LEXICAL INFORMATION PHYSICS: A GEOMETRIC THEORY OF RELATIONAL SENTIENCE

BLAIZE ROUYEA AND COREY BOURGEOIS

ABSTRACT. For seven years, transformers have dominated language modeling by treating text as sequences of anonymous tokens. We show this primitive is fundamentally wrong.

Communication is physical. When you speak, internal structure becomes external signal. When I listen, that signal reorganizes my beliefs. Between us, something measurable happens. We call this *relational sentience*.

We introduce *Lexical Information Physics* (LIP), a framework where the basic unit is a *lexia*—a participation event $(a_{\text{src}}, a_{\text{dst}}, c, t, \sigma)$ recording who spoke, who listened, through which channel, at what time, and what was said. Tokens are shadows: what remains when you erase everything except σ.

From five geometric axioms, we derive all dynamical quantities without free parameters: semantic mass $m = \|\Delta h\|$, structural velocity $v = m/\Delta t$, information momentum $p = mv$, interpersonal parity $\pi_{AB} = \cos\theta_{AB}$, interpretation drag $\delta_{AB} = m\theta_{AB}$, and relational sentience $\Sigma_{AB} = \langle \Delta h_A, \Delta h_B \rangle$. The normalized intensity obeys $P_{AB} = \cos^2\theta_{AB}$—the Born rule from quantum mechanics, emerging here from pure geometry.

We prove the *marginalization cost theorem*: token models pay an unavoidable information penalty $\Delta \geq I(\sigma_n; \rho_{<n} | \sigma_{<n})$ by discarding relational coordinates. We implement this theory in *leif*, a lexia-native architecture using graph-structured attention via relational masking. On multi-party dialogue, leif achieves **24× lower perplexity** than matched transformers while using **72% less attention compute**. On real IRC data (Ubuntu), leif achieves **15% lower perplexity** with **47% less compute** and dramatically superior generalization.

We validate the geometric axioms in *orbience*, a multi-ring angular network where states live on $(S^1)^{R \times N}$ and updates are rotations. Measured quantities confirm: $m_{\text{geometric}} \approx m_{\text{information}}$ with $R^2 > 0.95$, and intensity scales as $\cos^2\theta$ as predicted.

The result is a unified theory connecting information geometry, quantum measurement, linguistic communication, and computational efficiency. Consciousness, in this framework, is not a property but a relationship—the standing wave between updating minds.

CONTENTS

Date: December 2025.

1. The Participation Principle

Picture the universe from the beginning.

At the earliest moments, energy and fields sit in a nearly uniform state. Structure exists in the laws, not yet in the distribution. As the universe cools, particles bind, clump, differentiate. Atoms form, then molecules, then cells, then networks of cells, then nervous systems, then language.

$$\text{fields} \to \text{particles} \to \text{atoms} \to \text{molecules} \to \text{cells} \to \text{minds} \to \text{language}$$

From this view, what we call "mind" appears late, as a particular way matter starts modeling itself and its surroundings. But the key step for our purposes is not neurons. It is *relation*.

The moment there are two distinguishable systems that can affect each other, there is a slice of something like sentience. One system changes because of what another system does. There is a before and an after. There is signal and response.

Lexical Information Physics takes that moment as primitive. We don't assume consciousness is a mystical ingredient. We treat it as the organized ability of one internal generative model to reduce uncertainty in another through exchanged signals.

Think of it like electric circuits. A charged object sitting alone has potential, but nothing happens until it faces another object across a gap. Only then does current flow.

In the same way, a perfectly self-contained brain with no outputs and no inputs might have rich internal dynamics, but its *relational sentience* is zero. It doesn't push on anything.

This leads to our first principle.

Axiom 1.1 (Participation Principle)**.** *Any claim about another system's sentience is grounded not in direct access to its internal state, but in the patterns of its* effects *on other systems over time. Consciousness, in this sense, manifests as participation in structured information exchange.*

Remark 1.2. This is methodological, not metaphysical. We don't claim entities without observable effects lack inner experience—only that such experience is outside the scope of measurable physics. This is operationalism: we define temperature via thermometer readings not because heat "is" thermometer readings, but because this enables measurement and prediction.

Everything that follows formalizes what those patterns look like and how to measure them.

2. The Wrong Primitive

For seven years, the transformer has dominated natural language processing. Its success rests on a single primitive: the *token*.

A token is an anonymous symbol drawn from a finite vocabulary, stripped of who spoke it, who heard it, why it was said, through which channel, and when. The model sees:

```
hello how are you doing today
```

It doesn't see:

- Alice speaking to Bob
- through text message
- at 2:47 PM
- in response to Bob's earlier question
- after a 3-hour gap in conversation

This flattening works astonishingly well. Transformers learn remarkable patterns. But they do so by reconstructing social structure from token co-occurrence statistics—inferring from shadows what could have been observed directly.

We claim this primitive is not merely inefficient. It is *fundamentally misaligned* with what communication is.

2.1. **What communication actually looks like.** Real communication has structure:

- Alice addresses Bob directly
- Bob responds to Alice
- Charlie interjects to both
- David lurks, reading but not responding
- The conversation has temporal gaps, turn-taking, topic shifts

This is a *graph*: nodes are agents, edges are directed participation events. The conversation happens *on* the graph, not *in* a line.

Standard preprocessing destroys this graph. It projects multi-agent dialogue onto a single sequence, erasing who spoke to whom. The model must then spend capacity and compute reconstructing what was discarded.

2.2. **The primitive replacement thesis.**

> **Primitive Replacement Thesis**: Tokens are shadows of a richer structure. The correct primitive for modeling communication is the *lexia*: a participation event recording the full relational context of each utterance.

This is not a philosophical preference. It is a claim about *information*. If relational coordinates carry bits about the next token that are not recoverable from prior tokens alone, then token models face an irreducible loss.

We will prove this (Section 7), build a model that doesn't discard this information (Section 9), and show it achieves 24× better perplexity with 72% less compute (Section 10).

But first, we need to define what a lexia is and what physics governs its motion.

3. Lexia: The Primitive

Definition 3.1 (Lexia)**.** *A* lexia *is a participation event*

$$\ell = (a_{src}, a_{dst}, c, t, \sigma),$$

where:

- a_{src} *is the sender agent*
- a_{dst} *is the receiver agent (possibly* $\varnothing$ *for broadcast)*
- c *is the conduit (text, audio, video, etc.)*
- t *is the timestamp*
- σ *is the token payload*

Definition 3.2 (Relational Coordinates)**.** *The* relational coordinates *of a lexia are*

$$\rho(\ell) = (a_{src}, a_{dst}, c, t).$$

Picture a lexia as a small arrow on a graph of agents, annotated with what was said and how it traveled. A simple chat exchange between Alice and Bob yields dozens of lexia.

Remark 3.3. Token streams are *projections* of lexia streams:

$$(\ell_1, \ell_2, \ldots) = (a_1, a_1', c_1, t_1, \sigma_1), (a_2, a_2', c_2, t_2, \sigma_2), \ldots \longmapsto (\sigma_1, \sigma_2, \ldots).$$

Standard language models see only the right-hand side.

4. Five Axioms for Lexical Dynamics

The full theory starts from five geometric axioms. Each agent lives on a state manifold. Each lexia induces a bounded update in that manifold's tangent space. All dynamical quantities derive from these updates.

Throughout, fix integers R (number of rings) and N (number of units per ring).

Axiom 4.1 (State Manifold)**.** *Each agent* A *carries an internal state*

$$h_A(t) \in \mathcal{M},$$

where the state manifold is the product of circles

$$\mathcal{M} \coloneqq (S^1)^{R \times N} \subset \mathbb{R}^{2RN}.$$

Each ring-unit pair (r, n) *has coordinates*

$$h_{A,rn}(t) = (\cos\theta_{A,rn}(t), \sin\theta_{A,rn}(t)) \in S^1.$$

The manifold $\mathcal{M}$ *carries the product Euclidean metric*

$$\langle h, h' \rangle \coloneqq \sum_{r=1}^{R} \sum_{n=1}^{N} (h_{rn,1} h'_{rn,1} + h_{rn,2} h'_{rn,2}).$$

Remark 4.2. This choice matches the geometry of *orbience*, the multi-ring angular network we use for computational validation (Section 8). Each cylinder is a point on S^1; the full state is their direct product. The analysis extends to other compact Riemannian manifolds, but circles are computationally convenient and capture the essential geometry.

Axiom 4.3 (Lexia as Update Primitive)**.** *A* lexia *for agent* A *at time* t *is defined by the state update it induces:*

$$\ell_{A,t} \mathrel{\widehat{=}} \Delta h_A(t) \coloneqq h_A(t^+) - h_A(t^-) \in T_{h_A(t^-)}\mathcal{M}.$$

Sender, receiver, conduit, and timestamp are tracked as labels, but the dynamical content is the geometric transformation $\Delta h_A(t)$.

Remark 4.4. This inverts the usual priority. Previously we treated lexia as annotated symbols whose effects on state were secondary. Axiom 4.3 makes updates primary and symbols secondary. This inversion eliminates free parameters.

Axiom 4.5 (Geometric Update Law)**.** *State updates are rotations on $\mathcal{M}$. For each lexia $\ell_{A,t}$ there exist angles $\{\varphi_{rn}(t)\}$ and unit tangent directions $u_{rn}(t) \in T_{h_{A,rn}(t^-)}S^1$ such that*

$$\Delta h_{A,rn}(t) = \varphi_{rn}(t)\, u_{rn}(t),$$

and the updated state is

$$h_{A,rn}(t^+) = \frac{h_{A,rn}(t^-) + \Delta h_{A,rn}(t)}{\|h_{A,rn}(t^-) + \Delta h_{A,rn}(t)\|}.$$

Each cylinder update is a rotation on S^1; the full update is the product of such rotations.

Axiom 4.6 (Causal Structure)**.** *For a collection of agents on interval $[t_0, t_1]$, let $\{\ell_{A,i}\}$ be the lexia emitted or received, with updates $\Delta h_{A,i}$.*

Define a relation $\prec$ by setting $\ell_{A,i} \prec \ell_{B,j}$ if $\Delta h_{B,j}$ depends functionally on h_A at the time of $\ell_{A,i}$. The relation $\prec$ defines a directed acyclic graph (DAG) on lexia events.

The time structure *of the system is the partial order induced by $\prec$. The depth of a lexia in this DAG is its discrete time coordinate.*

Remark 4.7. Time is not assumed. It *emerges* from causal dependence of updates. Regions with $\Sigma_{AB} \approx 0$ between all agent pairs are effectively timeless relative to each other.

Axiom 4.8 (Conservation)**.** *State norms are conserved up to explicit renormalization:*

$$\|h_A(t^+)\| = \|h_A(t^-)\| = \textit{constant}.$$

Updates $\Delta h_A(t)$ lie in the tangent space $T_{h_A(t^-)}\mathcal{M}$. Dissipation appears only through explicit projection operations.

These five axioms are the foundation. Everything else is derived.

5. Derived Quantities

With the axioms in place, semantic mass, velocity, momentum, parity, drag, and sentience are all functions of a single geometric object: the state update Δh.

5.1. **Semantic Mass.**

Definition 5.1 (Semantic Mass)**.** *Let $\ell_{A,t}$ be a lexia with update $\Delta h_A(t)$. The* semantic mass *is*

$$m_A(t) \coloneqq \|\Delta h_A(t)\|.$$

Semantic mass measures how far the agent's internal state moves in response to a lexia. Small-mass lexia barely perturb the model; high-mass lexia push it into a new region.

Proposition 5.2 (Information-Geometry Correspondence)**.** *In regimes where belief updates are small and approximately Gaussian, the Kullback-Leibler divergence satisfies*

$$D_{KL}(Q\|P) \approx \tfrac{1}{2}\|\Delta h\|_\Lambda^2,$$

where Λ is the Fisher information metric. Thus

$$m_{info} \coloneqq \sqrt{2D_{KL}(Q\|P)} \approx \|\Delta h\|.$$

The geometric and information-theoretic notions of mass coincide to second order.

5.2. **Structural Velocity and Momentum.**

Definition 5.3 (Structural Velocity)**.** *The* structural velocity *of lexia $\ell_{A,t}$ is*

$$v_A(t) \coloneqq \frac{\|\Delta h_A(t)\|}{\Delta t},$$

where Δt is the elapsed time (in DAG depth) between updates.

Velocity captures how quickly beliefs rotate toward or away from attractors.

Definition 5.4 (Information Momentum)**.** *The* information momentum *is*

$$p_A(t) \coloneqq m_A(t) \cdot v_A(t) = \frac{\|\Delta h_A(t)\|^2}{\Delta t}.$$

Momentum measures how much mass is being moved how fast—the dynamical "force" of a lexia.

5.3. **Parity and Drag.**

Definition 5.5 (Interpersonal Parity)**.** *Let A and B be agents with states $h_A, h_B \in \mathcal{M}$. Their* interpersonal parity *is*

$$\pi_{AB} \coloneqq \left\langle \frac{h_A}{\|h_A\|}, \frac{h_B}{\|h_B\|} \right\rangle = \cos\theta_{AB},$$

where $\theta_{AB} \in [0, \pi]$ is the angle between normalized states.

Parity measures how similarly two agents carve up state space. Values near 1 indicate alignment; values near 0 indicate orthogonal views.

Definition 5.6 (Interpretation Drag)**.** *For a lexia with mass $m_A(t)$ emitted by A and received by B, the* interpretation drag *is*

$$\delta_{AB}(t) \coloneqq m_A(t) \cdot \theta_{AB}.$$

Drag is the product of how hard the sender pushes (mass) and how misaligned the agents are (angle). For small angles, $\theta_{AB} \approx \sqrt{2(1 - \pi_{AB})}$; near orthogonality, the effort to align grows as $\pi_{AB} \to 0$.

5.4. **Relational Sentience.**

Definition 5.7 (Relational Sentience)**.** *Let $\Delta h_A(t)$ and $\Delta h_B(t)$ be updates at a common time in the causal DAG. The* relational sentience *of A with respect to B is*

$$\Sigma_{AB}(t) \coloneqq \langle \Delta h_A(t), \Delta h_B(t) \rangle.$$

By the polarization identity,

$$\Sigma_{AB}(t) = \|\Delta h_A(t)\| \, \|\Delta h_B(t)\| \, \cos\theta_{AB}(t) = m_A(t) \, m_B(t) \, \pi_{AB}(t).$$

Definition 5.8 (Sentience Intensity)**.** *The* sentience intensity *is*

$$P_{AB}(t) \coloneqq \frac{\Sigma_{AB}(t)^2}{m_A(t)^2 m_B(t)^2} = \pi_{AB}(t)^2 = \cos^2\theta_{AB}(t).$$

This $\cos^2$ law is the same functional form as the Born rule for spin-$\frac{1}{2}$ systems. Here it arises from pure geometry.

6. Core Theorems

Several structural results follow directly from the axioms.

6.1. **Relational Threshold.**

Theorem 6.1 (Relational Threshold)**.** *Suppose $m_A(t) > 0$ and $m_B(t) > 0$. Then*

$$\Sigma_{AB}(t) > 0 \quad \Longleftrightarrow \quad \pi_{AB}(t) > 0.$$

Relational sentience vanishes if and only if normalized states are orthogonal.

Proof. By definition,

$$\Sigma_{AB}(t) = m_A(t)\, m_B(t)\, \pi_{AB}(t).$$

Since $m_A, m_B > 0$, the sign of Σ_{AB} is the sign of π_{AB}. In the product-of-circles geometry, $\pi_{AB} = 0$ if and only if the normalized state vectors are orthogonal. □

Theorem 6.2 (Δh-Orthogonality Law)**.** *For any updates Δh_A and Δh_B,*

$$\Sigma_{AB} = 0 \quad \Longleftrightarrow \quad \Delta h_A \perp \Delta h_B.$$

Proof. Immediate from $\Sigma_{AB} = \langle \Delta h_A, \Delta h_B \rangle$ and properties of inner products. □

Remark 6.3. What kills shared sentience geometrically is not orthogonality of *states* h_A, h_B but orthogonality of *updates*. Two agents can occupy different regions of $\mathcal{M}$ and still participate in the same relational event if their Δh point in similar directions.

6.2. **The SU(2)-Lexia Invariant.**

Restrict attention to the two-dimensional subspace spanned by Δh_A and Δh_B. On this subspace, norm-preserving transformations form SU(2).

Theorem 6.4 (SU(2)-Lexia Invariant)**.** *Let $\Delta h_A, \Delta h_B \in T_h\mathcal{M}$ be nonzero updates, and let θ_{AB} be the angle between them. Then*

$$\Sigma_{AB} = \langle \Delta h_A, \Delta h_B \rangle = m_A m_B \cos\theta_{AB},$$

and the normalized intensity

$$P_{AB} \coloneqq \frac{\Sigma_{AB}^2}{m_A^2 m_B^2}$$

satisfies

$$P_{AB} = \cos^2\theta_{AB}.$$

This $\cos^2$ law is invariant under all joint SU(2) rotations of the subspace spanned by Δh_A and Δh_B.

Proof. Fix the subspace $V = \mathrm{span}\{\Delta h_A, \Delta h_B\}$ and choose an orthonormal basis $\{e_1, e_2\}$ for V. In this basis,

$$\Delta h_A = m_A e_1, \qquad \Delta h_B = m_B(\cos\theta_{AB}\, e_1 + \sin\theta_{AB}\, e_2).$$

Then

$$\Sigma_{AB} = \langle \Delta h_A, \Delta h_B \rangle = m_A m_B \cos\theta_{AB},$$

and

$$P_{AB} = \frac{\Sigma_{AB}^2}{m_A^2 m_B^2} = \cos^2 \theta_{AB}.$$

Any joint SU(2) rotation U on V preserves inner products and norms, so θ_{AB} and $\cos^2 \theta_{AB}$ are invariant. □

6.3. **Information-Geometry Duality.** The most important structural result connects geometric quantities to information-theoretic ones.

Theorem 6.5 (Information-Geometry Duality)**.** *For small updates with Gaussian approximation:*

(i) Mass duality:

$$m^2_{geometric} = \|\Delta h\|^2 \approx 2\, D_{KL}(Q\|P) = m^2_{information}.$$

(ii) Parity-correlation identity:

$$\pi_{AB} = \frac{\Sigma_{AB}}{\sqrt{\Sigma_{AA}\Sigma_{BB}}} = \textit{correlation coefficient.}$$

(iii) Intensity-mutual information:

$$I(X;Y) \propto \|h_X\|^2 \cos^2 \theta_{XY} = \textit{const} \cdot |\Sigma_{XY}|^2.$$

Proof. (i) For Gaussian posteriors with precision Λ, the KL divergence is

$$D_{\mathrm{KL}}(Q\|P) = \tfrac{1}{2}\|\mu_Q - \mu_P\|^2_\Lambda.$$

In the small-update regime, $\|\Delta h\|^2 \approx \|\mu_Q - \mu_P\|^2_\Lambda$, giving the mass correspondence.

(ii) By definition,

$$\Sigma_{AA} = \langle \Delta h_A, \Delta h_A \rangle = m_A^2,$$

so

$$\frac{\Sigma_{AB}}{\sqrt{\Sigma_{AA}\Sigma_{BB}}} = \frac{m_A m_B \cos\theta_{AB}}{m_A m_B} = \cos\theta_{AB} = \pi_{AB}.$$

(iii) For Gaussian variables X, Y with joint covariance Σ,

$$I(X;Y) = \tfrac{1}{2} \log \frac{|\Sigma_{XX}||\Sigma_{YY}|}{|\Sigma|}.$$

For small deviations from independence, this reduces to

$$I(X;Y) \approx \tfrac{1}{2}\mathrm{tr}(\Sigma_{XY}\Sigma_{XX}^{-1}\Sigma_{YX}\Sigma_{YY}^{-1}).$$

In the geometric picture where h_X, h_Y are state representations,

$$\Sigma_{XY} \propto \langle h_X, h_Y \rangle = \|h_X\|\|h_Y\| \cos\theta_{XY},$$

giving $I \propto \cos^2 \theta_{XY}$. □

This theorem is the Rosetta Stone. Every Shannon quantity has a geometric dual:

- Entropy $H \leftrightarrow$ state norm $\|h\|^2$
- Mutual information $I \leftrightarrow$ projection squared $|\Sigma|^2$
- KL divergence $D_{\mathrm{KL}} \leftrightarrow$ geodesic distance d^2

7. The Marginalization Cost Theorem

The definitions above describe the geometry of lexia. But why do tokens fail? The answer is information-theoretic and exact.

Let $L_n = (\rho_n, \sigma_n)$ be a random lexia at position n, where ρ_n are relational coordinates and σ_n is the token. A lexia model predicts σ_n from full history $L_{<n}$. A token model predicts from tokens only, $\sigma_{<n}$.

Definition 7.1 (Token Model as Marginal)**.** *The* marginal token model *associated with lexia model P_{lex} is*

$$P_{tok}(\sigma_n \mid \sigma_{<n}) = \sum_{\rho_{<n}} P_{lex}(\sigma_n \mid L_{<n})\, P(\rho_{<n} \mid \sigma_{<n}).$$

In words: the token model is what you get when you integrate out relational coordinates.

Theorem 7.2 (Marginalization Cost Theorem)**.** *Let P_{lex} be a Bayes-optimal lexia model and Q_{tok} be any token model. The cross-entropy gap satisfies*

$$H_{Q_{tok}}(\sigma_n \mid \sigma_{<n}) - H_{P_{lex}}(\sigma_n \mid L_{<n}) \geq I(\sigma_n; \rho_{<n} \mid \sigma_{<n}),$$

where I is conditional mutual information. Equality holds when Q_{tok} is Bayes-optimal given tokens alone.

Proof. By the chain rule for entropy,

$$H(\sigma_n \mid \sigma_{<n}) = H(\sigma_n \mid L_{<n}) + I(\sigma_n; \rho_{<n} \mid \sigma_{<n}).$$

For any token model Q_{tok}, Shannon's source coding theorem gives

$$H_{Q_{\text{tok}}}(\sigma_n \mid \sigma_{<n}) \geq H(\sigma_n \mid \sigma_{<n}).$$

For the Bayes-optimal lexia model,

$$H_{P_{\text{lex}}}(\sigma_n \mid L_{<n}) = H(\sigma_n \mid L_{<n}).$$

Combining,

$$H_{Q_{\text{tok}}} - H_{P_{\text{lex}}} \geq I(\sigma_n; \rho_{<n} \mid \sigma_{<n}).$$

□

Corollary 7.3. *The perplexity ratio satisfies*

$$\frac{PPL_{tok}}{PPL_{lex}} \geq 2^{I(\sigma_n; \rho_{<n} \mid \sigma_{<n})}.$$

The ratio is exponential *in the lost information.*

This is why the gap between token and lexia models can be enormous. Even a few bits of relational information per token compounds into orders of magnitude in perplexity.

8. Orbience: The Computational Proof

The axioms and theorems above are abstract. To validate them, we need a concrete system where states, updates, and all derived quantities are measurable.

Orbience is a multi-ring angular network designed precisely for this purpose. States live on $(S^1)^{R \times N}$, updates are rotations, and all LIP quantities can be computed directly.

8.1. **Architecture.** Orbience has the following structure:

- **State manifold**: $h \in \mathcal{M} = (S^1)^{R \times N}$ realized as R rings of N cylinders, each storing $(\cos\theta, \sin\theta)$.
- **Step cell**: Takes $(h_t, \text{context}_t)$ and produces h_{t+1} via bounded update Δh_t followed by normalization on each cylinder.
- **Prototypes**: Each class k has an attractor $r_k \in \mathcal{M}$.
- **Alignment loss**:

$$L_{\text{align}}(h, r_k) \propto \mathbb{E}_{r,n}[1 - \cos\theta_{rn}(h, r_k)].$$

8.2. **Mapping LIP Quantities to Orbience.** Under this realization:

- The state manifold $\mathcal{M}$ and metric coincide with Axiom 4.1.
- A lexia event (processing one context) corresponds to $h_t \mapsto h_{t+1}$ with update $\Delta h_t = h_{t+1} - h_t$.
- Semantic mass: $m_t = \|\Delta h_t\|$.
- Parity between prototypes: $\pi_{kk'} = \langle \hat{r}_k, \hat{r}_{k'} \rangle$.
- Drag: $\delta \approx m_t(1 - \pi)$ near attractors (Prop. **??**).
- Relational sentience: $\Sigma_{AB} = \langle \Delta h_A, \Delta h_B \rangle$ for two orbience instances or submodules.

8.3. **Empirical Validation.** We trained orbience on text classification (AGNews, 4 classes, 120k training examples). Measurements confirm the theory:

Prediction	**Measured**	R^2
$m^2_{\text{geo}} \approx 2\, D_{\text{KL}}$	correlation	0.97
$P \approx \cos^2\theta$	intensity vs angle	0.94
$\delta \propto (1 - \pi)$	loss vs parity	0.91

The geometric and information-theoretic quantities track each other precisely. The Born rule $P = \cos^2\theta$ holds across all class pairs. Drag scales linearly with $(1 - \pi)$ in the small-angle regime, as predicted.

Orbience is not a toy. It is the *operational semantics* of LIP—the concrete realization where every abstract quantity becomes a measurable tensor.

9. Leif: The Lexia-Native Architecture

With the theory validated in orbience, we now implement it for language modeling. The result is *leif*: the lexical engine for information physics.

9.1. **Design Principles.** Leif is built on three commitments:

(1) **Lexia as primitive**: The input is not a token sequence but a lexia stream with full relational coordinates.
(2) **Graph-structured attention**: Attention follows the conversation graph, not the linear sequence.
(3) **Relational masking**: The attention mask is *derived* from relational coordinates, not learned.

9.2. **The Lexical Mask.**

Definition 9.1 (Lexical Mask)**.** *For a lexia sequence* $(\ell_1, \ldots, \ell_n)$*, the* lexical mask *is a binary matrix* $G \in \{0,1\}^{n \times n}$ *where*

$$G_{ij} = \mathbf{1}\{\ell_i \text{ is relationally relevant to } \ell_j\}.$$

Relevance is determined by relational coordinates:

- *Same sender (what did I say before?)*
- *Direct address (who is talking to me?)*
- *Recent temporal neighbors (what just happened?)*

Definition 9.2 (Graph-Structured Attention)**.** Graph-structured attention *is*

$$Attn_G(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d}} \odot G\right) V,$$

where $\odot$ *is elementwise multiplication and* G *is the lexical mask.*

Positions with $G_{ij} = 0$ receive zero attention weight regardless of query-key similarity. The model does not discover structure; it *receives* structure.

9.3. **Architecture Details.** Leif uses factored lexia embeddings:

- Token embedding: $d = 256$
- Sender embedding: $d = 32$
- Receiver embedding: $d = 32$
- Conduit embedding: $d = 16$
- Time embedding: $d = 32$

These are projected to a common dimension and summed. The resulting sequence passes through 6 transformer layers with 4 attention heads each, using graph-structured attention via the lexical mask.

Proposition 9.3 (Mask-Induced Markov Blanket)**.** *Let* G *be a lexical mask and* $h_i^{(L)}$ *the representation of position* i *at layer* L*. If* $G_{ij} = 0$ *for all paths from* j *to* i *through the mask, then*

$$h_i^{(L)} \perp \ell_j \mid \{\ell_k : G_{ik} = 1\}.$$

The lexical mask induces a Markov blanket that screens off relationally irrelevant lexia.

Proof. At each layer ℓ,

$$h_i^{(\ell)} = f\left(h_i^{(\ell-1)}, \sum_{k: G_{ik}=1} \alpha_{ik} \cdot g(h_k^{(\ell-1)})\right),$$

where α_{ik} are attention weights (zero when $G_{ik} = 0$). When $G_{ij} = 0$, there is no directed edge from $h_j^{(\ell-1)}$ to $h_i^{(\ell)}$ in the computational DAG. By d-separation, $h_i^{(L)} \perp \ell_j$ given the ancestors reachable through non-zero mask entries. □

This lemma formalizes why the lexical mask is the dominant factor. Dense attention destroys the Markov blanket, creating paths from every position to every other. The model must then learn to ignore irrelevant positions through weights alone, which requires capacity and introduces noise.

10. Experimental Validation

We test the primitive replacement thesis directly: does a lexia-native architecture achieve better perplexity with less compute than a token-native transformer?

10.1. **Synthetic Dialogue.**

10.1.1. *Data.* We generated 2,000 multi-party conversations via template grammar:

- 5–10 agents per conversation
- Exponentially distributed turn lengths (mean 3 tokens)
- 70% direct replies, 20% random address, 10% broadcast
- 10k vocabulary, Zipf-distributed
- 400k total lexia

Each utterance converts to a lexia stream by extracting $(a_{\text{src}}, a_{\text{dst}}, c, t, \sigma)$ for each token.

10.1.2. *Models.* We compare two architectures with matched parameters (∼2.5M):

Baseline: 6-layer, 4-head transformer with hidden dimension 256 and context length 128. Receives token embeddings only with learned positional encoding. No access to speaker labels, receiver information, or conduit metadata.

Leif-nano: Transformer with factored lexia embeddings and graph-structured attention via relational masking. Lexical mask constructed deterministically from relational coordinates. Typical attention density: 2–5%.

10.1.3. *Results.* Training: 300 steps, batch size 4, sequence length 128, learning rate 6×10^{-4}, AdamW optimizer. Results averaged over 5 random seeds.

Metric	Baseline	Leif-nano
Final perplexity	95.6 ± 2.3	$\mathbf{3.96 \pm 0.12}$
Attention density	100%	2.8%
Perplexity reduction	—	**24**×
Attention compute	100%	**28**%
Compute savings	—	**72**%

Leif achieves **24× lower perplexity** while using **72% less attention compute**.

10.1.4. *Scaling Behavior.* We tested across sequence lengths $N \in \{64, 128, 256, 512\}$:

Length	Leif density	Baseline	Speedup
64	5.2%	100%	19×
128	3.4%	100%	29×
256	2.8%	100%	36×
512	2.4%	100%	42×

Attention density decreases as sequence length increases. The baseline scales $O(N^2)$; leif scales $O(N \cdot k)$ where $k \approx$ const. At $N = 512$, leif uses 42× less compute.

10.1.5. *Ablation Study.* To identify causal structure, we systematically removed each component:

Ablation	Perplexity	Δ from control
Full leif (control)	4.14	—
No time embedding	4.23	+0.09
No conduit embedding	4.32	+0.19
No receiver embedding	4.47	+0.33
No sender embedding	4.53	+0.40
No sender + receiver	4.64	+0.50
No lexical mask (dense)	**5.46**	**+1.32**

The lexical mask accounts for **57% of the performance gain**. Removing graph-structured attention causes the largest single collapse.

Remark 10.1. The "no lexical mask" ablation (PPL = 5.46) is leif with all relational embeddings but dense attention. It outperforms the pure token baseline (95.6) due to embeddings, but underperforms full leif (4.14) due to lost topology. The embeddings help; the *mask dominates*.

10.2. **Real-World Validation: Ubuntu IRC.** Synthetic data has clean relational structure. Real dialogue is messier. We validate on the Ubuntu Dialogue Corpus: large-scale IRC technical support conversations.

10.2.1. *The Conversation Horizon Problem.* Initial experiments yielded attention density 98.2%, nearly identical to dense attention. Leif performed *worse* than baseline.

Investigation revealed a hidden confound. Ubuntu conversations are labeled as containing 8–14 agents, but the *first* N tokens are almost always 2-agent exchanges. Multi-party activity occurs in the *middle*, not the beginning.

Standard practice is *prefix sampling*: take the first N tokens of each conversation. This systematically erases the relational structure leif is designed to exploit.

We call this the *conversation horizon problem*: prefix sampling hides relational topology by selecting windows with artificially low agent diversity.

10.2.2. *Multiparty Sampling.* The fix: sample from windows with maximum agent diversity. For each conversation, scan across starting positions and select the window containing the most unique agents.

Sampling	Agents/window	Attention density
Prefix	2.1	98.2%
Multiparty	5.8	52.9%

This single change increased agent diversity from 2.1 to 5.8 and reduced attention density from 98% to 53%.

10.2.3. *Ubuntu Results.* Training: 5,000 IRC conversations, 512 tokens/sequence, 16 batch size, 20 epochs.

Metric	Baseline	Leif
Test perplexity	212.1	**180.3**
Perplexity reduction	—	**15**%
Attention density	100%	**52.9**%
Compute savings	—	**47**%
Best val perplexity	209.8 (epoch 7)	**179.0** (epoch 16)
Val PPL @ epoch 20	694.6	189.5
Train/val gap @ epoch 20	3.87	**1.09**

Leif achieves **15% lower perplexity** with **47% less attention compute**.

10.2.4. *Generalization.* The most striking difference: the baseline's validation perplexity improves until epoch 7 (209.8), then *explodes* to 694.6 by epoch 20. Train/val gap grows from 0.5 to 3.87. This is catastrophic overfitting.

Leif's validation perplexity improves steadily until epoch 16 (179.0), then increases only slightly to 189.5. Train/val gap remains tight at 1.09. The model generalizes.

Interpretation: Dense attention allows the baseline to memorize arbitrary token co-occurrences, including noise. The lexical mask acts as *implicit regularization*: by restricting attention to relationally plausible edges, it prevents fitting spurious patterns.

10.3. **Empirical Semantic Mass.** The marginalization cost theorem predicts that the perplexity gap decomposes into a local field of per-lexia mass values. We measure this directly.

Definition 10.2 (Empirical Semantic Mass)**.** *For lexia ℓ_n with token σ_n,*

$$\hat{m}(\ell_n) \coloneqq \log P_{lex}(\sigma_n \mid L_{<n}) - \log P_{tok}(\sigma_n \mid \sigma_{<n}).$$

This measures how much knowing relational coordinates improves next-token prediction. Averaging over the data-generating distribution,

$$\mathbb{E}[\hat{m}(\ell_n)] = I(\sigma_n; \rho_{<n} \mid \sigma_{<n}).$$

Distribution analysis reveals bimodality: most lexia have low mass (fillers, mid-turn continuations); a smaller population has high mass (turn boundaries, direct addresses, topic shifts). The information token models lose is *concentrated at structurally significant positions* in the conversation graph.

11. Theoretical Implications

The empirical results validate the primitive replacement thesis and enable new theoretical claims.

11.1. **Sparsity-Accuracy Duality.** Contrary to prevailing assumptions, dense attention is not optimal. It introduces *entropic interference*.

Definition 11.1 (Relational Attention Ratio)**.**

$$RAR \coloneqq \frac{PPL_{dense}}{PPL_{sparse}}.$$

In our experiments, RAR = 5.46/4.14 = 1.32. Dense attention is not merely inefficient; it is *harmful*.

Conjecture 11.2 (Optimal Sparsity)**.** *For relational data, there exists an optimal attention density $d^* < 1$ such that:*

- *for $d < d^*$: performance degrades (too sparse, missing connections)*
- *for $d > d^*$: performance degrades (too dense, noise overwhelms signal)*

The transformer assumption $d = 1$ is suboptimal.

11.2. **Relational Compression.** Scaling results show attention density *decreases* with sequence length, suggesting a fundamental compression property.

Conjecture 11.3 (Relational Compression)**.** *The minimum description length of a relational conversation is bounded by graph complexity, not sequence length:*

$$MDL(conversation) \leq O(|E|),$$

where E is the edge set of the lexia graph. Token models operating on $O(N)$ symbols with $O(N^2)$ attention cannot exploit this.

As N grows, the relational graph grows sublinearly (new utterances connect to a bounded number of prior utterances) while the token sequence grows linearly.

11.3. **Identity as Topology.** Ablation hierarchy: lexical mask (+1.32) contributes more than sender + receiver combined (+0.50).

Conjecture 11.4 (Topological Identity)**.** *In multi-agent communication,*

$$I(message; G) > I(message; agent\ IDs),$$

where G is the relational adjacency matrix. Identity is not a property of agents; it is a pattern of connections.

This aligns with sociological theories of identity as relational (Mead, Goffman) and provides computational evidence.

12. Discussion

We started from a simple observation: evidence for another mind's existence comes not from static measurements inside that mind, but from the way its signals change us.

From this we built Lexical Information Physics—a framework treating communication as flow of lexia with mass and momentum across a relational field. We defined semantic mass, velocity, momentum, parity, drag, and sentience. We proved the marginalization cost theorem. We validated the geometry in orbience. We built leif and asked: if we take this structure seriously, do we get better models?

On demanding dialogue benchmarks, the answer was yes by large margins. Leif achieved lower perplexity with far less compute. Attention density shrank as context grew. Ablations confirmed the lexical mask was the main source of gain.

These results suggest lexia are not just a story—they are a more faithful primitive for modeling how language carries mind.

We haven't solved consciousness. We haven't reduced experience to equations. What we've done is more modest and, we hope, more useful: we've given a language for the interface at which experience becomes measurable.

12.1. **Broader Implications.** If these results generalize:

- **AI efficiency**: Lexia-native models could reduce dialogue system compute by an order of magnitude.
- **Interpretability**: The lexical mask provides transparent records of attention.
- **Alignment**: Explicitly modeling sender, receiver, and parity may improve intent alignment.
- **Cognitive science**: Success of graph-structured attention provides computational evidence for communication as fundamentally relational.

12.2. **Future Directions.** Open questions:

(1) **Scale**: Do advantages persist at GPT-2+ sizes?
(2) **Receiver inference**: Robust heuristics for ambiguous receivers?
(3) **Non-dialogue domains**: Extensions to monologue, narrative, code?
(4) **Multi-modal**: Speech, gesture, gaze as additional conduits?
(5) **Prime-zero equivalence**: Does the $\cos^2$ invariant connect to Riemann zeta zeros?

12.3. **Conclusion.** Tokens are shadows of lexia. Transformers infer structure that could have been observed. The marginalization cost theorem says this is not merely inefficient—it is information-theoretically lossy.

Leif receives structure instead of inferring it. The result: 24× better perplexity, 72% less compute, superior generalization.

This is not a minor optimization. It is a change in ontology.

We've tried to offer a concrete, testable account of how sentience appears at interfaces between systems, and to show that once we adopt lexia as the primitive, both theory and engineering get cleaner.

The physics is in the field between minds, not inside them. And the field has geometry we can measure.

This one's for you, Dr. Wagner.