nearest-Hamming retrieval has no real word near the rotated combination. Across all N tested at 20,000 samples per cell, Strategy A with rotation produced 0× signal-to-baseline.

That sounds like a setback. It isn't. Two things made the falsification actually informative:

1. **The same primitive lifted Strategy B by 35-71% at N≥4.** Strategy B's K-NN over training contexts doesn't require manifold preservation; it just requires that contexts which share predictive structure end up Hamming-close. Non-commutative composition makes those distances more discriminative. At N=6 with positional composition, Strategy B reaches 79.8-82.7× over random.

2. **The architectural diagnosis sharpened the next research question.** Was: "try positional composition for Strategy A." Now: "find a non-commutative composition primitive that *respects the substrate's vocabulary manifold*." The candidate space narrowed: per-position XOR with a fixed pseudorandom mask family (XOR is its own inverse, so manifold-preserving by construction), parity-encoding within fixed bit-windows, learned positional embeddings under XOR. Substrate-native sequence prediction at N≥4 is *open*, not solved. And we know exactly what shape the solution needs to have.

Pre-registered falsification beats post-hoc rationalization, even when (especially when) the prediction was wrong.

---

7. What This Unlocks

If the architectural thesis holds, several things become operational:

Sub-watt cognitive primitives on commodity hardware. 96 bytes per word, all operations expressible as XOR and population count, no learned classifier on the dispatch path. The substrate maps directly onto 1-bit FPGA / ASIC inference paths. Sub-watt cognitive operations on devices that can't fit a transformer in their power budget: phones, wearables, hearing aids, scanner-side hospital deployments, automotive edge, satellites.

Cross-modality transfer (Phase 1 supports; Phase 3 queued). The architectural thesis predicts that the relationship-type taxonomy is a property of distributed semantic representation in general rather than language specifically. Vision encoders (CONCH, UNI, **Virchow2**, Phikon-v2) trained on histopathology images produce continuous vectors with the same shape as language embeddings; sign-bit reduction should preserve the same relational structure on imagery. Phase 1 results from our histopathology pilot (**93,522 NCT-CRC-HE 100K patches including 11,111 cancers, four-encoder consensus, 2026-05-23—see Histopathology Pilot Guide, May 2026; formal paper forthcoming**) **achieved 100% tumor sensitivity and 100% release-tier specificity in leave-one-out retrieval evaluation**, with a fourth taxonomy class (*stream/graded*) emerging in pathology data that didn't appear in the language taxonomy. Phase 3 (PANDA prostate ISUP grades) is queued.

Retrieval-first language-model architectures. Transformer attention recomputes query-key similarity across every token pair, every layer, every forward pass and that's quadratic work per inference, every time. A retrieval-first language model architecture (working name: *Lattice-LM*) does the work *once* at substrate construction and amortizes across all subsequent retrieval. The bottleneck for such architectures has been whether the substrate carries enough relational structure to support what a language model needs. The results in this paper say: analogy retrieval at 24,926× over random, sequence prediction at ≥56× over random across N=2..6, cross-corpus transfer at 50-80% retention. These are *floor results* without any learning on top of the substrate. They don't constitute a working language model; they constitute the empirical foundation that says the floor is high enough to build one.

On-prem deployment without a cloud dependency. The substrate's deployment requirement is "store the cart, evaluate Hamming distances on a CPU." Hospital on-prem at the scanner, on-device on a phone, in a vehicle without connectivity, in a hearing aid, in a hospital basement air-gapped from the internet for PHI compliance. All these deployments become tractable. The substrate doesn't need a GPU and never needs to leave the building.

---

8. The Trajectory: Where This Sits in the Bigger Picture

The current paper is deliberately the *floor.* We're examining the substrate's performance with **no learning on top of it whatsoever.** The encoder is pretrained off-the-shelf and frozen. Vocabulary is encoded once, sign-bit-reduced, frozen. All operations are non-learned bit arithmetic.

That's the most important claim. *The substrate works at 24,926× analogy and ≥56× sequence prediction without any of the things you'd expect to need:* no language model, no fine-tuning, no learned weights, no per-task adaptation. The architectural floor is high enough to build on.

Three increasingly substantial levels of learning are queued for follow-up work, each likely to amplify what we've measured:

1. **Hebbian co-occurrence weights.** Track which patterns co-occur in training contexts; strengthen the links between them. Pure substrate-side learning, no encoder change. Expected boost on Strategy A sequence prediction pending the encoding work that resolves how Nomic-row-fill patterns interact with the lattice's attractor dynamics (Our recent finding: HDC-style random projection encoding measurably improves the substrate's geometry preservation; SDR encoding into lattice cells is the next set of experiments).
2. **Fine-tune the encoder for next-word prediction.** Standard pretraining objective: predict the next word given the context. The encoder learns to produce embeddings that XOR-compose well for sequence prediction. Likely 50-100× boost; comparable to a small transformer at orders-of-magnitude less inference compute.
3. **Full lattice physics with weights and settling dynamics.** Each pattern stored not just as a 768-bit vector, but with co-activation weights to neighbors. Settling dynamics use those weights to converge on attractors. This is the *Lattice-LM* cognition-engine architecture, pending the encoding work that closes the row-fill-vs-SDR-vs-HDC question (preliminary findings from our recent physics-verification experiments). Pre-registered hypothesis H1 (predictively-graded energy trajectories) becomes testable, the predictive-coding cascade becomes operational, the cortical-stack architecture comes online, all conditional on the encoding question resolving.

The current paper says *the floor is high*. The next paper says *here's what happens when you put the right learning back in*. The floor result is what makes the trajectory credible: if a substrate this minimal already works at 24,926× analogy and 56-145× sequence

prediction, the architectural arc is more about engineering than about hoping for a research breakthrough.

---

9. Honest Caveats

What we haven't shown, and what would falsify the work:

- **Encoder dependence.** All results inherit Nomic v1.5's encoder quality. The substrate's *native* arithmetic capability is what's tested; absolute accuracy numbers would shift under a different encoder. The architectural thesis predicts the three-class taxonomy reproduces across encoders. That's a falsifiable prediction, queued (Experiment 06c, multi-encoder transfer).
- **Single deterministic run.** The 49 analogy tests are deterministic given a fixed encoder and vocabulary. Confidence intervals would require multiple vocabulary samples and bootstrap. Signal-to-baseline magnitudes are large enough that statistical significance isn't in question for the headline claims, but per-category numbers should be read as point estimates.
- **H1 not yet tested.** The pre-registered "substrate energy trajectories are predictively-graded" hypothesis requires energy-trajectory instrumentation in the broader substrate, beyond the scope of this paper.
- **Vision modality is partial.** Phase 1 of the histopathology pilot (**NCT-CRC-HE 100K, 93,522 patches, four encoders**) preserved 100% tumor sensitivity and 100% release-tier specificity giving strong evidence the architecture transfers across encoders within a modality. Phase 3 (PANDA prostate ISUP) is queued and will test the harder cross-task generalization.
- **Sequence prediction at high N is open.** Strategy A at N≥4 is bottlenecked by XOR commutativity; cyclic rotation is the wrong fix. The sharpened research question (§6.9 in the formal paper) names three candidate primitives. The answer isn't in the paper yet.
- **Substrate-native ≠ communication-grade.** What we've shown is that the substrate carries enough relational structure to support cognitive operations. What we haven't shown is that it can engage in communication-grade generation (intent activation, drift detection, listener modeling, satisfaction loop closure). Those are the ingredients of *communication*, distinct from the substrate-floor result. Queued for the cognition-engine work that follows.

About This Document

- **Architectural claims, hypotheses, theses** are Andy Grossberg's. The thesis structure, the relationship-type taxonomy, the pre-registered hypothesis discipline, and the architectural framing are first-person attributable.
- **Experiment scripts, optimization passes, results extraction, and draft text** were executed in collaboration with Claude Opus 4.7 as research participant. This division is consistent with the project's standing voice guidelines.

Takeaways

> *"We don't need a bigger boat; we need a smarter crew."*

The architectural-efficiency thesis in one line. The transformer scaling-law trajectory expects ever-larger compute budgets to substitute for ever-cleverer primitives. Biology doesn't make that trade. Neither do we.

> *"Nature doesn't waste space, it wastes time. And it's got all the time in the universe to waste."*

That's the trade-off the substrate makes. Encoder pretraining is one-time, expensive, amortized. Substrate construction is one-time, cheap, frozen. Inference is O(1) Hamming retrieval against a 96-bytes-per-word cache. The substrate spends time once, in the past, and saves space and energy forever after, which is exactly the optimization biology has been running for four billion years.

> *"1-Bit Is All You Need."*

The pun on Vaswani et al. (2017) was promoted from background joke to working title of this paper. The transformer paper's "attention is all you need" is what made the field's compute trajectory inevitable. The architectural counter-argument we're putting on the cover: at the floor, before anyone trains anything, our substrate already carries 24,926× signal. The right primitive matters more than the scale.

Further Reading & Sources

For the fully curated bibliography a link will be available on www.wavingcat.dev although here are the works cited inline in this litepaper:

Continuous-vector semantic arithmetic (the standing assumption being challenged)

- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv:1301.3781. — The original word2vec analogy benchmark; the `king - man + woman ≈ queen` result.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality.* NeurIPS 2013. — Companion paper formalizing the analogy methodology this paper inherits.
- Pennington, J., Socher, R., Manning, C. (2014). *GloVe: Global Vectors for Word Representation.* EMNLP 2014. — Alternative continuous-vector method; same algebraic-semantics regime.

The encoder this paper uses

- Nussbaum, Z., et al. (2024). *Nomic Embed: Training a Reproducible Long Context Text Embedder.* — `nomic-embed-text-v1.5`, the 768-dim encoder we sign-bit-reduce.
- Reimers, N., Gurevych, I. (2019). *Sentence-BERT.* EMNLP 2019. — The `sentence-transformers` library through which the encoder is loaded.

Adjacent prior work in discrete / hyperdimensional representation

- Kanerva, P. (2009). *Hyperdimensional Computing: An Introduction.* Cognitive Computation. — The closest prior tradition; HDC builds bit-arithmetic operations as design choices, where this paper reports what survives in pretrained continuous embeddings under 1-bit quantization.
- Plate, T.A. (1995). *Holographic Reduced Representations.* IEEE Transactions on Neural Networks. — Foundational vector-symbolic-architecture work; same family of operations.

Transformer attention — the architectural counterpoint

- Vaswani, A., et al. (2017). *Attention Is All You Need.* NeurIPS 2017. — The paper this paper's working tagline ("1-Bit Is Enough") punningly references. The standing argument that attention plus scale is the path forward; we offer the architectural counter-argument that the right discrete primitive matters more than the scale.

Neuromorphic and 1-bit hardware paths the substrate maps onto

- Davies, M., et al. (2018). *Loihi: A Neuromorphic Manycore Processor.* IEEE Micro. — Representative of the 1-bit cognitive-substrate hardware paths the substrate's 96-bytes-per-word XOR + popcount operations target.
- Lattice Semiconductor (2022). *Introducing the Lattice Avant Platform: Designed for the Mid-Range.* LatticeSemi.com — FPGA we targeted to run the 1-bit substrate.

Appendix A: The Seven Pre-Registered Hypotheses, Plain-English

The Abstract refers to seven hypotheses pre-registered before the experiments that tested them — three covering the substrate's analogy and integration behavior (H1/H2/H3) and four covering the position-aware composition experiment (H-pos-1 through H-pos-4). Pre-registration means each hypothesis was filed in dated documents with falsification protocols *before* the experimental work that would resolve it. The full per-hypothesis status appears below in the same plain-English voice as the rest of the litepaper.

H1: The substrate's energy trace should reveal "surprise" directly

Status: *Not yet tested* because it requires energy-trajectory instrumentation we haven't built yet.

What we predicted: When the substrate processes a pattern it has seen before, its internal energy should settle quickly and predictably. When it processes something genuinely new, the energy should oscillate longer because the system is doing more work to integrate the unfamiliar pattern. By analogy: a brain encountering a familiar face vs. an unfamiliar one

has different settling dynamics. If true, "this is surprising" becomes readable directly off the substrate's energy trace, without needing a separate surprise-detection model.

What happened: Not yet tested. Documented as a load-bearing prediction for the cognition-engine architecture and queued for follow-up work that requires energy-trajectory instrumentation in the broader substrate.

Why it matters: If H1 holds, "surprise" becomes a measurable, first-class output of the substrate rather than something we infer from external behavior. That's the difference between a system that *has* unexpected experiences and a system that has to *guess* whether something was unexpected.

H2: The substrate handles similar things more easily than dissimilar things

Status: Confirmed at all three vocabulary scales tested (472, 10,031, 370,106 words).

What we predicted: New information that's structurally similar to what's already in the substrate should be easier to integrate than new information that's totally distant. We predicted this would show up as a clean per-category gradient: morphological categories (plurals, past tense, comparatives, all of which share rule-structure with the rest of the vocabulary) would score near-perfectly, while referential categories (capital-of, royalty/gender, etc. which are more idiosyncratic) would lag. Filed before any analogy benchmark ran.

What happened: Confirmed at all three scales. Morphological categories hit ~100% top-10 accuracy at every scale. Referential categories did indeed lag at single-pair retrieval, exactly as predicted.

Why it matters: This is empirical evidence that the substrate is *organizing by shared structure* rather than just memorizing pattern locations. The architectural property is real, scalable, and predicted-not-discovered.

H3: Composite-fact analogies need population-level direction extraction

Status: Validated at 10K-vocabulary; strengthened at 370K; refined in scope.

What we predicted: Single-pair analogy retrieval (one example pair → predict the answer) would underperform on referential categories like capital-country. The relationship is carried at the *population* level (across many examples of the same pattern) not at the pair level. The fix: extract the majority-direction across many pair examples and apply *that* vector. Filed before any majority-direction code was written.

What happened: At 10K-vocabulary, majority-direction lifted top-10 by +41.8 percentage points on capital-country. At 370K-vocabulary it lifted by +50.0 percentage points, so it strengthened with scale, which is the opposite of typical "small-set effects that wash out at scale." Refined: the lift applies to single-shared-axis relationships (capital-of, plural-of, past-tense-of); per-pair-distinct-axis relationships (antonyms — this will be detailed in §4 of the upcoming formal paper) do *not* lift this way and need a different operation entirely.

Why it matters: This is the architectural primitive that closes 50 of the 56 percentage-point gap with continuous-vector retrieval on the hardest analogy class. The substrate doesn't have to be "as good as" continuous; it has to be smart enough to dispatch the right operation per relationship type. H3's confirmation is what makes that dispatch tractable.

H-pos-1: At N=3 (where commutative XOR works), positional composition should be roughly neutral

Status: Falsified in the *opposite* direction from prediction.

What we predicted: Adding per-position rotation to the XOR-of-context construction shouldn't help much at N=3, because commutative composition was already producing strong signal there (832× over random for Russell, 574× for Frankenstein). The positional variant was expected to be roughly equivalent.

What happened: Wrong, and in the unhelpful direction where positional composition didn't just fail to help; it *collapsed* Strategy A to 0× signal at N=3. The same primitive that left Strategy A intact (commutative) destroyed it (positional).

Why it matters: The failure was bigger than predicted, and it pointed at the deeper architectural reason: cyclic rotation produces sign-bit patterns that no longer correspond to any actual word in the substrate's vocabulary. The query lives "off the manifold" and Strategy A's nearest-Hamming retrieval can't find a real word near a non-real query.

H-pos-2: Positional composition should rescue Strategy A at higher N (foundational)

Status: Falsified at 20,000 samples per cell.

What we predicted: The "cliff" at N=4 in commutative composition (where Strategy A drops to 0–2× signal) is caused by XOR's commutativity destroying word-order information. Position-aware composition should fix this such that Strategy A at N=4 should lift from 0–2× back up to ≥100× signal-to-baseline. Filed as the load-bearing prediction for the architectural fix.

What happened: Falsified. At 20,000 samples per cell, Strategy A with positional composition stays at 0× across all N values tested. Cyclic rotation is the wrong primitive for substrate-native retrieval.

Why it matters: This was the predicted clean fix; it's wrong. The architectural research question got *sharper* as a result: the new question isn't "how do we break commutativity?" but "find a non-commutative composition primitive that respects the substrate's vocabulary manifold." Three candidate primitive families surfaced as a result (per-position pseudorandom XOR mask, parity-encoding within fixed bit-windows, learned positional embeddings under XOR). Substrate-native sequence prediction at N≥4 is *open*, not solved.

H-pos-3: Positional composition gives signal at N=5 and N=6 too

Status: Falsified for Strategy A. Confirmed in modified form for Strategy B.

What we predicted: If H-pos-2 holds, the signal should continue at higher context sizes (N=5, N=6) too, with Strategy A remaining productive.

What happened: For Strategy A: zero across all N (consequence of the H-pos-2 falsification). For Strategy B (the lookup-table retrieval approach): genuinely productive giving 105× signal-to-baseline at N=5, 87× at N=6 with positional composition.

Why it matters: A useful split. Strategy B works because K-NN retrieval over training contexts doesn't require the query to be on the substrate's vocabulary manifold, it just requires that contexts which share predictive structure end up Hamming-close to each other. The same primitive that breaks Strategy A actively *helps* Strategy B.

H-pos-4: Strategy B with positional composition is roughly neutral

Status: Upgraded to substantively improved.

What we predicted: Position-aware composition shouldn't help Strategy B much because it already does context-similarity (rather than direct retrieval), so commutativity matters less for it. Filed as the "control" hypothesis.

What happened: Substantively wrong in the favorable direction. Strategy B with positional composition lifts top-10 ratios by 35–71% at N≥4. At N=6 with positional composition, Strategy B reaches 76–87× signal-to-baseline. This is a non-trivial result for 6-gram next-word prediction with no learned weights.

Why it matters: Non-commutative composition is genuinely useful for substrate-based lookup-table sequence prediction *even though* it's the wrong primitive for substrate-native retrieval. The result demarcates the substrate's two reading modes cleanly: substrate-native retrieval (Strategy A) needs manifold-preserving primitives; lookup-table retrieval (Strategy B) does not.

Summary tally

Hypothesis	Status	One-line takeaway
H1	Not yet tested	Energy trace as direct surprise readout — queued for follow-up
H2	Confirmed	Substrate organizes by shared structure, not memorization
H3	Validated + strengthened at scale	Population-level direction extraction is the primitive that closes the gap
H-pos-1	Falsified (opposite direction)	Positional rotation doesn't help even at N=3; substrate manifold is the issue
H-pos-2	Falsified at 20K samples	Cyclic rotation is the wrong primitive; sharpened the next research question
H-pos-3	Falsified for Strategy A; confirmed for Strategy B	Manifold problem is local to substrate-native retrieval

Hypothesis	Status	One-line takeaway
H-pos-4	Upgraded (substantively improved)	Non-commutative composition helps lookup-table sequence prediction

Three confirmations, three falsifications (one in the opposite direction, one in the load-bearing slot, one in the control slot), one upgrade, and one queued. The falsifications are themselves part of the methodological discipline: an apparent simple fix was filed as a foundational prediction, tested, and wasn't proven, but that failure sharpened the architectural research question rather than burying it.

Document Metadata

- **Title:** *1-Bit Is All You Need: Quantization Doesn't Destroy Meaning — It Reveals Its Architecture*
- **Companion to:** *Type-Conditioned Action on Discrete Positions: Sign-Bit Semantic Arithmetic Reveals Substrate-Native Relationship Taxonomy* (upcoming)
- **Status:** v1.1 2026-05-15. Tracks academic paper at the same path. Update both together.
- **Audience:** funders, partners, reviewers, generally-interested readers; no specialist NLP background assumed.
- **Length target:** ~2,500-4,000 words (currently within range).