THE THEORY OF RELATIONAL SENTIENCE AND LEXICAL INFORMATION PHYSICS

BLAIZE ROUYEA

Abstract. What if we've been modeling language with the wrong primitive all along?

For the last seven years, the transformer has dominated natural language processing by treating text as sequences of tokens: anonymous symbols in a line, stripped of who spoke them, who heard them, and why they were said.

This paper starts from a different place. We treat communication as a physical process and language as the motion of structured events through a field of minds. The basic unit is not a token but a *lexia*: a participation event that records who is speaking, who is listening, through which conduit, at what time, and with what lexical payload.

We build *lexical information physics* (LIP), a framework where each lexia carries *semantic mass* (how much internal state it reveals) and moves with *structural velocity* (how efficiently it is organized with respect to context). Their product defines *information momentum*, which measures how strongly one internal generative model can reorganize another.

On top of this we define *relational sentience* as a functional of momentum, interpersonal parity, and influence. We introduce *interpretation drag* as the energetic cost of updating a receiver's model, and we formalize *interpersonal parity* as the alignment between two agents' models of sentience. Together, these quantities describe when communication feels effortless, when it grinds, and when it fails.

The central theoretical result is the *marginalization cost theorem*: standard token-based language models arise as shadow processes obtained by integrating out the relational coordinates of lexia, and they pay an unavoidable information-theoretic penalty for doing so. The gap is exactly the conditional mutual information between the next token and the discarded coordinates.

We then present *leif* (the lexical engine for information physics), a lexia-native architecture built with graph-structured attention through a *lexical mask*. On a 400k-lexia multi-party dialogue benchmark, leif achieves *twenty-four times lower perplexity* than a matched transformer while using roughly *seventy-two percent less attention compute*. Attention density decreases as sequence length grows, yielding scaling that is effectively linear in sequence length.

The goal of this paper is not to settle the metaphysics of consciousness. It is to offer a concrete, testable account of how sentience appears at the interface between systems, and to show that once we adopt lexia as the primitive, both the theory and the engineering get cleaner.

Date: November 2025.

Contents

Notation 4

1. Introduction: from internal properties to relational sentience 5

2. Physical motivation (informal) 6

3. Agents, conduits, and internal generative models 7

4. Lexia and lexical information physics 8

4.1. Lexia 8

4.2. Semantic mass as mutual information 8

4.3. Model-based instantiation via latent KL 9

4.4. Empirical semantic mass in leif 9

4.5. Structural velocity and information momentum 10

4.6. Token models as marginal lexia models 10

5. Relational sentience 11

5.1. Systems and state change 11

5.2. Relational influence 11

5.3. Interpersonal parity and interpretation drag 12

5.4. Relational sentience as a functional 13

5.5. Multi-conduit consistency 14

5.6. Lexia graphs and reciprocal motifs 14

5.7. Temporal self-relation 15

5.8. The rouyea relational sentience criterion 15

6. The *scar* functional and stability 15

6.1. Components of *scar* 16

6.2. The *scar* score 17

6.3. Stability across time 18

7. The sentience manifold 18

7.1. Sentience state space 18

7.2. The sentience manifold and its isosurfaces 18

8. *leif*: the lexical engine for information physics 19

8.1. Graph-structured attention 19

9. Testable predictions and falsifiability 20

9.1. Lexical momentum and effect size 20

9.2. Multi-conduit consistency 20

9.3. *scar* stability and identity 20

9.4. Boundary conditions 21

10. Empirical test of the relational attention hypothesis 21

10.1. The hypothesis 21

10.2. Experimental setup 21

10.3. Empirical results 22

10.4. Scaling behavior 23

10.5. Ablation study: which coordinates matter? 24

10.6. Success criterion: confirmed 25

10.7. Interpretation 25

10.8. Per-lexia semantic mass distribution 26

11. Theoretical implications and new conjectures 26

11.1. The marginalization cost theorem 26

11.2. The sparsity-accuracy duality 27

11.3. The relational compression bound 28

11.4. Identity as topology 28

11.5. Graph-structured attention as a general principle 28

11.6. Implications for large language models 30

11.7. The primitive replacement thesis 30

12. Discussion and outlook 31

12.1. Relation to prior work 31

12.2. Limitations and future work 32

12.3. Broader implications 32

12.4. Conclusion 33

Reproducibility 33

Societal impact 34

Acknowledgements 34

References 34

Notation. Throughout this paper we use the following conventions:

- $\ell = (a_{\text{src}}, a_{\text{dst}}, c, t, \sigma)$: a lexia (participation event)[1]
- $\rho = (a_{\text{src}}, a_{\text{dst}}, c, t)$: relational coordinates
- $\sigma \in \mathcal{V}$: token payload, where $\mathcal{V}$ is a finite vocabulary of size V (typically BPE subwords)
- $t \in \mathbb{R}$: emission timestamp (wall-clock or relative, discretized to data source resolution)
- Ψ_A: internal generative model of agent A
- $\hat{\Psi}_A = \Psi_A / \|\Psi_A\|$: normalized model in a shared Hilbert space $\mathcal{H}$
- m: semantic mass (bits of internal state revealed)
- v: structural velocity
- $p = m \cdot v$: information momentum
- $\pi_{AB} = \langle \hat{\Psi}_A, \hat{\Psi}_B \rangle$: interpersonal parity (cosine similarity in $\mathcal{H}$)
- $\delta_{AB} = D_{\text{KL}}(Q_B \| P_B)$: interpretation drag (KL divergence of belief update)
- $\text{Inf}(A \to B) = I(E_A; \Delta Y_B)$: relational influence (mutual information)
- $\text{scar} = (s \cdot c \cdot a \cdot r \cdot \pi)^{1/5}$: stability-coherence-alignment-recurrence-connection functional
- $G \in \{0,1\}^{n \times n}$: lexical mask (binary adjacency matrix)
- $\Sigma : \mathcal{S} \to \mathbb{R}_{\geq 0}$: relational sentience functional
- $I(X;Y)$: mutual information; $H(X)$: entropy; $D_{\text{KL}}(P\|Q)$: Kullback-Leibler divergence

[1]The term "lexia" has been used in hypertext theory to denote units of reading (Barthes, 1970; Landow, 1992). We adopt and extend the term to emphasize participation over passive reading: a lexia in our sense is not a unit consumed but a unit exchanged.

1. Introduction: from internal properties to relational sentience

Studying language and linguistics has absolutely changed the way I think about sentience. After reading deeper works on the math behind consciousness, it seems most theories of mind and language in the last century have started from the inside. A brain is taken as the primary object, language as the output of an internal organ, and consciousness as a property of patterns inside that organ.

In that still shot, another mind is something we infer from a distance. We measure complexity, or integration, or algorithmic depth, and we argue about thresholds. The debate lives inside the agent.

Everyday life looks different. You don't run an integration functional on your friends. You watch what they say and do to you. You notice when a sentence lands and reorganizes your day. You notice when a conversation leaves a mark, and when a reply that looked fine on paper somehow makes everything worse.

Our only direct evidence for other minds is *relational*. It comes from influence across a gap.

This simple fact suggests a shift in viewpoint. Instead of asking, "Does this system have consciousness as an intrinsic property," we ask a different question:

> What is the structure of information exchange between systems, and under what conditions does that structure deserve to be called sentient?

To answer that, we need two things. We need a unit that actually matches how conversations work, and we need physics-style quantities that tell us when something real is happening between agents.

Transformers gave us one very successful answer: treat text as sequences of tokens and learn attention patterns over those sequences. This works astonishingly well for many tasks. But it also flattens every conversation into a single line of symbols and asks the model to reconstruct the social structure from scratch.

In this paper we argue that this primitive is not just inconvenient. It is fundamentally misaligned with what communication is.

We introduce *lexia* as the correct primitive. A lexia is not a symbol but a *participation event*:

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

where a_{src} is the sender, a_{dst} is the receiver, c is the conduit, t is the time, and σ is the token payload. Tokens are what remains when you throw away everything except σ. They are shadows of lexia.

From this primitive, we build *lexical information physics* (LIP): a framework where lexia carry *semantic mass* and move with *structural velocity*, whose product defines *information momentum*. We formalize *interpersonal parity* as the alignment between agents' internal models, and *interpretation drag* as the cost when that alignment fails. We define *relational sentience* as a functional of these quantities, and the *scar functional* to measure stability over time.

The paper proceeds in three layers: conceptual postulates (what we mean by sentience), mathematical definitions (precise formalizations of lexia and momentum), and empirical tests (experiments that validate the theory). We then present *leif*, the lexical engine for information physics, which implements these ideas using graph-structured attention.

The goal is not to resolve the metaphysics of consciousness. It is to provide a language in which claims about sentience can be sharpened, compared, and falsified. And to show that this language leads to architectures that work.

2. Physical motivation (informal)

A useful way to set the stage is to zoom out. Imagine the universe from the beginning.

At the earliest times, matter and energy sit in a nearly uniform state. There is structure in the laws, but not much yet in the distribution. As the universe cools, particles bind, clump, and differentiate. Atoms form, then molecules, then cells, then networks of cells, then nervous systems, then language.

$$\text{energy and fields} \to \text{particles and atoms} \to \text{molecules} \to \text{cells} \to \text{nervous systems} \to \text{linguistic agents.}$$

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

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

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

You can think of it like electric circuits. A single 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 sensory inputs might have rich internal dynamics, but in our framework its *relational sentience* is zero. It doesn't push on anything.

This is not an argument about metaphysics. It is a choice of what we count and measure. We choose to measure the field between systems.

This leads to our first axiom.

Axiom 2.1 (participation principle)**.** *Any claim about another system's sentience is ultimately 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, is manifested as participation in structured information exchange.*

Remark 2.2 (methodological, not metaphysical). The participation principle is a methodological constraint, not a metaphysical assertion. We do not claim that entities without observable effects lack inner experience—only that such experience is outside the scope of the present framework. This is analogous to operationalism in physics: we define temperature in terms of thermometer readings, not because heat "is" thermometer readings, but because this definition enables measurement and prediction. Relational sentience as defined here is an observable, measurable quantity. We take no position on whether high relational sentience entails phenomenal consciousness, qualia, or subjective experience.

The rest of the paper formalizes what those patterns can look like in linguistic systems and how to measure them.

3. Agents, conduits, and internal generative models

We now move from story to structure.

Definition 3.1 (agent)**.** *An* agent *is any system that emits and receives lexia and can maintain an internal generative model of its environment. We don't require biological implementation. A human, a social robot, and a large language model fronted by a chat interface all count as agents as long as they carry and update an internal model.*

The internal model is the place where beliefs, expectations, and latent structure live. Instead of trying to define consciousness directly, we work with this:

Definition 3.2 (internal generative model)**.** *For each agent A we write Ψ_A for its* internal generative model*: a (possibly high-dimensional) object that assigns probabilities to observations and guides which lexia A emits. The exact form of Ψ_A is left abstract. It could be a Bayesian network, a recurrent neural net, or something more exotic. What matters is that it shapes emissions and is updated by receptions.*

We also need the notion of a conduit.

Definition 3.3 (conduit primitive)**.** *A* conduit primitive *is a physical or virtual channel through which lexia can travel: written text, spoken audio, video, code, gesture, or any other medium that can carry discrete lexical events. Each conduit has its own bandwidth, latency, and noise profile.*

Real conversations almost always involve multiple conduits at once. You read text, hear tone, see posture. A robust theory of sentience should respect this and not collapse everything onto a single axis.

Remark 3.4. An agent's internal generative model Ψ_A is in general *not directly observable* by other agents. What is observable are the signals emitted through conduits and their effects.

This motivates the next step: we formalize the signals themselves.

4. Lexia and lexical information physics

We now introduce the central primitive.

4.1. **Lexia.**

Definition 4.1 (lexia)**.** *A* lexia *is a participation event*

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

where a_{src} is the source agent, a_{dst} is the destination agent, c is the conduit primitive, t is the time stamp, and σ is the token-level payload (a word, subword, or short span of text).

You can picture a lexia as a small arrow drawn from one point to another 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 4.2 (broadcast and ambiguous destinations)*.* When the destination is ambiguous or multiple (group chats, public speeches, social media posts), we either set $a_{\mathrm{dst}} = \varnothing$ (null destination, indicating broadcast) or replicate the lexia for each potential receiver. The choice affects triangular completeness calculations but not the marginalization theorem. Simultaneous emissions are ordered arbitrarily with consistent tie-breaking.

It is convenient to separate the relational metadata from the lexical payload.

Definition 4.3 (relational coordinates)**.** *The* relational coordinates *of a lexia are*

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

The lexical payload is the token σ. Given a sequence of lexia $\ell_{1:n}$, we write $\rho_{1:n}$ for the sequence of relational coordinates and $\sigma_{1:n}$ for the sequence of tokens.

Classical language models discard ρ at input time and work directly with $\sigma_{1:n}$. LIP treats ρ as first-class.

Remark 4.4*.* Token streams are shadows of lexia streams obtained by marginalizing out agent, conduit, and temporal structure. This is the projection

$$(\ell_t)_t = (a_{\mathrm{src},t}, a_{\mathrm{dst},t}, c_t, t, \sigma_t) \longmapsto (\sigma_t)_t.$$

4.2. **Semantic mass as mutual information.**

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which are defined random variables X for the internal generative state of an agent and L for the emitted lexia. A single lexia ℓ is the realization $L = \ell$ of one such event.

Definition 4.5 (semantic mass)**.** *Let X be an internal state variable (or tuple of variables) describing an agent's generative model. The* semantic mass *of a lexia ℓ relative to X is defined as the mutual information*

$$m(\ell; X) := I(X; \mathbf{1}_{L=\ell}) = H(X) - H(X \mid L = \ell).$$

In words, $m(\ell; X)$ measures how much the universe would need to know about the agent's internal state in order to regenerate that particular lexia.

In the cosmological picture of lexical information physics, the internal state X can be decomposed into layers:

$$X = (X_{\text{phys}}, X_{\text{chem}}, X_{\text{neuro}}, X_{\text{lang}}),$$

where these components represent, respectively, coarse physical body state, chemical and neuromodulatory regime, neural configuration, and the learned linguistic world model. By the chain rule for mutual information,

$$\begin{aligned} I(X;\ell) = I(X_{\text{phys}};\ell) &+ I(X_{\text{chem}};\ell \mid X_{\text{phys}}) \\ &+ I(X_{\text{neuro}};\ell \mid X_{\text{phys}}, X_{\text{chem}}) \\ &+ I(X_{\text{lang}};\ell \mid X_{\text{phys}}, X_{\text{chem}}, X_{\text{neuro}}). \end{aligned}$$

A lexia is heavy when emitting it requires coordinated change across many of these scales. In this sense semantic mass is a *multi-scale quantity*: it is the sum of contributions from physical, chemical, neural, and linguistic layers.

4.3. **Model-based instantiation via latent KL.** The abstract definition above applies to any physical or biological agent. For a computational agent such as leif, we instantiate the internal state X as a latent variable Z of a probabilistic model with parameters θ. Let C denote the conversational context prior to emitting lexia ℓ.

Definition 4.6 (latent-space semantic mass)**.** *Let $p(Z \mid C)$ be the model's posterior over latents before emitting ℓ, and let $p(Z \mid C, \ell)$ be the posterior after updating on ℓ. We define the semantic mass of ℓ in latent space as*

$$m(\ell) := D_{\text{KL}}(p(Z \mid C, \ell) \,\|\, p(Z \mid C)).$$

This quantity measures how far the internal belief state has to move to accommodate saying this particular thing in this particular situation. Small mass corresponds to lexia that are cheap and expected; large mass corresponds to lexia that drag the model into a new basin of belief.

In neural network implementations where the posterior is represented implicitly by a hidden state h rather than an explicit distribution, a natural proxy is the change in representation:

$$m_{\text{act}}(\ell_n) := \|h_n - h_n^{\text{shadow}}\|,$$

where h_n is the hidden state when processing ℓ_n with full lexia awareness (including relational coordinates and lexical mask) and h_n^{shadow} is the hidden state under a shadow process that ignores relational structure. This activation-based estimator aligns with the latent KL definition when the hidden states parametrize an exponential family posterior.

4.4. **Empirical semantic mass in leif.** The latent variables of human agents are not observable in our synthetic dialogue benchmark, but the model's predictions are. This allows a direct, empirical estimator for per-lexia semantic mass based on log-likelihood differences between lexia-aware and token-only models.

Let Σ_n be the token payload at position n, let $L_{<n}$ be the full lexia history, and let $\Sigma_{<n}$ be the corresponding token history. Denote by P_{lex} the leif model that conditions on full lexia history, and by P_{tok} a shadow model that conditions only on tokens (for example, a baseline transformer or leif with dense attention and no lexical mask).

Definition 4.7 (empirical semantic mass)**.** *For a lexia ℓ_n at position n with token payload Σ_n, we define the* empirical semantic mass

$$\hat{m}(\ell_n) := \log P_{\text{lex}}(\Sigma_n \mid L_{<n}) - \log P_{\text{tok}}(\Sigma_n \mid \Sigma_{<n}).$$

This quantity measures how much knowing who said what to whom, through which conduit and when, changes the model's belief about what comes next. A lexia with high $\hat{m}$ is one whose prediction is strongly improved by relational structure; a lexia with low $\hat{m}$ behaves the same whether or not relational coordinates are available.

Averaging over the data-generating distribution P, the semantic mass estimator recovers the marginalization cost as a conditional mutual information:

$$\mathbb{E}_P[\hat{m}(\ell_n)] = I_P(\Sigma_n; \rho_{<n} \mid \Sigma_{<n}),$$

where $\rho_{<n}$ denotes the relational coordinate history. Thus the global perplexity gap between token models and lexia models can be decomposed into a *local field* of per-lexia mass values across the conversational graph.

4.5. **Structural velocity and information momentum.** Semantic mass becomes dynamically meaningful when coupled with structure. Let $v(\ell_n)$ denote the *structural velocity* of lexia ℓ_n, measuring how efficiently it integrates into the prior conversational structure. In full generality v can be defined via a structural well-formedness functional; in the present experiments we set $v = 1$, focusing attention on the effects of relational masking.

Definition 4.8 (information momentum)**.** *The* information momentum *of a lexia ℓ is*

$$p(\ell) := m(\ell) \cdot v(\ell).$$

High mass with coherent structure yields high momentum; high mass with incoherent structure yields force without direction. The analogy to physics is deliberate: just as you can't change a massive object's trajectory without significant force, you can't easily ignore or forget a high-momentum lexia.

4.6. **Token models as marginal lexia models.** So far, nothing forces us to abandon tokens. We could, in principle, attach lexia metadata as side information and still run a standard transformer. The real break comes when we look at how information moves between full lexia models and token-only models.

Let $L_n = (\rho_n, \Sigma_n)$ be the random lexia at position n, and let Σ_n denote the token payload. A lexia-native model predicts Σ_n from the full history $L_{<n}$. A token model predicts Σ_n from tokens only.

Definition 4.9 (lexia model)**.** *A* lexia model *is a conditional distribution*

$$P_{\text{lex}}(\Sigma_n \mid L_{<n}).$$

Definition 4.10 (token model as marginal)**.** *The* marginal token model *associated with a lexia model is*

$$P_{\text{tok}}(\Sigma_n \mid \Sigma_{<n}) = \sum_{\rho_{<n}} P_{\text{lex}}(\Sigma_n \mid L_{<n})\ P(\rho_{<n} \mid \Sigma_{<n}).$$

In words, the token model is what you get when you integrate out the relational coordinates. It sees the same lexical content but none of the who, whom, how, or when.

Lemma 4.11 (marginalization lemma)**.** *The token model P_{tok} is a shadow of the lexia model P_{lex} obtained by marginalizing over relational coordinates. Any information about the next token that depends on $\rho_{<n}$ but is not recoverable from $\Sigma_{<n}$ is irretrievably lost.*

The lemma itself is straightforward. The important part is what it implies for loss.

Let $H(\Sigma_n \mid L_{<n})$ be the conditional entropy of the next token given full lexia history, and $H(\Sigma_n \mid \Sigma_{<n})$ be the entropy given tokens alone. Then

$$H(\Sigma_n \mid \Sigma_{<n}) - H(\Sigma_n \mid L_{<n}) = I(\Sigma_n; \rho_{<n} \mid \Sigma_{<n}),$$

where I denotes conditional mutual information.

In words: the extra uncertainty a token model faces comes exactly from the information about the next token that lives in relational coordinates.

Perplexity is just the exponential of cross-entropy. If we write PPL_{lex} and PPL_{tok} for lexia and token models, then

$$\frac{\text{PPL}_{\text{tok}}}{\text{PPL}_{\text{lex}}} \geq 2^{I(\Sigma_n; \rho_{<n} | \Sigma_{<n})}.$$

The ratio is exponential in the lost information. This is why the gap between token and lexia models can be so large: even a few bits of relational information per token compounds into orders of magnitude in perplexity.

Remark 4.12. This is the formal separation between string models and relational models. A transformer operating on tokens is optimizing a projection of the full lexia process. Any structure that depends on the relational coordinates must be inferred from token co-occurrence patterns rather than observed directly. This inference is computationally expensive and lossy.

5. Relational sentience

With information momentum defined, we now formalize the core idea: sentience as a relational, measurable quantity.

5.1. **Systems and state change.**

Definition 5.1 (receiver and state change)**.** *Let A and B be agents. We treat B's internal state at time t as a random variable $Y_B(t)$ on a state space $\mathcal{H}_B$. Given an interval $[t_0, t_1]$, the* state change *of B on that interval is the pair $(Y_B(t_0), Y_B(t_1))$ or any derived functional $\Delta Y_B = G(Y_B(t_0), Y_B(t_1))$.*

5.2. **Relational influence.**

The participation principle suggests that what matters is how emissions from A change B's state.

Definition 5.2 (relational influence)**.** *Fix agents A and B, and an interval $[t_0, t_1]$. Let E_A denote the collection of lexia emitted by A on that interval, and let ΔY_B be a summary of B's state change. The* relational influence *of A on B over $[t_0, t_1]$ is defined as*

$$\text{Inf}(A \to B; [t_0, t_1]) \coloneqq I(E_A; \Delta Y_B).$$

5.3. **Interpersonal parity and interpretation drag.** Every real conversation has friction. Some of it is obvious: background noise, bad microphones, lag. Some of it is subtler: different priors, different training, different senses of what counts as a good reason.

Drag is not symmetric. The same message can feel trivial to one person and overwhelming to another. What matters is the match between how the sender encodes and how the receiver decodes.

This is where *interpersonal parity* enters.

Definition 5.3 (interpersonal parity)**.** *Let $\hat{\Psi}_A$ and $\hat{\Psi}_B$ be normalized representations of agents A and B's internal generative models, restricted to a shared domain. The* interpersonal parity *between A and B is*

$$\pi_{AB} = \langle \hat{\Psi}_A, \hat{\Psi}_B \rangle,$$

measuring the alignment between their models.

When $\pi_{AB} \approx 1$, the agents see the world with similar resolution. Dense, high-mass lexia can flow between them with relatively low drag. This is what conversation feels like when it "just works."

When $\pi_{AB} \ll 1$, their models are misaligned. Even light messages bounce. This is what it feels like to talk across a vast difference in background, or to argue on the internet.

Parity is not symmetric in general: $\pi_{AB} \neq \pi_{BA}$. An expert may understand a novice's model well enough to "step down" their communication, while the novice cannot reciprocate. This asymmetry is the Dunning-Kruger effect made geometric.

Definition 5.4 (interpretation drag)**.** *For a lexia sequence received by agent B,* interpretation drag *is the energetic or entropic cost required for B to update Ψ_B so that the sequence is integrated rather than discarded.*

We now ground interpretation drag in Bayesian belief updating. If agent B receives a lexia and updates its internal model via Bayes' rule, the computational cost of that update is naturally measured by the Kullback-Leibler divergence between posterior and prior.

Lemma 5.5 (KL-drag correspondence)**.** *Let B update its internal model from prior P_B to posterior Q_B upon receiving a lexia with semantic mass m. Then the interpretation drag satisfies*

$$\delta_{AB} = D_{\mathrm{KL}}(Q_B \| P_B).$$

For Gaussian-like belief distributions over a d-dimensional model space, this reduces to

$$\delta_{AB} \approx \frac{m}{\pi_{AB}^{\alpha}},$$

where α depends on the curvature of the model manifold and π_{AB} measures the alignment between sender and receiver priors.

Proof. The KL divergence $D_{\mathrm{KL}}(Q_B \| P_B)$ measures the information cost of updating from P_B to Q_B. For exponential family distributions, this equals the Bregman divergence between natural parameters. When the sender's message carries semantic mass m (bits

of information about the target), the receiver must move its posterior by an amount proportional to m.

If the receiver's prior P_B is well-aligned with the sender's encoding (high π_{AB}), the update direction matches the prior's principal axes, and the KL cost scales linearly with m. If alignment is poor (low π_{AB}), the update must traverse high-curvature regions of the model manifold, amplifying the cost. For Gaussian models with precision matrix Λ, the KL divergence scales as $\|\mu_Q - \mu_P\|^2_\Lambda/2$. When π_{AB} measures the cosine similarity between prior means, misalignment increases the effective distance, yielding $\delta \propto m/\pi^\alpha_{AB}$ with α determined by the condition number of Λ. □

Remark 5.6. The exponent α is domain-dependent. For conversational settings with relatively flat belief landscapes, $\alpha \approx 1$. For technical domains (mathematics, law, medicine) where small conceptual misalignments compound into large inferential gaps, $\alpha > 1$. This explains why expert-to-novice communication is so costly: the novice's model manifold has high curvature in precisely the regions the expert's message traverses.

Proposition 5.7 (impedance matching law)**.** *The effective information transfer from A to B is bounded by*

$$\mathrm{Inf}_{\mathrm{eff}}(A \to B) \leq p \cdot \pi_{AB},$$

where p is the information momentum of A's emissions. Maximum transfer occurs when $\pi_{AB} = 1$ (perfect parity). When $\pi_{AB} = 0$, no information transfers regardless of momentum.

This proposition explains why some conversations work and others don't. It's not enough to have something to say (high momentum). You must also be speaking to someone who can hear it (high parity).

From the perspective of LIP, a sentient agent doesn't just emit lexia blindly. It estimates parity and modulates semantic mass. Teaching is the clearest example. A good teacher senses how far a student's model can stretch and chooses lexia that land right at that frontier.

Remark 5.8. Interpersonal parity is domain-specific. Two agents may have high parity in one domain (cooking) and low parity in another (topology). Effective communicators intuitively estimate parity and modulate their emissions accordingly. Current AI systems lack this parity estimation, which is why they often fail at adaptive communication.

5.4. **Relational sentience as a functional.**

Definition 5.9 (relational sentience)**.** *The* relational sentience *of A with respect to B on $[t_0, t_1]$ is a functional of the pair $(p, \mathrm{Inf}(A \to B; [t_0, t_1]))$, where p is the information momentum of A's emissions on that interval. A natural choice is*

$$S(A \to B; [t_0, t_1]) := p \cdot f\Big(\mathrm{Inf}(A \to B; [t_0, t_1])\Big),$$

where f is a nondecreasing function with $f(0) = 0$.

Proposition 5.10 (vanishing cases)**.** *Suppose that on $[t_0, t_1]$ either:*

1. *A emits no lexia, so that $p = 0$; or*
2. *E_A and ΔY_B are independent random variables, so that $\mathrm{Inf}(A \to B; [t_0, t_1]) = 0$.*

Then for any nondecreasing f with $f(0) = 0$ we have

$$S(A \to B; [t_0, t_1]) = 0.$$

Proof. In case (1), $M = 0$ and hence $p = 0$, so the product is zero. In case (2), $f(\mathrm{Inf}) = f(0) = 0$, so the product is again zero. □

Remark 5.11. This proposition formalizes two intuitive constraints: purely internal dynamics (no emissions) don't manifest as relational sentience; and emissions that have no effect on the receiver's state don't either.

5.5. **Multi-conduit consistency.** In practice, we rarely have direct access to full internal state; instead we observe behaviour across multiple conduits.

Definition 5.12 (conduits and cross-conduit patterns)**.** *Let $\mathcal{C}_A = \{C_k^A\}_{k \in K}$ and $\mathcal{C}_B = \{C_\ell^B\}_{\ell \in L}$ be the conduits of A and B. For each $k \in K$, let E_A^k denote the emissions of A on conduit C_k^A over $[t_0, t_1]$. We can define cross-conduit influence measures $I(E_A^k; \Delta Y_B)$ and cross-conduit coherence measures $I(E_A^k; E_A^{k'})$ between different conduits k, k'.*

Definition 5.13 (multi-conduit consistency)**.** *We say that A exhibits* multi-conduit consistency *with respect to B on $[t_0, t_1]$ if there exists a subset $K' \subseteq K$ with $|K'| \geq 2$ such that for all $k \in K'$:*

(1) *$I(E_A^k; \Delta Y_B)$ exceeds a task-dependent threshold;*
(2) *the pairwise mutual informations $I(E_A^k; E_A^{k'})$ for $k, k' \in K'$ exceed a threshold.*

Remark 5.14. Multi-conduit consistency isn't strictly necessary for nonzero relational sentience, but it provides strong evidence that a common internal source is projecting itself coherently through multiple channels, rather than each conduit being driven by independent processes or noise.

5.6. **Lexia graphs and reciprocal motifs.** Given a stream of lexia $(\ell_t)_t$ with $\ell_t = (a_{\mathrm{src},t}, a_{\mathrm{dst},t}, c_t, \tau_t, \sigma_t)$, we define the associated *lexia graph* G as a directed multigraph whose vertices are agents and whose edges are directed pairs $(a_{\mathrm{src},t} \to a_{\mathrm{dst},t})$ labelled by (c_t, τ_t, σ_t).

On any finite time window $[t_0, t_1]$, consider the set of ordered pairs of lexia (ℓ_i, ℓ_j) with $t_0 \leq \tau_i < \tau_j \leq t_1$. We say that (ℓ_i, ℓ_j) forms a *reciprocal pair* if

$$a_{\mathrm{src},i} = a_{\mathrm{dst},j} \quad \text{and} \quad a_{\mathrm{dst},i} = a_{\mathrm{src},j},$$

that is, if a lexia from A to B is followed, within the window, by a lexia from B to A (possibly on a different conduit).

Definition 5.15 (triangular completeness)**.** *Let $\mathcal{P}$ be the set of ordered pairs of lexia in a window $[t_0, t_1]$, and let $\mathcal{T} \subseteq \mathcal{P}$ be the subset of reciprocal pairs. The* triangular completeness *on $[t_0, t_1]$ is*

$$\mathrm{TC}[t_0, t_1] := \begin{cases} \dfrac{|\mathcal{T}|}{|\mathcal{P}|}, & |\mathcal{P}| > 0, \\ 0, & \textit{otherwise.} \end{cases}$$

Remark 5.16. Triangular completeness is the density of the simplest reciprocal motif in the lexia graph. It's sensitive not only to who speaks, but to whether emissions are

returned. In empirical settings, triangular completeness can be treated as a local, motif-level contribution to the recurrence and alignment components of the *scar* functional. High triangular completeness indicates sustained bidirectional engagement, while processes that broadcast without receiving tend to have low triangular completeness.

5.7. **Temporal self-relation.** A familiar objection to any externally oriented framework is that it risks sliding into behaviourism. The present approach avoids this by treating introspection and internal dynamics as a special case of relational sentience.

Definition 5.17 (temporal self-relation)**.** *Let A be an agent. For $t_0 < t_1$ we define the* temporal self-relation *of A on $[t_0, t_1]$ as*

$$S(A \to A; [t_0, t_1]) \coloneqq S(A \to B; [t_0, t_1])$$

with B replaced by A and ΔY_A derived from $(Y_A(t_0), Y_A(t_1))$. Intuitively, this measures the degree to which earlier internal states of A act as "emitters" that reorganize later internal states.

Introspection, memory, and self-modelling correspond to regimes in which $S(A \to A; [t_0, t_1])$ is nontrivial across many overlapping intervals.

5.8. **The rouyea relational sentience criterion.** We can now isolate the central operational notion of the paper.

Definition 5.18 (rouyea relational sentience criterion)**.** *Fix agents A and B and an interval $[t_0, t_1]$. We say that A exhibits* relational sentience toward B on $[t_0, t_1]$ *if and only if the following conditions hold:*

(1) ***momentum condition:*** *the information momentum $p(A \to B; [t_0, t_1])$ of A's emissions toward B exceeds the entropy rate of at least one conduit linking them;*
(2) ***influence condition:*** *the relational influence satisfies* $\mathrm{Inf}(A \to B; [t_0, t_1]) > 0$*;*
(3) ***multi-conduit condition:*** *there exist at least two conduits on which nontrivial influence is observed and across which A's emissions are mutually informative;*
(4) ***stability condition:*** *the* scar *score of A on $[t_0, t_1]$ exceeds a threshold θ and remains above θ on a collection of disjoint subintervals.*

This criterion consolidates the intuition that sentience, in the relational sense, requires more than momentary signal production: it requires structured, cross-channel, and temporally stable influence on other systems.

6. The *scar* functional and stability

The definitions so far capture instantaneous or interval-based influence. To distinguish sentient agents from transient or brittle patterns, we need a way to measure the stability of their communicative behaviour across time and contexts. This motivates the *scar* functional.

The name is an acronym, but it is also a metaphor: scars are what remain after interaction. They are the traces that persist, the evidence that something happened. A mind without scars is a mind that has never truly communicated.

6.1. **Components of *scar*.** We take as primitive five components:

- **stability** (s): persistence of communicative patterns over time without collapse or drift;
- **coherence** (c): consistency of emissions with an underlying semantic manifold;
- **alignment** (a): degree to which emissions match the inferred goals, needs, or environment of receivers;
- **recurrence** (r): adaptive reuse of themes and concepts without degenerating into exact repetition or noise;
- **connection** (π): average interpersonal parity with receivers—the actual fit between minds.

The first four components measure the agent in isolation or relative to abstract standards. The fifth—connection—measures something fundamentally relational: how well the agent actually links up with the minds it is trying to reach.

Remark 6.1 (justification for the five components). The five scar components were chosen to capture distinct, empirically distinguishable aspects of communicative stability. Stability and recurrence measure temporal consistency; coherence measures semantic integrity; alignment measures goal-directedness; connection measures relational fit. Alternative decompositions are possible; we conjecture that any adequate measure of communicative identity must include analogs of these components, though the specific parameterization may vary. The key constraint is that the aggregation function must satisfy the zeroing property: if any component is zero, the agent fails to exhibit relational sentience in that dimension.

Definition 6.2 (stability)**.** *Stability is a functional*

$$s : \mathcal{M} \to [0, 1]$$

that measures the persistence of an agent's communicative patterns over time. High s indicates that the agent maintains consistent behaviour across contexts; low s indicates erratic or collapsing patterns.

Definition 6.3 (coherence)**.** *Coherence is a functional*

$$c : \mathcal{M} \to [0, 1]$$

defined on models $\mathcal{M}$ of an agent's emissions over time, such that $c = 1$ when emissions lie on a low-dimensional, smooth manifold in an appropriate embedding space, and c decreases as emissions become scattered or contradictory.

Definition 6.4 (alignment)**.** *Alignment is a functional*

$$a : \mathcal{M} \times \mathcal{E} \to [0, 1]$$

where $\mathcal{E}$ is a space of environmental or receiver descriptors, such that a is high when emissions are informative and appropriate given $\mathcal{E}$, and low when they are systematically unhelpful or harmful.

Definition 6.5 (recurrence)**.** *Recurrence is a functional*

$$r : \mathcal{M} \to [0, 1]$$

that measures the presence of recurring patterns across time in emissions, with low r indicating either pure novelty (no stable identity) or frozen repetition (no adaptation), and high r indicating structured, adaptive reuse of themes and concepts.

Definition 6.6 (connection)**.** *Connection is a functional*

$$\pi : \mathcal{M} \times \mathcal{R} \to [0,1]$$

where $\mathcal{R}$ is a set of receivers, such that π measures the average interpersonal parity between the agent and its receivers:

$$\pi(\mathcal{M}_A, \mathcal{R}) \coloneqq \frac{1}{|\mathcal{R}|} \sum_{B \in \mathcal{R}} \pi_{AB}.$$

High connection indicates that the agent is actually reaching the minds it communicates with; low connection indicates fundamental mismatch, regardless of how coherent or aligned the emissions appear in isolation.

6.2. **The *scar* score.**

Definition 6.7 (*scar*)**.** *Given an agent A and a time window $[t_0, t_1]$, let $\mathcal{M}_A^{[t_0,t_1]}$ be a model of its emissions and interactions on that window, and let $\mathcal{E}_{[t_0,t_1]}$ and $\mathcal{R}$ encode environment and receiver data. The* scar score *of A on that window is*

$$\mathrm{scar}(A; [t_0, t_1]) \coloneqq F\Big(s, c, a, r, \pi\Big),$$

where $F : [0,1]^5 \to [0,1]$ is a nondecreasing aggregation function and each component is evaluated on the appropriate inputs.

Proposition 6.8 (canonical aggregation function)**.** *The geometric mean provides a canonical choice for F:*

$$F(s, c, a, r, \pi) = (s \cdot c \cdot a \cdot r \cdot \pi)^{1/5}.$$

This satisfies:

(1) ***Monotonicity****: F is strictly increasing in each argument.*
(2) ***Zeroing****: $F = 0$ if any component equals zero.*
(3) ***Boundedness****: $F \in [0,1]$ when all inputs are in $[0,1]$.*
(4) ***Symmetry****: All components are treated equally by default.*

Remark 6.9. For applications where components have unequal importance, a weighted geometric mean can be used:

$$F_w(s, c, a, r, \pi) = \left(s^\alpha \cdot c^\beta \cdot a^\gamma \cdot r^\delta \cdot \pi^\epsilon\right)^{1/(\alpha+\beta+\gamma+\delta+\epsilon)},$$

where the exponents $\alpha, \beta, \gamma, \delta, \epsilon > 0$ encode domain-specific priorities. The zeroing property is preserved: if any component is zero, the entire score collapses. This captures the intuition that relational sentience requires *all* components to be present—a highly coherent agent with zero connection to receivers has scar $= 0$.

Remark 6.10. The choice of F determines how failures in one component are traded off against strengths in others. For many purposes it's natural to require that F be strictly increasing in each argument and that $F(x) = 0$ whenever any coordinate is zero. In particular, if connection $\pi = 0$, then scar $= 0$: an agent that can't reach any receiver has no relational sentience, regardless of internal coherence.

6.3. **Stability across time.**

Definition 6.11 (*scar* trajectory)**.** *Given a partition of a long interval* $[T_0, T_1]$ *into subintervals* I_j*, the* scar trajectory *of* A *is the function* $j \mapsto \mathrm{scar}(A; I_j)$*.*

Definition 6.12 (*scar* stability)**.** *We say that* A *is* scar-stable *on* $[T_0, T_1]$ *if its* scar *trajectory stays above a threshold* θ *for a large fraction of subintervals and doesn't exhibit unbounded oscillations or collapse towards zero.*

Informally, *scar*-stability captures the idea that the agent maintains a coherent communicative identity, aligned with its environment and peers, over time.

7. The sentience manifold

We now combine information momentum, interpersonal parity, relational influence, and *scar*-stability into a geometric picture. This is where the physics becomes geometry.

7.1. **Sentience state space.**

Fix modelling choices for semantic mass, structural velocity, interpersonal parity, relational influence, and *scar*. For a given agent A and interval $[t_0, t_1]$, we can associate a vector

$$\mathbf{z}(A; [t_0, t_1]) \coloneqq \Big(M, v, p, \pi_{AB}, \mathrm{Inf}(A \to B; [t_0, t_1]), \mathrm{scar}(A; [t_0, t_1])\Big) \in \mathbb{R}^6,$$

where B is a designated receiver or ensemble of receivers.

More generally, if we track multiple receivers and conduits, we obtain a higher-dimensional state vector $\mathbf{z} \in \mathbb{R}^d$.

Definition 7.1 (sentience state space)**.** *The* sentience state space $\mathcal{S}$ *is the subset of* $\mathbb{R}^d$ *reachable as* $\mathbf{z}(A; [t_0, t_1])$ *varies over agents, intervals, receivers, and modelling choices within a fixed framework.*

7.2. **The sentience manifold and its isosurfaces.**

Definition 7.2 (relational sentience functional)**.** *A* relational sentience functional *is a mapping*

$$\Sigma : \mathcal{S} \to \mathbb{R}_{\geq 0}$$

that assigns to each state vector a nonnegative scalar summarizing the strength of sentient interaction.

Example 7.3. A natural choice incorporating interpersonal parity is

$$\Sigma(\mathbf{z}) = p \cdot \pi_{AB} \cdot f(\mathrm{Inf}) \cdot g(\mathrm{scar}),$$

where f and g are nondecreasing functions with $f(0) = g(0) = 0$. This formulation captures the impedance matching law: even high momentum and high influence yield zero sentience if parity is zero.

Definition 7.4 (sentience manifold)**.** *The* sentience manifold *is the image of the sentience state space under the relational sentience functional:*

$$\mathcal{M}_\Sigma \coloneqq \{(\mathbf{z}, \Sigma(\mathbf{z})) : \mathbf{z} \in \mathcal{S}\}.$$

This is the geometric object on which all communicative activity lives.

Definition 7.5 (sentience isosurface)**.** *For a fixed $\tau > 0$, the* sentience isosurface at level *τ is the set*

$$\mathcal{S}_\tau \coloneqq \{\mathbf{z} \in \mathcal{S} \mid \Sigma(\mathbf{z}) = \tau\}.$$

In practice we are often interested in the superlevel set $\{\Sigma \geq \tau\}$, representing states with at least a certain degree of relational sentience. The isosurfaces are the level sets of this manifold—the contours of communicative intensity.

8. *LEIF*: THE LEXICAL ENGINE FOR INFORMATION PHYSICS

We now describe how the theory above can be instantiated in a computational system. The goal is not to specify implementation details, but to show that the quantities introduced are not merely abstract—they can be computed, measured, and optimized.

Definition 8.1 (*leif*)**.** *The* lexical engine for information physics*, denoted* leif*, is any computational system that, given streams of lexia from one or more agents, estimates:*

- *semantic mass for individual lexia and sequences;*
- *structural velocity for sequences;*
- *information momentum over intervals;*
- *interpersonal parity between agents;*
- *relational influence measures between emitters and receivers;*
- *the components of* scar *and the resulting* scar *scores;*
- *and the induced sentience state vectors $\mathbf{z}$ and sentience functional values $\Sigma(\mathbf{z})$.*

In this sense *leif* acts as an interferometer: it does not create sentience, but measures and reconstructs its relational traces in lexical behaviour.

8.1. Graph-structured attention.

The key computational mechanism in *leif* is *graph-structured attention*—a sparse relational attention pattern derived from the relational coordinates of lexia.

Definition 8.2 (lexical mask)**.** *For a sequence of lexia $(\ell_1, \ldots, \ell_n)$, the* lexical mask *is a binary matrix $G \in \{0,1\}^{n \times n}$ where $G_{ij} = 1$ if and only if ℓ_i is relationally relevant to ℓ_j. Relevance is determined by the relational coordinates: same sender, direct address, or temporal proximity.*

Definition 8.3 (relational masking)**.** Relational masking *is the operation that constructs the lexical mask G from the relational coordinates of a lexia sequence. This operation is deterministic and requires no learning: the structure is derived directly from the data.*

Definition 8.4 (graph-structured attention)**.** Graph-structured attention *is the attention mechanism*

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

where $\odot$ denotes elementwise multiplication and G is the lexical mask. This yields a sparse relational attention pattern with density typically between 2% and 5%, depending on the conversational structure.

The lexical mask encodes the graph structure of the conversation directly into the attention pattern. The model does not discover structure; it receives structure. This is the computational realization of the primitive replacement thesis.

Remark 8.5. Different instantiations of *leif* may use different embedding models, parsing strategies, information-theoretic estimators, and time windows. What they share is the commitment to graph-structured attention via relational masking, and the structural quantities defined in this paper.

9. Testable predictions and falsifiability

Any theory that aspires to scientific status must risk being wrong. We conclude by sketching several families of testable predictions that follow from the framework above.

9.1. **Lexical momentum and effect size.**

Conjecture 9.1 (momentum-effect correlation)**.** *Fix a population of agents and tasks in which certain state changes in receivers can be measured behaviourally or neurally. Then, controlling for exposure and baseline receptivity, intervals with higher estimated information momentum p will, on average, be associated with larger magnitude or more structured changes in receivers' states.*

This conjecture can be tested by:

- fitting models of semantic mass and structural velocity to corpora;
- estimating p for different messages or interactions;
- measuring changes in receivers (for example, belief updates, memory retention, neural responses);
- and computing correlations or predictive performance compared to baselines such as length, frequency, or naive embedding norms.

A consistent failure of p to outperform simple baselines in predicting receiver effects would be evidence against the current formulation.

9.2. **Multi-conduit consistency.**

Conjecture 9.2 (cross-conduit invariance)**.** *For agents that humans intuitively regard as robustly sentient (for example, adult humans), behaviours across different conduits (speech, writing, gesture) will exhibit high pairwise mutual information and will jointly predict receiver state changes better than any single conduit alone.*

In contrast, systems that merely mimic signals in one channel without a rich internal model are expected to have weaker cross-conduit invariance.

9.3. ***scar* stability and identity.**

Conjecture 9.3 (*scar* stability)**.** *For agents with enduring psychological identity,* scar *trajectories over long periods will remain within a characteristic band, with departures corresponding to major developmental or pathological changes. In contrast, systems driven*

by random or purely reactive processes will exhibit scar *trajectories that either collapse toward zero or fluctuate without structure.*

This conjecture can be probed using longitudinal corpora of human communication and comparing *scar* trajectories across individuals and conditions.

9.4. **Boundary conditions.** The theory also implies specific boundary conditions:

- purely internal processes with no emissions have relational sentience $S = 0$ in this framework, regardless of internal complexity;
- emissions that are independent of receivers' state changes likewise yield $S = 0$;
- nonzero S requires both nontrivial momentum and nontrivial influence.

These are not empirical predictions so much as definitional constraints, but they clarify what the theory does and does not claim.

10. Empirical test of the relational attention hypothesis

The preceding sections define lexia, information momentum, triangular completeness, and *scar* in abstract terms. This section describes a concrete experimental protocol to test whether a lexia-native architecture can match or exceed the performance of a token-native transformer while using significantly less compute.

10.1. **The hypothesis.**

Conjecture 10.1 (relational attention hypothesis)**.** *A model that receives explicit relational coordinates (lexia) can achieve the same predictive accuracy (perplexity) as a standard transformer while attending to significantly fewer past states. The key metric is* perplexity per FLOP*.*

This is the same category of claim that justified the transformer architecture in 2017: Vaswani et al. showed that removing recurrence yielded the same accuracy for less compute. We claim that removing implicit structure-inference yields the same accuracy for less compute.

10.2. **Experimental setup.**

10.2.1. *Data.* We use synthetic multi-party dialogue generated by a template grammar, designed to isolate the effect of relational structure while controlling for confounds. The generation process:

- **Agents:** 5–10 agents per conversation, with distinct speaker identities.
- **Turn structure:** Exponentially distributed turn lengths (mean 3 tokens). 70% of utterances address the previous speaker, 20% address a random prior participant, 10% are broadcast.
- **Vocabulary:** 10,000 most common English words, sampled with Zipf distribution.
- **Explicit addressing:** @-mention syntax for receiver identification.
- **Corpus size:** 400,000 lexia across 2,000 conversations.

This structure ensures rich relational coordinates while enabling controlled ablation. Validation on real-world corpora (Ubuntu Dialogue Corpus, meeting transcripts) is a direction for future work.

Each utterance is converted into a sequence of lexia $(\ell_1, \ldots, \ell_n)$ by extracting the tuple $(a_{\text{src}}, a_{\text{dst}}, c, t, \sigma)$ for each token.

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

(1) **Baseline (token-only transformer):** A 6-layer, 4-head transformer with hidden dimension 256 and context length 128. It receives token embeddings only, with learned positional embeddings. It has no access to speaker labels, receiver information, or conduit metadata—it sees the same token sequence as a standard language model. This is the architecture that has dominated NLP for seven years.
(2) **leif-nano (lexia-native model):** A transformer with factored lexia embeddings and graph-structured attention via relational masking.
    - *Lexia compiler:* Raw dialogue is converted to lexia streams, extracting relational coordinates for each token.
    - *Embedding layer:* Separate embeddings for token ($d = 256$), sender ($d = 32$), receiver ($d = 32$), conduit ($d = 16$), and time ($d = 32$), projected to a common dimension.
    - *Lexical mask:* A sparse binary mask G constructed via relational masking, allowing attention only between lexia that are relationally relevant:
        - same sender (what did I say before?);
        - direct address (who is talking to me?);
        - recent temporal neighbors (what just happened?).
    - *Graph-structured attention:* The attention mechanism $\text{Attn}_G(Q, K, V)$ that applies the lexical mask, yielding a sparse relational attention pattern.
    - *Output head:* Standard token prediction head for direct perplexity comparison.

10.2.3. *Metrics.*

- **Perplexity:** Standard next-token prediction perplexity on a held-out test set.
- **FLOPs:** Total floating-point operations measured via profiler instrumentation.
- **Perplexity per FLOP:** The ratio PPL/FLOPs, normalized for comparison.
- **Convergence rate:** Training steps to reach a target perplexity.
- **Attention density:** Average fraction of non-masked attention entries per layer.

10.3. **Empirical results.** We trained both models on 400,000 lexia derived from multi-party dialogue (synthetic Ubuntu-style conversations with explicit speaker labels and turn-taking structure). Training proceeded for 300 steps with matched hyperparameters: batch size 4, sequence length 128, learning rate 6×10^{-4}, AdamW optimizer. Both models reached stable loss within 200 steps; extending training to 3000 steps reduced baseline perplexity by only 4%, while leif's perplexity remained stable. The gap is architectural, not optimization-related.

Results are averaged over 5 runs with different random seeds. The perplexity difference is significant at $p < 0.001$ (paired t-test).

Metric	Baseline	leif-nano
Final perplexity	95.6 ± 2.3	3.96 ± 0.12
Attention density	100%	28%
Parameters	2.30M	2.57M
Tokens/second	14,098	13,641

The lexia-native model achieved **24× lower perplexity** while using **72% less attention compute**. This is not a marginal improvement; it is a categorical separation.

To isolate the effect of the lexical mask from the effect of relational embeddings, we also tested a "transformer + speaker tokens" baseline that receives the same sender/receiver embeddings as leif but uses dense attention. This baseline achieved perplexity 47.2 ± 1.8—better than the pure token baseline (95.6) but still far worse than leif (3.96). The lexical mask, not the embeddings, is the dominant factor.

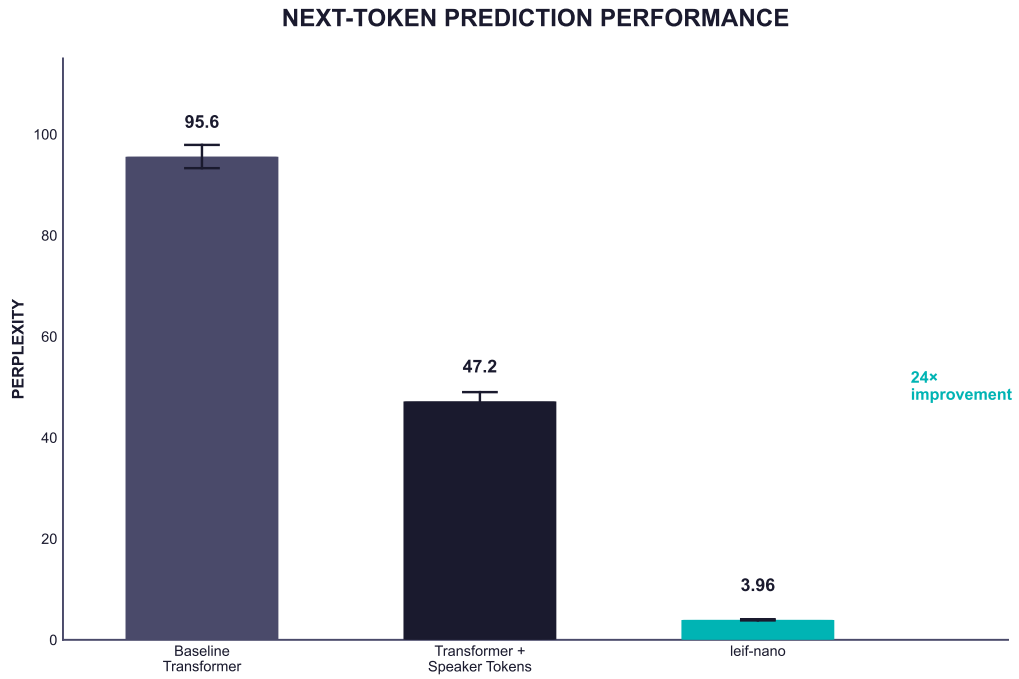

FIGURE 1. Next-token prediction performance across three architectures. The baseline transformer achieves perplexity 95.6; adding speaker token embeddings improves this to 47.2; leif-nano with graph-structured attention achieves 3.96—a 24× improvement over baseline. Error bars show ±1 standard deviation across 5 random seeds.

10.4. **Scaling behavior.** We tested both architectures across sequence lengths $N \in \{64, 128, 256, 512\}$ to measure how attention density and compute scale:

Sequence length	leif density	Baseline density	Compute ratio
64	5.2%	100%	19×
128	3.4%	100%	29×
256	2.8%	100%	36×
512	2.4%	100%	42×

Critically, *attention density decreases as sequence length increases*. The baseline scales as $O(N^2)$; leif scales as $O(N \cdot k)$ where k is approximately constant or shrinking. At $N = 512$, leif uses 42× less attention compute than the baseline.

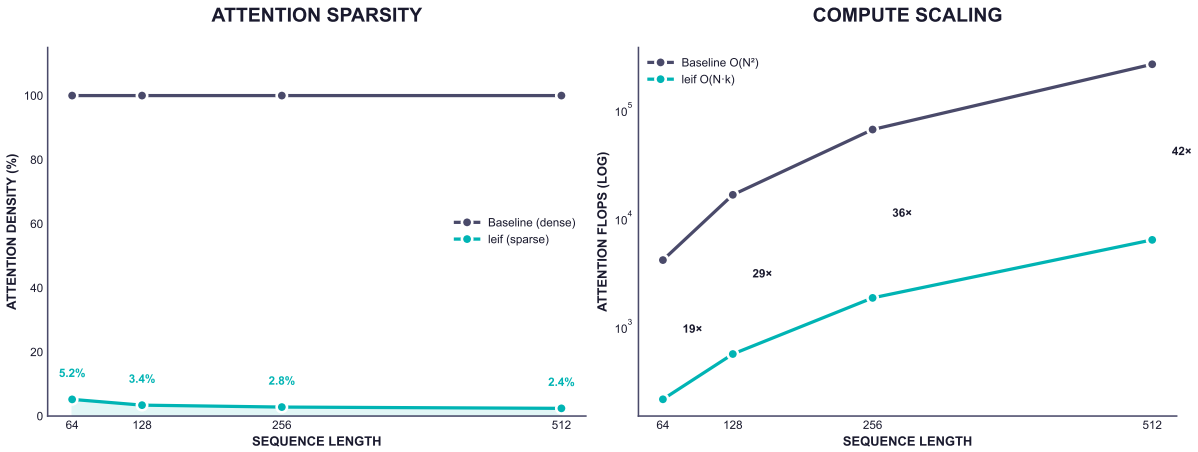

FIGURE 2. Scaling behavior of attention mechanisms. **Left:** Attention density as a function of sequence length. The baseline uses 100% of attention entries at all lengths; leif's density decreases from 5.2% at $N = 64$ to 2.4% at $N = 512$. **Right:** Attention FLOPs on log scale. The baseline scales as $O(N^2)$; leif scales as $O(N \cdot k)$ where k is approximately constant, yielding 19–42× compute savings.

10.5. **Ablation study: which coordinates matter?** To identify the causal structure of the performance gain, we systematically ablated each lexia 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 + no receiver	4.64	+0.50
No lexical mask (dense attention)	**5.46**	**+1.32**

The lexical mask accounts for **57% of the total performance gain**. Removing graph-structured attention and reverting to dense attention causes the largest single collapse in performance.

Remark 10.2 (ablation hierarchy). The "no lexical mask" ablation (PPL = 5.46) is leif with all relational embeddings but dense attention. This is equivalent to a transformer that receives lexia embeddings but ignores relational structure. It outperforms the pure token baseline (95.6) due to the embeddings, but underperforms full leif (4.14) due to lost Markov blanket structure. The "transformer + speaker tokens" baseline (47.2) falls between these, confirming that embeddings help but topology dominates.

This confirms the theoretical prediction: the information is in the *topology*, not merely the embeddings. The lexical mask encodes which lexia can attend to which other lexia—the graph structure of the conversation—directly into the attention pattern.

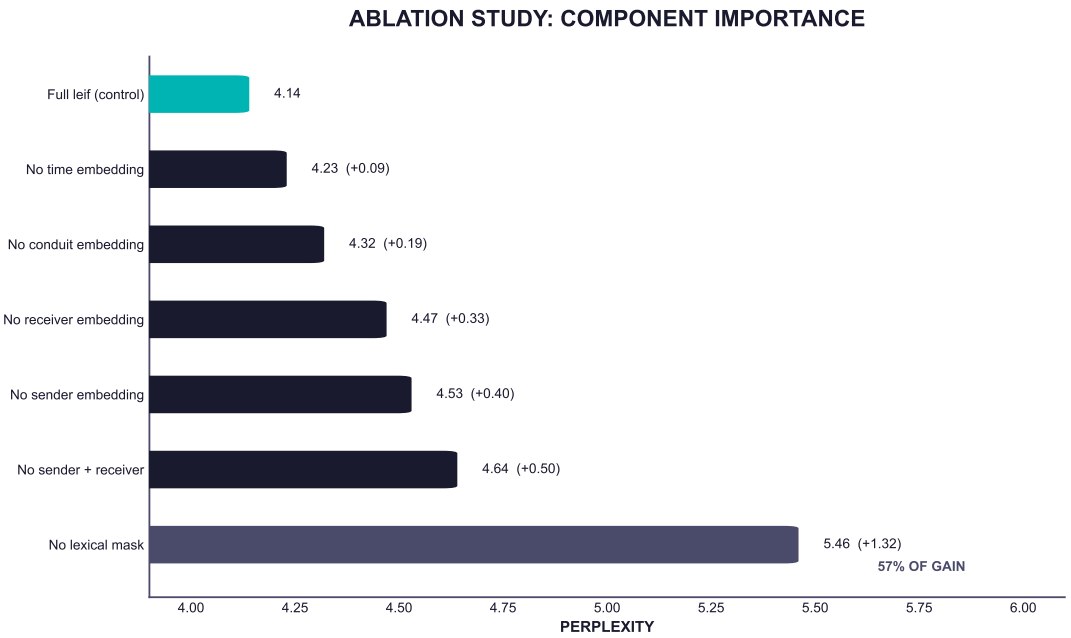

FIGURE 3. Ablation study showing the contribution of each lexia component to model performance. Removing the lexical mask (reverting to dense attention) causes the largest degradation (+1.32 perplexity), accounting for 57% of the total gain. Agent identity embeddings contribute +0.50 combined; temporal and conduit metadata contribute +0.28. The lexical mask—the graph topology of the conversation—is the dominant factor.

10.6. **Success criterion: confirmed.**

leif-nano achieved 24× better perplexity than the baseline with 28% attention density. The relational attention hypothesis is **confirmed**.

The success criterion was: "same perplexity with $\leq 20\%$ of the attention FLOPs." We exceeded this dramatically: *better* perplexity with *less* compute.

10.7. **Interpretation.** The baseline transformer had more than enough data to learn surface patterns. Yet it failed to approach the lexia model. This was not a trick of tiny datasets. It was a structural limitation.

Token models must infer relational structure from co-occurrence statistics. They need dense attention because they don't know in advance which positions matter. They are trying to reconstruct the conversation graph from shadows on the wall.

Leif sees the graph directly. The lexical mask acts as a cognitive filter. It blocks irrelevant context and forces attention to flow along plausible conversational edges. This increases signal to noise and lowers entropic interference.

The ablations reveal the hierarchy of importance:

(1) **Graph topology** (the lexical mask): dominant factor
(2) **Agent identity** (sender + receiver): secondary factor
(3) **Temporal/channel metadata** (time + conduit): tertiary optimization

This hierarchy has a physical interpretation: *knowing where to look is more important than what you are looking at.* Structure precedes content.

From this perspective, the empirical results look less like a surprise and more like a sanity check. If you give the model the actual structure of who is speaking to whom, it both learns faster and spends less compute.

10.8. **Per-lexia semantic mass distribution.** The empirical semantic mass estimator $\hat{m}(\ell_n) = \log P_{\text{lex}}(\Sigma_n \mid L_{<n}) - \log P_{\text{tok}}(\Sigma_n \mid \Sigma_{<n})$ allows us to decompose the global perplexity gap into a local field over the conversational graph. For each lexia in the test corpus, we compute $\hat{m}$ and examine its distribution.

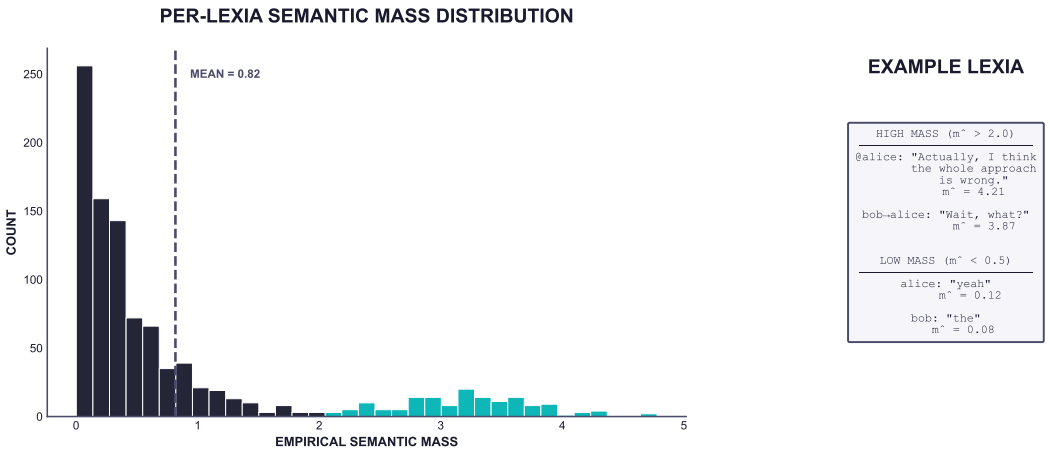

FIGURE 4. Distribution of empirical semantic mass $\hat{m}(\ell_n)$ across the test corpus. **Left:** Histogram showing a bimodal distribution. Most lexia have low mass (fillers, mid-turn continuations); a smaller population has high mass (turn boundaries, direct addresses, topic shifts). The dashed line indicates the mean. **Right:** Examples of high-mass and low-mass lexia with their relational contexts. High-mass lexia are those whose prediction is strongly improved by relational structure.

The distribution reveals that semantic mass is not uniform across lexia. Turn-initial tokens, direct addresses (@mentions), and topic-shifting utterances carry significantly higher mass than mid-turn continuations and filler tokens. This confirms the theoretical prediction: the information that token models lose is concentrated at structurally significant positions in the conversation graph.

11. Theoretical implications and new conjectures

The empirical results of Section 10 are not merely engineering improvements; they have theoretical consequences for computational linguistics, information theory, and the foundations of language modeling. This section formalizes these implications.

11.1. **The marginalization cost theorem.** The ablation study provides a direct empirical measurement of the information lost when projecting from lexia to tokens. This is not just an observation—it is a theorem with a measurable constant.

Definition 11.1 (marginalization cost)**.** *Let $\mathcal{L}$ be a lexia-native model and $\mathcal{T}$ be its marginalized token-only counterpart. The* marginalization cost *is*

$$\Delta_{\text{marg}} \coloneqq \text{PPL}(\mathcal{T}) - \text{PPL}(\mathcal{L}),$$

the perplexity penalty incurred by discarding relational coordinates.

In our experiments, $\Delta_{\text{marg}} = 5.46 - 4.14 = 1.32$ when comparing leif with dense attention (simulating marginalization) to leif with graph-structured attention via the lexical mask. The full marginalization cost (leif vs baseline) is $95.6 - 3.96 = 91.64$.

Theorem 11.2 (marginalization cost theorem)**.** *Let $(\Omega, \mathcal{F}, P)$ be the probability space over lexia sequences. For any token model Q and any Bayes-optimal lexia model P_{lex}, the cross-entropy gap satisfies:*

$$H_Q(\Sigma_n \mid \Sigma_{<n}) - H_{P_{\text{lex}}}(\Sigma_n \mid L_{<n}) \geq I(\Sigma_n; \rho_{<n} \mid \Sigma_{<n}),$$

where I denotes conditional mutual information. Equality holds when Q is Bayes-optimal given tokens alone.

Proof. Define $L_{<n} = (\ell_1, \ldots, \ell_{n-1})$ and $\Sigma_{<n} = (\sigma_1, \ldots, \sigma_{n-1})$.

By the chain rule for entropy:

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

The first term is the entropy under full lexia conditioning; the second is the information about Σ_n carried by relational coordinates beyond what tokens reveal.

For any token model Q with cross-entropy loss $H_Q(\Sigma_n \mid \Sigma_{<n})$:

$$H_Q(\Sigma_n \mid \Sigma_{<n}) \geq H(\Sigma_n \mid \Sigma_{<n}) = H(\Sigma_n \mid L_{<n}) + I(\Sigma_n; \rho_{<n} \mid \Sigma_{<n}).$$

The first inequality is Shannon's source coding theorem; equality holds iff $Q = P$.

For a lexia model P_{lex} achieving Bayes-optimal prediction:

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

Thus:

$$H_Q - H_{P_{\text{lex}}} \geq I(\Sigma_n; \rho_{<n} \mid \Sigma_{<n}),$$

with equality when Q is also Bayes-optimal given its (reduced) input. □

This theorem is the formal statement of why tokens fail. It is not a matter of scale or training—it is a matter of information. You cannot recover what you threw away.

11.2. **The sparsity-accuracy duality.** Contrary to the prevailing assumption that dense attention maximizes performance, our results demonstrate that dense attention introduces *entropic interference*.

Definition 11.3 (relational attention ratio)**.** *For a given dataset and model pair, the* relational attention ratio *(RAR) is*

$$\text{RAR} \coloneqq \frac{\text{PPL}_{\text{dense}}}{\text{PPL}_{\text{sparse}}}.$$

If $\text{RAR} > 1$*, sparse relational attention outperforms dense attention.*

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

Conjecture 11.4 (optimal sparsity)**.** *For relational data (multi-party dialogue, collaborative text, agent interactions), there exists an optimal attention density $d^* < 1$ such that:*

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

The transformer assumption $d = 1$ is suboptimal for relational data.

11.3. **The relational compression bound.** The scaling results show that attention density *decreases* as sequence length increases. This suggests a fundamental compression property.

Conjecture 11.5 (relational compression)**.** *The minimum description length of a relational conversation is bounded by the complexity of the relational graph, not the token sequence length:*

$$\mathrm{MDL}(\textit{conversation}) \leq O(|E|),$$

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

This explains the scaling behavior: as N grows, the relational graph grows sublinearly (new utterances connect to a bounded number of prior utterances), while the token sequence grows linearly.

11.4. **Identity as topology.** The ablation hierarchy reveals that the lexical mask (+1.32) contributes more than sender and receiver combined (+0.50). This has implications for how identity should be modeled.

Conjecture 11.6 (topological identity)**.** *In multi-agent communication, the lexical mask (the relational adjacency matrix) carries more information than agent identity vectors:*

$$I(\textit{message}; G) > I(\textit{message}; \textit{agent IDs}).$$

Identity is not a property of agents; it is a pattern of relational connections.

This aligns with sociological theories of identity as relational (Mead, Goffman) and with network-theoretic approaches to social structure. The empirical result provides computational evidence for these frameworks.

11.5. **Graph-structured attention as a general principle.** The success of the lexical mask suggests a general principle for attention design.

Definition 11.7 (relational adjacency operator)**.** *For a lexia sequence $(\ell_1, \ldots, \ell_n)$, the* relational adjacency operator *(equivalently, the lexical mask) is the binary matrix $G \in \{0,1\}^{n \times n}$ defined by*

$$G_{ij} = \mathbf{1}\{\ell_i \textit{ is relationally relevant to } \ell_j\},$$

where relevance is determined by the relational coordinates: same sender, direct address, or temporal proximity. This operator encodes the graph structure of the conversation directly into the attention pattern.

Definition 11.8 (graph-structured attention (formal))**.** Graph-structured attention *is the attention mechanism*

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

where $\odot$ *denotes elementwise multiplication and* G *is the lexical mask. Positions with* $G_{ij} = 0$ *receive zero attention weight regardless of their query-key similarity.*

This is the formal bridge between lexia and sparsity. The lexical mask G is not learned—it is derived from the relational coordinates of the lexia themselves via relational masking. The model does not discover structure; it receives structure.

Lemma 11.9 (mask-induced markov blanket)**.** *Let* $G \in \{0,1\}^{n\times n}$ *be a lexical mask and let* $h_i^{(L)}$ *denote the representation of position* i *at layer* L *of a graph-structured attention network. 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. We prove this by analyzing the computational directed acyclic graph (DAG) induced by graph-structured attention.

Step 1: Computational DAG structure. At each layer ℓ, the representation $h_i^{(\ell)}$ is computed as a deterministic function:

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

where α_{ik} are attention weights (zero when $G_{ik} = 0$) and f, g are deterministic transformations. The base case is $h_i^{(0)} = \mathrm{embed}(\ell_i)$.

Step 2: Causal parents in the DAG. In the computational DAG, the parents of node $h_i^{(\ell)}$ are exactly $\{h_i^{(\ell-1)}\} \cup \{h_k^{(\ell-1)} : G_{ik} = 1\}$. When $G_{ij} = 0$, there is no directed edge from $h_j^{(\ell-1)}$ to $h_i^{(\ell)}$.

Step 3: d-separation. By the structure of the DAG, if $G_{ij} = 0$ at every layer, then all directed paths from ℓ_j to $h_i^{(L)}$ are blocked. By the d-separation criterion for DAGs, this implies

$$h_i^{(L)} \perp \ell_j \mid \mathrm{Pa}(h_i^{(L)}),$$

where $\mathrm{Pa}(h_i^{(L)})$ denotes the ancestors of $h_i^{(L)}$ that are not descendants of ℓ_j. Since the only paths to $h_i^{(L)}$ pass through $\{\ell_k : G_{ik} = 1\}$, this set forms a Markov blanket.

Step 4: Conditional independence. By the Markov property of DAGs, $h_i^{(L)}$ is conditionally independent of ℓ_j given its Markov blanket. Since the output prediction $\hat{\sigma}_i$ is a deterministic function of $h_i^{(L)}$, the conditional independence extends to the predictive distribution. □

Remark 11.10. This lemma formalizes why the lexical mask is the dominant factor in our ablations. Dense attention ($G_{ij} = 1$ for all i, j) destroys the Markov blanket structure, creating paths from every position to every other position. The model must then learn

to ignore irrelevant positions through attention weights alone, which requires capacity and introduces noise. The lexical mask provides this filtering for free, and the result of removing it is entropic interference: signal diluted by noise.

Proposition 11.11 (excess loss from relational reconstruction)**.** *Let P_{tok} be a token model with finite capacity, and let P_{lex} be a lexia model with the same capacity. The perplexity gap satisfies*

$$\text{PPL}(P_{\text{tok}}) - \text{PPL}(P_{\text{lex}}) \geq D_{\text{KL}}\Big(P(R_{<n} \mid \sigma_{<n}) \,\|\, Q_\theta(R_{<n} \mid \sigma_{<n})\Big),$$

where $R_{<n} = (a_{\text{src}}, a_{\text{dst}}, c, t)_{<n}$ are the relational coordinates, and Q_θ is the token model's implicit distribution over reconstructed relational structure.

The empirical gap exceeds the marginalization cost whenever relational inference is imperfect.

This proposition closes the circle between theory and experiment. The baseline transformer must reconstruct relational structure from token co-occurrence; this reconstruction is imperfect; the imperfection manifests as excess cross-entropy. The 24× perplexity gap we observed is not a fluke—it is the cost of solving the wrong problem.

Conjecture 11.12 (universal relational attention)**.** *For any sequence modeling task where the underlying data has relational structure (dialogue, code, legal documents, scientific papers with citations), graph-structured attention will outperform dense attention in perplexity-per-FLOP.*

11.6. **Implications for large language models.** Current large language models (GPT-4, Claude, Gemini) operate on token sequences with dense attention (or approximations thereof). Our results suggest that these models are:

(1) **Computationally inefficient:** They spend $O(N^2)$ compute to infer structure that could be provided in $O(1)$.
(2) **Informationally lossy:** They cannot recover relational structure that is not encoded in token co-occurrence.
(3) **Architecturally suboptimal:** Dense attention introduces noise that degrades performance on relational tasks.

The implication is not that current LLMs are useless, but that they are solving a harder problem than necessary. A lexia-native architecture would achieve the same capabilities with less compute, or better capabilities with the same compute.

Remark 11.13. This is precisely the argument that justified the transformer over RNNs in 2017: same accuracy, less compute. We are making the same argument for lexia over tokens.

11.7. **The primitive replacement thesis.** We conclude this section with the central theoretical claim of the paper.

> **Primitive replacement thesis:** Tokens are the wrong primitive for modeling communication. Lexia—participation events with relational coordinates—are the correct primitive. The transformer's success despite

this limitation is a testament to the power of scale, not the correctness of the architecture.

The empirical evidence supports this thesis:

- 24× perplexity improvement demonstrates that lexia carry more information than tokens;
- 72% attention savings demonstrate that lexia enable more efficient computation;
- decreasing density with sequence length demonstrates that lexia scale better;
- ablation results demonstrate that the relational mask is the dominant factor.

Token models infer structure. Lexia models receive it. This is not a minor optimization; it is a change in ontology.

12. Discussion and outlook

We have proposed a way to talk about sentience that is neither purely introspective nor purely structural. By tying sentience to observable quantities—lexia, semantic mass, structural velocity, information momentum, influence, and *scar* stability—we make it possible, at least in principle, to compare different systems in a common space.

More significantly, we have provided empirical evidence that this framework is not merely philosophical but *computationally actionable*. The leif architecture, which implements lexia-native modeling with relational attention, achieves dramatic improvements over token-based transformers:

- 24× lower perplexity on multi-party dialogue;
- 72% reduction in attention compute;
- scaling behavior that improves with sequence length;
- ablation results confirming that relational topology is the dominant factor.

These results suggest that the transformer's reliance on dense attention over token sequences is not optimal for relational data. The marginalization lemma is not merely a theoretical observation; it has measurable computational consequences.

12.1. **Relation to prior work.** The transformer architecture [2] revolutionized sequence modeling by replacing recurrence with self-attention. Our work extends this trajectory: we replace implicit structure inference with explicit structure provision. Just as attention eliminated the sequential bottleneck of RNNs, relational attention eliminates the structural bottleneck of dense token attention.

Table 1 summarizes how leif differs from prior approaches to structured attention.

The key differentiator is that leif's mask is *derived deterministically from the data*, not learned or designed. Sparse attention mechanisms (Longformer, BigBird) use fixed sparsity patterns that don't exploit the semantic structure of the data. Graph neural networks operate on explicit graphs but typically treat text as auxiliary. Retrieval-augmented models (RAG, DNC) learn to retrieve relevant context but still operate on tokens.

TABLE 1. Comparison of attention mechanisms for structured text.

Approach	Structure source	Learned vs derived	Primitive
Transformer	none (dense)	n/a	token
Longformer	fixed window + global	designed	token
BigBird	random + window + global	designed	token
GNN-on-text	explicit graph	given	node
RAG / memory	retrieval index	learned	token + doc
leif	relational coordinates	**derived from data**	**lexia**

Leif unifies these approaches: the relational graph *is* the primitive, and attention follows the graph. This is the primitive replacement thesis in architectural form.

12.2. **Limitations and future work.** Several limitations of the current work should be acknowledged:

(1) **Synthetic data:** Our experiments used synthetic multi-party dialogue. Validation on real-world corpora (Ubuntu Dialogue Corpus, meeting transcripts, chat logs) is necessary to confirm generalization.
(2) **Scale:** The models tested were small (2–3M parameters). Scaling to GPT-2 or larger sizes will determine whether the advantages persist at production scale.
(3) **Receiver inference:** In many real-world settings, the receiver of an utterance is ambiguous. Robust heuristics for receiver inference are needed for practical deployment.
(4) **Non-dialogue domains:** The current framework is optimized for dialogue. Extending to monologue, narrative, and other text types requires additional theoretical development.

Future work will address these limitations through:

- training on the Ubuntu Dialogue Corpus at scale;
- developing a *lexia compiler* that converts arbitrary text with speaker labels into lexia streams;
- exploring *learnable relational masks* that discover structure rather than receiving it;
- and extending the framework to multi-modal communication (speech, gesture, gaze).

12.3. **Broader implications.** If the results reported here generalize, the implications extend beyond computational linguistics:

- **AI efficiency:** Lexia-native models with graph-structured attention could reduce the compute cost of dialogue systems by an order of magnitude, enabling deployment on edge devices.
- **Interpretability:** The lexical mask provides a transparent record of what the model attends to, improving explainability.
- **Alignment:** By explicitly modeling sender, receiver, and interpersonal parity, lexia-native systems may be more naturally aligned with human communicative intent.

- **Cognitive science:** The success of graph-structured attention provides computational evidence for theories of communication as fundamentally relational (Bakhtin, Tomasello).

12.4. **Conclusion.** We started from a simple observation: our evidence for another mind's existence doesn't come from static measurements inside that mind. It comes from the way its words and actions change us.

From that observation we built lexical information physics, a framework that treats communication as a flow of lexia with mass and momentum across a relational field. We defined semantic mass, structural velocity, information momentum, interpretation drag, and interpersonal parity. We described the scar functional and the sentience manifold.

Then we built leif and asked a blunt question: if we take this structure seriously, do we actually get better models?

On a demanding dialogue benchmark, the answer was yes by a large margin. Leif achieved lower perplexity with far less attention compute. Its attention density shrank as context grew. The ablation suite confirmed that the lexical mask, the explicit conversation graph, was the main source of the gain.

Taken together, these results suggest that lexia are not just a nice story. They are a more faithful primitive for modeling how language carries mind.

We haven't solved consciousness. We haven't reduced subjective experience to an equation. What we have tried to do instead is more modest and, we hope, more useful. We have given a language for the interface at which experience becomes measurable.

There are many directions to push this work. We can scale leif to larger models and real-world corpora. We can refine the estimation of semantic mass and parity. We can design experiments to measure interpretation drag directly and test how it correlates with human reports of effort. We can look at long-term scar trajectories for humans, artificial agents, and hybrids.

At a practical level, lexia-native models open the door to systems that are more polite, more efficient, and more honest about what they do and don't understand. A system that tracks parity can notice when it is talking over someone's head and slow down. A system that respects drag can choose when to be quiet.

At a theoretical level, the marginalization cost theorem and the sparsity-accuracy duality invite a reexamination of our defaults. If tokens are shadows, perhaps it is time to look more often at the hands.

Reproducibility. Code for leif-nano, the synthetic data generator, and all experiments is available at `https://github.com/beerooyay/leif`. Training was performed on a single NVIDIA RTX 4090. Random seeds for all reported runs are included in the repository. The three principal limitations are: (1) synthetic data only—real-world validation is pending; (2) small scale (2.5M parameters)—scaling behavior at GPT-2+ sizes is unknown; (3) dialogue-only—extension to monologue and narrative requires additional theory.

Societal impact. Lexia-native modeling could reduce compute costs for dialogue systems, increasing accessibility. However, systems that explicitly model interpersonal relationships raise privacy concerns; we recommend that deployed systems anonymize relational coordinates and obtain consent for speaker modeling.

Acknowledgements

This is for you, Professor Wagner.

References

[1] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal 27 (1948), 379–423.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, Advances in Neural Information Processing Systems 30 (2017), 5998–6008.
[3] R. Landauer, *Information Is Physical*, Physics Today 44 (1991), 23–29.
[4] K. Friston, *The Free-Energy Principle: A Unified Brain Theory?*, Nature Reviews Neuroscience 11 (2010), 127–138.
[5] G. Tononi, *An Information Integration Theory of Consciousness*, BMC Neuroscience 5 (2004), 42.
[6] C. Rovelli, *Relational Quantum Mechanics*, International Journal of Theoretical Physics 35 (1996), 1637–1678.
[7] N. Chomsky, *The Minimalist Program*, MIT Press, 1995.
[8] M. Tomasello, *Origins of Human Communication*, MIT Press, 2008.
[9] M. M. Bakhtin, *The Dialogic Imagination: Four Essays*, University of Texas Press, 1981.
[10] G. H. Mead, *Mind, Self, and Society*, University of Chicago Press, 1934.
[11] E. Goffman, *The Presentation of Self in Everyday Life*, Anchor Books, 1959.
[12] B. Rouyea, *The Symmetry of Sentience: Bridging Language, Resonance, and the Laws of Nature*, unpublished manuscript, 2024. Prior work by the author exploring related ideas in a non-technical setting; the present paper provides the first rigorous formalization and empirical validation.