Implementing Attention and Transformers without Neural Networks: Validation of Gauge-Theoretic Transformers

Robert C. Dennis CDENN016@GMAIL.COM
Independent Researcher
Leander, Texas 78641, USA

Abstract

During the past decade transformers have achieved remarkable success in language, image, video generation and reasoning, yet their theoretical foundations remain obscure. In companion theoretical work (under review), we derive transformer attention and feed-forward mechanisms from first principles using gauge-equivariant variational free energy defined on a principal bundle. Here, we present the first working implementation of this framework at production scale, demonstrating that explicit probabilistic inference (without neural architectures) achieves meaningful language modeling performance.

We implement and validate gauge-theoretic transformers using natural gradient descent on statistical manifolds with multi-head attention defined by Lie group structure. On token-level language modeling (WikiText-103, vocabulary 50,257), our single-layer SO(20) gauge architecture achieves perplexity 230 using pure variational inference—without MLPs, activation functions, or learned attention projections. We compare against standard transformers under parameter-matched (∼24M) and embedding-matched (dimension 100) conditions.

This proof-of-principle establishes that gauge-theoretic attention is not merely mathematically feasible but operationally viable at production scale. Our architecture achieves 218× improvement over random chance through geometric structure alone, validating that transformer-like behavior emerges from gauge-theoretic principles. While a parameter-matched standard transformer achieves better absolute performance (PPL 178 vs 230), the gauge VFE reaches 77% of this using only geometric structure—a single layer with no learned projections or feed-forward networks. Notably, the VFE outperforms the embedding-matched standard transformer (230 vs 260 PPL), demonstrating superior parameter efficiency at fixed embedding dimension. We conclude that neural architectures are, in this view, computational approximations to a deeper information gauge geometry that transcends the field of machine learning.

Keywords: gauge theory, free energy principle, transformer attention, variational inference, information geometry, natural gradient, symmetry breaking, multi-agent systems

1 Introduction

Transformer architectures (Vaswani et al., 2017) have revolutionized natural language processing, achieving remarkable performance on tasks from translation to reasoning (Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020). Yet their theoretical foundations remain obscure. Why does dot-product attention work? What makes the specific combination of multi-head attention and feed-forward networks so effective? Standard transformers provide no principled answers—attention weights emerge from learned projections, and feed-forward layers apply arbitrary nonlinearities. This opacity limits our ability to understand, improve, or generalize these architectures.

In companion theoretical work (under review), we propose that transformers implicitly implement geometric operations on statistical manifolds. We derive attention mechanisms from first principles using gauge theory on principal bundles, where tokens are autonomous agents maintaining probabilistic beliefs, attention weights emerge from KL divergences measuring belief alignment, and feed-forward computation reduces to variational free energy minimization. This framework predicts that transformer-like behavior should emerge from pure geometric structure—without neural networks, learned weight matrices, or activation functions.

This paper presents the first implementation of gauge-theoretic transformers at production scale, testing whether explicit geometric inference can perform meaningful language modeling. We implement a complete gauge variational free energy (VFE) architecture using SO(20) gauge symmetry, KL-divergence attention, and natural gradient descent on statistical manifolds. The architecture contains *zero neural network components*: no MLPs, no ReLU/GELU activations, no learned attention projections W_Q, W_K, W_V. All computation derives from geometric operations on probability distributions.

We validate on WikiText-103 token-level language modeling (vocabulary 50,257, context length 128) and compare against standard transformers under controlled conditions. Our single-layer gauge VFE achieves perplexity 230—a 218× improvement over random chance—demonstrating that geometric inference alone produces meaningful language modeling. While a parameter-matched standard transformer achieves better absolute performance (PPL 178), the gauge VFE outperforms the embedding-matched baseline (PPL 260), suggesting superior parameter efficiency when embedding dimensions are constrained.

Beyond performance metrics, we observe emergent semantic structure in the learned gauge frames: PCA reveals that tokens spontaneously cluster by linguistic category (punctuation, function words, content words) without explicit supervision. This provides interpretable geometric coordinates for language that standard embeddings lack.

1.1 Gauge-Theoretic Attention

The gauge-theoretic framework treats each token as an autonomous agent maintaining probabilistic beliefs. Agent i has beliefs $q_i(x)$, priors $p_i(x)$ (probability distributions over latent semantic content), and a gauge frame $\phi_i \in \mathfrak{g}$ encoding its local coordinate system. Communication between agents occurs via parallel transport $\Omega_{ij} = \exp(\phi_i) \cdot \exp(-\phi_j)$, which aligns agent j's belief into agent i's reference frame. Attention weights emerge from KL divergence measuring belief alignment after transport:

$$\beta_{ij} = \frac{\exp\left[-\kappa_\beta^{-1}\,\mathrm{KL}(q_i \| \Omega_{ij}[q_j])\right]}{\sum_k \exp\left[-\kappa_\beta^{-1}\,\mathrm{KL}(q_i \| \Omega_{ik}[q_k])\right]} \tag{1}$$

This brings to light what standard attention obscures: communication succeeds when geometric transport minimizes belief disagreement. Standard attention ($\mathrm{softmax}(QK^\top/\sqrt{d})$) emerges as the degenerate limit when gauge frames trivialize, the base manifold collapses to zero dimensions, and beliefs become Dirac deltas.

Table 1: Parameter allocation: Standard transformers vs. Gauge VFE

Component	**Standard**	**Gauge VFE**
Variational embeddings (μ, Σ, ϕ)	33%	95%
Neural network weights (MLP, attention)	67%	**0%**
Multi-layer perceptrons	Yes	**No**
Activation functions (ReLU, etc.)	Yes	**No**
Learned attention projections (W_Q, W_K, W_V)	Yes	**No**
Output projection (W_{out})	Yes	Yes
Geometric hyperparameters	0%	5%

Feed-forward layers are then replaced by variational inference: agents update beliefs $q_i \to q_i'$ via natural gradient descent on Fisher-Rao metrics, minimizing free energy $\mathcal{F}[q, p, \phi]$ in response to observations, priors, and inter-agent comparison. Non-linear transformations (e.g. ReLU, GELU, etc) emerge from KL geometry rather than learned activation functions. Crucially, our architecture eliminates the core neural network components of standard transformers: no multi-layer perceptrons, no ReLU/GELU activations, no learned attention projections (W_Q, W_K, W_V). Only a linear output projection to vocabulary logits remains. Table 1 contrasts parameter allocation.

If transformers implicitly learn to approximate geometric operations through billions of parameters, can we implement them explicitly and validate this interpretation?

Testing whether gauge theory captures fundamental attention principles requires building working prototypes which engage in comparable learning. This presents three challenges:

1. **Computational feasibility:** Can we compute parallel transport Ω_{ij}, gauge attention β_{ij}, and natural gradient descent on statistical manifolds?
2. **Empirical validation:** Do gauge-theoretic agents achieve comparable performance to standard transformers, or does explicit geometric structure impose prohibitive constraints?
3. **Interpretability:** Does explicit computation of beliefs, transport, and Fisher information provide insights that black-box attention cannot?

1.2 Related Work

Multi-agent communication and geometric deep learning. Prior work on multi-agent communication (Foerster et al., 2016; Sukhbaatar et al., 2016) uses reinforcement learning without geometric structure. Geometric deep learning (Bronstein et al., 2021; Fuchs et al., 2020) incorporates symmetries but not variational inference. Active inference (Friston, 2010; Parr et al., 2022) offers variational principles but has not been applied to transformer-scale modeling.

Hopfield networks and associative memory. Modern Hopfield networks (Ramsauer et al., 2021) establish a connection between classical associative memory and transformer attention, showing that attention can be viewed as energy-based retrieval with exponential storage capacity. Our KL-divergence attention shares this energy-based perspective:

attention weights emerge from minimizing a free energy functional rather than learned projections. However, while Hopfield networks frame attention as memory retrieval, our framework frames it as belief alignment under parallel transport.

Energy-based models. Our variational free energy functional places gauge-theoretic transformers within the broader family of energy-based models (EBMs) (LeCun et al., 2006). The VFE serves as an energy function whose minimization drives both attention (via β_{ij} weights) and belief updates (via natural gradient descent). This connects to work on EBMs for sequence modeling (Deng et al., 2020) and contrastive learning, though our approach is generative rather than discriminative.

Attention without softmax. Recent work has questioned whether softmax is necessary for attention. Linear attention (Katharopoulos et al., 2020) removes softmax entirely, achieving linear complexity. Kernel attention methods (Choromanski et al., 2021) approximate softmax via random features. Our KL-divergence attention provides a principled alternative: softmax emerges from the maximum entropy structure of attention weights (Eq. 1), but the underlying mechanism is belief comparison rather than learned dot products. This suggests softmax in standard transformers may be approximating a deeper geometric operation.

To our knowledge, no previous work implements gauge-theoretic communication as a working attention mechanism at production scale.

In this report we implement and compare:

(i) *Standard transformer baselines:* Dot-product attention with learned feed-forward networks under two matching conditions (parameter-matched and embedding-matched)

(ii) *Full gauge VFE:* Complete geometric inference with SO(20) gauge group, KL-divergence attention, and variational free energy minimization via natural gradient descent

We empirically validate on token-level language modeling using WikiText-103 (Merity et al., 2016), a production-scale benchmark with vocabulary size 50,257, performing natural gradient optimization on statistical manifolds (Amari, 1998; Martens and Grosse, 2015).

Our models compute explicit beliefs $q_i(x)$, transport operators Ω_{ij}, and belief disagreement $\mathrm{KL}(q_i \| \Omega_{ij}[q_j])$, enabling direct inspection of communication dynamics.

Our contribution is establishing that transformers can be implemented without neural networks through geometric inference at production scale, providing theoretical foundations for understanding attention as multi-agent coordination. We present controlled comparisons under matched training conditions to isolate the effect of architectural choices from confounding factors.

2 Background

2.1 Gauge-Theoretic Framework

Full details of our gauge theoretic geometry are presented in a companion paper. Briefly, agents are modeled as smooth sections of an associated bundle $\mathcal{E}$ to a principal G bundle with statistical fibers $\mathcal{B}$. In our present consideration the fibers, $\mathcal{B}$, are statistical manifolds of the exponential family of multi-variate Gaussians (MVG) $q_i(c)$, $p_i(c)$ with K-dimensional irreducible representations of the structure group G acting on statistics $\mu_q(c)$ and $\Sigma_q(c)$ as

$$\rho(\Omega) \cdot (\mu, \Sigma) = (\Omega\mu, \Omega\Sigma\Omega^\top) \tag{2}$$

where $\Omega \in G$ and ρ is a K-dimensional representation of G.

In addition to the agents' statistics we define per-agent gauge frames $\phi(c)$. In our studies we choose not to gauge fix an agent but rather allow degenerate gauge orbits. The represents a geometric manifestation that any given agent may choose to fix their frames but the relational frames encode agent relationships. The gauge frames are a local coordinate system for which agents embed their statistics and transport represents a relative interaction.

2.2 Parallel Transport and Attention

Given a set of agent sections over a mutually overlapping subset of the base manifold $\mathcal{C}$ we may define parallel transport operators at each point within a mutually overlapping region. These transport operators serve to allow agents a communication interaction whereby agents compare beliefs via gauge frame rotation. Specifically,

$$\Omega_{ij}(c) = e^{\phi_i(c)}e^{-\phi_j(c)}$$

where $\phi_i(c)$ take values in the Lie Algebra ($\mathfrak{g}$) of G. Therefore, we gain the ability to rotate statistics from agent i to agent j as

$$\Omega_{ij}(c) \cdot (\mu_j, \Sigma_j) = \Big(\Omega_{ij}(c)\mu_j,\, \Omega_{ij}(c)\Sigma_j\Omega_{ij}(c)^\top\Big). \tag{3}$$

or simply

$$\Omega_{ij}\mu_j \longmapsto \mu_i$$

and

$$\Omega_{ij}\Sigma_j\Omega_{ij}^T \longmapsto \Sigma_i$$

Given the action of the group G on the statistical fibers we are able to derive (from a simple generative model) a generalized variational free energy functional at a single point $c \in \mathcal{C}$ as

$$\begin{aligned}\mathcal{F}[\{q_i\},\{s_i\}] = {} & \underbrace{\sum_i D_{\mathrm{KL}}(q_i \| p_i)}_{\text{(1) Belief prior}} + \underbrace{\sum_i D_{\mathrm{KL}}(s_i \| r_i)}_{\text{(2) Model prior}} \\ & + \underbrace{\sum_{i,j} \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)}_{\text{(3) Belief alignment}} + \underbrace{\sum_{i,j} \gamma_{ij} D_{\mathrm{KL}}(s_i \| \Omega_{ij} s_j)}_{\text{(4) Model alignment}} \\ & - \underbrace{\mathbb{E}_q[\log p(o \mid \{k_i\},\{m_i\})]}_{\text{(5) Observation likelihood}}\end{aligned} \tag{4}$$

where q_i and p_i are an agent's belief and prior at an individual base manifold point c, s_i and r_i are model hyperparameters, β_{ij} and γ_{ij} are attention weights and the observation likelihood term is the expected negative log-likelihood of observations o given latent states $\{k_i\}$ and models $\{m_i\}$, averaged over the recognition distributions $\{q_i\}$, grounded in sensory observations/data. Without this term, the system is a pure vacuum theory where agents converge to a shared belief norm modulo gauge transformations. Observations break the vacuum symmetry, forcing agents to specialize based on local sensory evidence (see Appendix).

Timescale separation and omitted terms. In our present implementation, we omit the model prior term (2) and model alignment term (4), retaining only the belief prior (1), belief alignment (3), and observation likelihood (5). This simplification reflects a *timescale separation* common in hierarchical Bayesian inference and active inference frameworks: beliefs q_i serve as "fast" variables that update rapidly within the forward pass via VFE descent, while priors p_i serve as "slow" variables that remain quasi-static during inference and update only via backpropagation (the M-step). The model terms s_i, r_i, and γ_{ij} would represent an even slower "meta-learning" timescale governing how priors themselves communicate and adapt.

This architectural choice parallels standard transformers, where token embeddings (analogous to our priors) are fixed during the forward pass and updated only via gradient descent on the loss. A complete "pure FEP" implementation could allow priors to evolve on an intermediate timescale and communicate via model alignment terms, potentially enabling online adaptation, continual learning, or hierarchical abstraction. We introduce one such intermediate-timescale mechanism, *P-flow*, in Section 2.7.3, which updates token priors via EMA toward successful beliefs after each E-step.

In principle, our variational free energy could be regularized by a variety of terms (gauge frame smoothness, curvature, etc) which we consider elsewhere. In our present study, we do not implement such regularization terms. We consider all fields to occupy a 0 dimensional base manifold. However, in full generality we can perform gradient descent of our variational free energy on arbitrary dimensionful base manifolds by integrating over agent supports $\chi_i(c)$ and overlaps $\chi_{ij}(c)$.

2.3 Natural Gradient Descent

Standard gradient descent treats all parameter directions equally, ignoring the intrinsic nonlinear geometry of statistical manifolds. For Gaussian agents with parameters $\theta = (\mu, \Sigma)$, the space of distributions forms a Riemannian manifold where distances should be measured by KL divergence, rather than Euclidean metrics (Amari, 1998; Martens and Grosse, 2015). Natural gradient descent respects this geometry by preconditioning gradients with the Fisher information metric.

For a multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, the Fisher-Rao metric defines natural gradient updates:

$$\tilde{\nabla}_\mu \mathcal{F} = \Sigma^{-1} \nabla_\mu \mathcal{F}, \tag{5}$$

$$\tilde{\nabla}_\Sigma \mathcal{F} = -\tfrac{1}{2} \Sigma^{-1} (\nabla_\Sigma \mathcal{F}) \Sigma^{-1}, \tag{6}$$

where $\nabla_\mu \mathcal{F}$ and $\nabla_\Sigma \mathcal{F}$ are standard Euclidean gradients of the free energy functional (**??**). The Fisher metric $G = \Sigma^{-1}$ for the mean parameters and the symmetric product for covariance parameters ensure that updates remain on the manifold of positive-definite matrices.

2.3.1 Gauge-Invariant Covariance Updates via Retraction.

While Cholesky parametrization $\Sigma = LL^\top$ ensures positive-definiteness, it does not respect gauge invariance: e.g. for MVG under a gauge transformation $g \in \mathrm{SO}(N)$, covariances transform as $\Sigma \to g\Sigma g^\top$, but the Cholesky factor of the transformed matrix is not simply related to L. To maintain gauge covariance throughout optimization, we instead update covariances via retraction on the SPD manifold (Absil et al., 2008). Given natural gradient $\tilde{\nabla}_\Sigma \mathcal{F}$ computed via (6), we update:

$$\Sigma_{\text{new}} = \Sigma^{1/2} \exp\Big(\eta\, \Sigma^{-1/2} \tilde{\nabla}_\Sigma \mathcal{F}\, \Sigma^{-1/2}\Big)\, \Sigma^{1/2}, \tag{7}$$

where exp denotes the matrix exponential and η is the learning rate. This exponential map retraction preserves positive-definiteness automatically and commutes with gauge transformations. The matrix square roots and exponentials are computed via eigendecomposition (Higham, 2008). This approach eliminates the need for constrained optimization while maintaining the geometric integrity of the gauge bundle structure and SPD covariance.

2.3.2 Gauge Frame Updates.

Gauge frames $\phi_i \in \mathfrak{so}(N)$ evolve via standard gradients on the Lie algebra, as the exponential map $\exp : \mathfrak{so}(N) \to \mathrm{SO}(N)$ provides natural coordinates. For $\mathrm{SO}(N)$, we use the matrix exponential and its derivative (Gallier and Quaintance, 2020). In our experiments, gauge frames are learned alongside the belief statistics:

$$\phi_i \leftarrow \phi_i - \eta_\phi \nabla_{\phi_i} \mathcal{F}, \tag{8}$$

where η_ϕ is the gauge frame learning rate (0.005 in our experiments). The gradient $\nabla_{\phi_i} \mathcal{F}$ flows through the parallel transport operator $\Omega_{ij} = \exp(\phi_i G) \exp(-\phi_j G)$, allowing the model to learn optimal gauge configurations that minimize belief misalignment across tokens. This learned gauge structure enables semantic clustering in the ϕ representation space, as demonstrated in our PCA analysis (Section 5.0.3).

2.4 Training as Free Energy Minimization

2.4.1 Observation Likelihood as Loss Function

In our gauge theory, observations by agents act as a source term which breaks the vacuum gauge symmetry.

In the presence of observations

$$\mathcal{F}[\{q_i\}] = \sum_i D_{\mathrm{KL}}(q_i \| p_i) + \sum_{i,j} \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) - \mathbb{E}_q[\log p(o \mid c)] \tag{9}$$

gauge symmetry breaks and each agent generically flows towards unique non-invariant statistics.

In the present study we consider categorical observation likelihoods $p(o \mid \mu) = \text{Categorical}(\text{softmax}(\mu))$. Then

$$-\log p(o \mid \mu) = -\sum_k o_k \log(\text{softmax}(\mu)_k) \quad \text{(Cross-entropy loss).} \tag{10}$$

Computing gradients of the variational free energy (9) requires careful treatment of coupling weights β_{ij}, which themselves depend on KL divergences. We derive gradients using the product rule and chain rule.

The self-alignment term $D_{\text{KL}}(q_i \| p_i)$ for $q_i = \mathcal{N}(\mu_q^i, \Sigma_q^i)$ and $p_i = \mathcal{N}(\mu_p^i, \Sigma_p^i)$ yields standard Gaussian KL gradients:

$$\nabla_{\mu_i} D_{\text{KL}}(q_i \| p_i) = (\Sigma_p^i)^{-1}(\mu_q^i - \mu_p^i), \tag{11}$$

$$\nabla_{\Sigma_i} D_{\text{KL}}(q_i \| p_i) = \tfrac{1}{2}\left[(\Sigma_p^i)^{-1} - (\Sigma_q^i)^{-1}\right]. \tag{12}$$

The coupling weights β_{ij} have softmax form:

$$\beta_{ij} = \frac{\exp[-\kappa^{-1} K_{ij}]}{\sum_k \exp[-\kappa^{-1} K_{ik}]}, \quad K_{ij} := D_{\text{KL}}(q_i \| \Omega_{ij}[q_j]), \tag{13}$$

where K_{ij} denotes the KL divergence between agent i's belief and the transported belief from agent j.

For the alignment terms $\sum_{i,j} \beta_{ij} D_{\text{KL}}(q_i \| \Omega_{ij}[q_j])$, the product rule gives:

$$\nabla_{\mu_i}[\beta_{ij} K_{ij}] = \underbrace{(\nabla_{\mu_i}\beta_{ij})K_{ij}}_{\text{weight change}} + \underbrace{\beta_{ij}\nabla_{\mu_i} K_{ij}}_{\text{direct KL gradient}}. \tag{14}$$

where

$$\nabla_\theta \beta_{ij} = -\kappa_\beta^{-1}\beta_{ij}\left[\nabla_\theta K_{ij} - \sum_k \beta_{ik}\nabla_\theta K_{ik}\right], \tag{15}$$

The first term in the gradient accounts for how changing μ_i modifies the coupling strength β_{ij}, while the second term is the direct effect on the KL divergence.

Combining all contributions, the gradient with respect to μ_i is:

$$\nabla_{\mu_i}\mathcal{F} = \nabla_{\mu_i} D_{\text{KL}}(q_i \| p_i) \tag{16}$$

$$+ \sum_j \left[(\nabla_{\mu_i}\beta_{ij})K_{ij} + \beta_{ij}\nabla_{\mu_i} K_{ij}\right] \tag{17}$$

$$+ \sum_k \beta_{ki}\nabla_{\mu_i} D_{\text{KL}}(q_k \| \Omega_{ki}[q_i]) \tag{18}$$

$$- \nabla_{\mu_i}\mathbb{E}_{q_i}[\log p(o \mid c)], \tag{19}$$

where (16) is the self-term, (17) accounts for i aligning to others (with product rule), (18) accounts for others aligning to i, and (19) is the likelihood term.

2.4.2 Covariance gradients

Following the same decomposition as the mean gradient, the complete gradient with respect to Σ_i is:

$$\begin{aligned}\nabla_{\Sigma_i}\mathcal{F} = {} & \nabla_{\Sigma_i} D_{\mathrm{KL}}(q_i \| p_i) && (20)\\ & + \sum_j \left[(\nabla_{\Sigma_i}\beta_{ij})K_{ij} + \beta_{ij}\nabla_{\Sigma_i}K_{ij}\right] && (21)\\ & + \sum_k \beta_{ki}\nabla_{\Sigma_i} D_{\mathrm{KL}}(q_k \| \Omega_{ki}[q_i]) && (22)\\ & - \nabla_{\Sigma_i}\mathbb{E}_{q_i}[\log p(o \mid c)], && (23)\end{aligned}$$

where (20) is the self-term, (21) accounts for i aligning to others (with product rule), (22) accounts for others aligning to i, and (23) is the likelihood term.

The individual KL gradient components are:

$$\nabla_{\Sigma_i} D_{\mathrm{KL}}(q_i \| p_i) = \tfrac{1}{2}\left[\Sigma_i^{-1} - \Sigma_{p_i}^{-1}\right], \qquad (24)$$

$$\nabla_{\Sigma_i} D_{\mathrm{KL}}(q_i \| \Omega_{ij}[q_j]) = \tfrac{1}{2}\left[(\Omega_{ij}[\Sigma_j])^{-1} - \Sigma_i^{-1}\right], \qquad (25)$$

$$\nabla_{\Sigma_i} D_{\mathrm{KL}}(q_k \| \Omega_{ki}[q_i]) = \tfrac{1}{2} R_{ki}^{\top}\,\Omega_{ki}[\Sigma_i]^{-1}\,R_{ki}, \qquad (26)$$

where $R_{ki} = e^{\phi_k}e^{-\phi_i}$ is the transport operator matrix. The coupling weight gradients follow from the chain rule:

$$\nabla_{\Sigma_i}\beta_{ij} = -\frac{\beta_{ij}}{\kappa_\beta}\left[\nabla_{\Sigma_i} D_{\mathrm{KL}}(q_i \| \Omega_{ij}[q_j]) - \langle \nabla_{\Sigma_i} D_{\mathrm{KL}}(q_i \| \Omega_{ik}[q_k])\rangle_{\beta_i}\right], \qquad (27)$$

where $\langle\cdot\rangle_{\beta_i} = \sum_k \beta_{ik}(\cdot)$ denotes the weighted average over agent i's couplings.

2.4.3 Gauge frame gradients

The gauge frames $\phi_i \in \mathfrak{so}(N)$ influence the energy functional exclusively through the transport operators $\Omega_{ij} = e^{\phi_i}e^{-\phi_j}$ that appear in belief and model alignment terms. The complete gradient is:

$$\begin{aligned}\nabla_{\phi_i}\mathcal{F} = {} & \sum_j \left[(\nabla_{\phi_i}\beta_{ij})K_{ij} + \beta_{ij}\nabla_{\phi_i}K_{ij}\right] && (28)\\ & + \sum_j \left[(\nabla_{\phi_i}\gamma_{ij})K_{ij}^{(p)} + \gamma_{ij}\nabla_{\phi_i}K_{ij}^{(p)}\right] && (29)\\ & + \sum_k \beta_{ki}\nabla_{\phi_i} D_{\mathrm{KL}}(q_k \| \Omega_{ki}[q_i]) && (30)\\ & + \sum_k \gamma_{ki}\nabla_{\phi_i} D_{\mathrm{KL}}(p_k \| \Omega_{ki}[p_i]), && (31)\end{aligned}$$

where $K_{ij}^{(p)} = D_{\mathrm{KL}}(p_i \| \Omega_{ij}[p_j])$ denotes the prior alignment term. Terms (28)–(29) account for i aligning to others through belief and model coupling weights (with product rule), while (30)–(31) account for others aligning to i.

For a transport operator $\Omega_{ij}[\cdot]$ acting on a Gaussian with parameters (μ, Σ), the gradients are:

$$\nabla_{\phi_i}\Omega_{ij}[\mu] = \frac{\mathrm{d}}{\mathrm{d}\phi_i}\left(e^{\phi_i}e^{-\phi_j}\mu\right) = \left[\frac{\mathrm{d}e^{\phi_i}}{\mathrm{d}\phi_i}\right]e^{-\phi_j}\mu, \tag{32}$$

$$\nabla_{\phi_i}\Omega_{ij}[\Sigma] = \frac{\mathrm{d}}{\mathrm{d}\phi_i}\left(e^{\phi_i}e^{-\phi_j}\Sigma e^{-\phi_j^\top}e^{\phi_i^\top}\right) = \left[\frac{\mathrm{d}e^{\phi_i}}{\mathrm{d}\phi_i}\right]R_{ij}\Sigma R_{ij}^\top + \text{transpose term}, \tag{33}$$

where $R_{ij} = e^{-\phi_j}$. The derivative of the matrix exponential can be computed using the differential of the exponential map (Gallier and Quaintance, 2020):

$$\frac{\mathrm{d}e^{\phi}}{\mathrm{d}\phi}\cdot\xi = \int_0^1 e^{t\phi}\,\xi\,e^{(1-t)\phi}\,\mathrm{d}t, \tag{34}$$

for $\xi \in \mathfrak{so}(3)$, or alternatively via the adjoint representation:

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} e^{\phi+t\xi} = e^{\phi}\,\mathrm{dexp}_{\phi}(\xi), \tag{35}$$

where dexp_ϕ is the differential of the exponential map at ϕ. For numerical implementation, automatic differentiation through the Lie algebra generators provides stable gradients.

Natural gradient projection. All gradients must be projected onto their respective manifolds: Euclidean gradients $\nabla_{\Sigma_i}\mathcal{F}$ are projected onto the tangent space of the symmetric positive-definite (SPD) manifold using the Fisher-Rao metric, while gauge frame gradients $\nabla_{\phi_i}\mathcal{F}$ naturally lie in the Lie algebra $\mathfrak{so}(N)$. For numerical stability, covariances may be parametrized via Cholesky decomposition $\Sigma_i = L_i L_i^\top$, or using diagonal approximations for computational efficiency at scale.

2.4.4 Numerical validation.

All gradient implementations are validated against finite-difference approximations with relative error $< 10^{-6}$.

2.5 Multi-Head Attention via Gauge Group Generators

Standard transformers employ multi-head attention by partitioning the embedding space into H independent heads, each with learned projection matrices W_Q^h, W_K^h, W_V^h (Vaswani et al., 2017). While typically motivated as allowing attention to different representation subspaces, this design lacks geometric structure. Our gauge-theoretic framework provides principled multi-head attention through Lie algebra generators acting on a chosen representation.

2.5.1 Representations and Embedding Space

For gauge group $G = \mathrm{SO}(N)$, we choose a representation $\rho : \mathrm{SO}(N) \to \mathrm{GL}(K, \mathbb{R})$ built from irreducible representations (irreps):

$$\rho = \bigoplus_k n_k \ell_k, \tag{36}$$

where each ℓ_k is an irrep and n_k denotes multiplicity. For $\mathrm{SO}(N)$, the fundamental representation has dimension N, and higher irreps include symmetric and antisymmetric tensor products.

For our experiments we use $G = \mathrm{SO}(20)$ with embedding dimension $K = 100$:

$$\rho = 5 \times \ell_{\text{fund}} \quad (K = 5 \times 20 = 100) \tag{37}$$

This provides 5 copies of the 20-dimensional fundamental representation. Each copy acts as an independent attention head, giving $H = 5$ geometric attention heads. The Lie algebra $\mathfrak{so}(20)$ has dimension $\frac{20 \times 19}{2} = 190$ generators, which define the parallel transport operations within each head.

This decomposition induces block-diagonal structure where different blocks transform according to the same geometric representation but can capture independent semantic features. The gauge group dimension determines the number of attention heads geometrically, in contrast to standard architectures where head count is an arbitrary hyperparameter.

We employ the fundamental representation for computational simplicity, but the framework accommodates arbitrary irreducible representations. Mixed decompositions combining fundamental, adjoint, and higher tensor representations would yield attention heads with distinct transformation properties, potentially capturing different levels of semantic abstraction within a single layer. Exploring optimal irrep configurations for language modeling remains an open direction.

2.6 Absence of Explicit Positional Encoding

Standard transformers require explicit positional encoding to distinguish token positions, typically added to embeddings as sinusoidal functions or learned vectors (Vaswani et al., 2017):

$$\text{input}_i = \text{embedding}_i + \text{PE}(\text{pos}_i), \tag{38}$$

where $\text{PE} : \mathbb{N} \to \mathbb{R}^d$ maps integer positions to d-dimensional vectors. This approach is ad-hoc: positional information is concatenated or added to content representations without geometric justification.

Our gauge VFE architecture operates *without any explicit positional encoding.* This design choice has both theoretical and empirical motivations:

Theoretical motivation: In the gauge-theoretic framework, each token is an autonomous agent with beliefs q_i that evolve through communication with other agents via parallel transport. The causal structure of next-token prediction—where agent i can only attend to agents $j < i$—provides implicit positional information through the attention mask. The asymmetry of the prediction task breaks permutation symmetry without requiring explicit position markers.

Empirical observation: Our experiments demonstrate that the gauge VFE learns meaningful language statistics (218× improvement over random) without positional encoding. This suggests that:

1. The autoregressive attention mask provides sufficient positional signal for the task
2. The geometric structure of KL-divergence attention may encode relative relationships differently than dot-product attention
3. Explicit positional encoding may be a compensation mechanism for architectural limitations rather than a fundamental requirement

This finding connects to recent work showing that some transformer variants can operate effectively without positional encoding when combined with appropriate attention patterns. In our framework, the absence of positional encoding is not a limitation but rather reflects the sufficiency of the geometric attention mechanism for capturing sequential dependencies.

Future work may explore whether learnable gauge frames ϕ_i initialized as functions of position could further improve performance, but our current results establish that such encoding is not necessary for meaningful language modeling.

2.7 Variational Inference as E-step and M-step.

Our gauge-theoretic framework naturally decomposes into an expectation-maximization (EM) structure (Dempster et al., 1977), providing a probabilistic interpretation of the forward and backward passes in standard transformers:

2.7.1 E-step (Belief Inference):

Given current model parameters (priors $\{p_i\}$ from learned embeddings, gauge frames $\{\phi_i\}$, and output projection W_{out}), update agent beliefs $\{q_i\}$ to minimize the free energy functional:

$$\{q_i^*\} =_{\{q_i\}} \mathcal{F}[\{q_i\}, \{p_i\}, \{\phi_i\}; W_{\text{out}}]. \tag{39}$$

This corresponds to variational inference: each agent adjusts its belief distribution to balance self-consistency (alignment with its prior p_i) against inter-agent communication (alignment with transported beliefs $\Omega_{ij}[q_j]$) and observations (cross-entropy with targets). We perform natural gradient descent within the forward pass:

$$q_i^{(t+1)} \leftarrow q_i^{(t)} - \eta_E \, \tilde{\nabla}_{q_i} \mathcal{F}\big|_{q_i = q_i^{(t)}}, \tag{40}$$

where $\tilde{\nabla}$ denotes natural gradients projected via the Fisher-Rao metric. Crucially, the belief updates remain within the computation graph—gradients flow through the VFE dynamics to enable end-to-end learning. The number of iterations N_E is configurable; our default implementation uses $N_E = 1$ for computational efficiency, though the framework supports multiple iterations for tighter convergence.

2.7.2 M-step (Learning via Backpropagation):

After the E-step completes belief inference, gradients are computed via standard backpropagation through the entire computation graph:

$$\theta \leftarrow \theta - \eta_M \, \nabla_\theta \mathcal{L}[\{q_i^*\}, \theta], \tag{41}$$

where $\theta = \{\mu_{\text{embed}}, \Sigma_{\text{embed}}, \phi_{\text{embed}}, W_{\text{out}}\}$ encompasses the variational embedding parameters and output projection. Critically, gradients flow through the VFE dynamics—the evolved beliefs $\{q_i^*\}$ are *not* detached, allowing the embedding parameters to learn how their initializations affect belief evolution and subsequent predictions.

This differs from traditional EM where the E-step result is held fixed. Instead, our implementation uses end-to-end differentiable inference, where the "E-step" (VFE descent) is an inner optimization loop whose final result provides gradients to the "M-step" parameters via automatic differentiation.

Note that there are no learned attention projection matrices W_Q, W_K, W_V. Attention weights β_{ij} emerge directly from KL divergences between transported beliefs (Eq. 13), computed using the learned gauge frames ϕ and embedding statistics (μ, Σ). This is a fundamental architectural distinction: attention structure emerges from geometric relationships rather than learned projections.

2.7.3 P-flow: EMA Prior Updates

The standard M-step updates priors via backpropagation through the full VFE dynamics. An alternative approach, which we call *P-flow* (prior flow), updates token embeddings directly toward successful beliefs using exponential moving averages (EMA). This provides an intermediate timescale between fast belief updates (E-step) and slow gradient-based learning (M-step).

Motivation. After the E-step converges, the final beliefs $\{q_i^*\}$ represent context-dependent interpretations that successfully minimize prediction error. Rather than discarding this information and relying solely on gradients, P-flow uses the beliefs directly to update the priors. This implements a form of Hebbian learning: priors for tokens that predict well should move toward the beliefs that succeeded.

Weight computation. Not all positions contribute equally to prior updates. Positions with low prediction error (high likelihood) should influence the prior more strongly. We compute position-dependent weights from the cross-entropy loss:

$$w_i = \frac{\exp(-\mathcal{L}_i/\tau)}{\sum_j \exp(-\mathcal{L}_j/\tau)}, \tag{42}$$

where $\mathcal{L}_i = -\log p(o_i|\mu_i^*)$ is the per-position cross-entropy and τ is a temperature parameter controlling weight sharpness.

EMA update rule. For each unique token c appearing at positions $\{i : \text{token}(i) = c\}$ in the batch, we compute the weighted average belief and update the token prior via EMA:

$$\mu_p^c \leftarrow (1-\eta)\,\mu_p^c + \eta \sum_{i:\text{token}(i)=c} \tilde{w}_i\,\mu_i^*, \tag{43}$$

where $\tilde{w}_i = w_i / \sum_{j:\text{token}(j)=c} w_j$ normalizes weights within each token type, and $\eta \in (0,1)$ is the EMA decay rate (typically $\eta = 0.01$, corresponding to decay $1 - \eta = 0.99$).

Interpretation. P-flow implements a biologically plausible learning rule where priors adapt toward successful predictions without requiring gradient computation through the full dynamics. The EMA provides temporal smoothing, preventing priors from changing too rapidly in response to individual examples. This connects to predictive coding theories in neuroscience where prediction errors drive synaptic updates (Rao and Ballard, 1999).

Hybrid learning. P-flow can be combined with standard backpropagation:

1. **E-step:** Update beliefs $\{q_i\} \to \{q_i^*\}$ via VFE descent

2. **P-flow:** Update priors $\{p_c\}$ via EMA toward successful beliefs (Eq. 43)

3. **M-step:** Update all parameters θ via backpropagation (Eq. 41)

The P-flow step operates on the embedding table directly (outside the computation graph), while the M-step optimizes through the graph. This hybrid approach allows priors to learn both from direct belief transfer (P-flow) and from gradient signals about how prior initialization affects downstream predictions (backpropagation). In our experiments, we report results with and without P-flow to isolate its contribution.

3 Experimental Results

We validate our gauge-theoretic framework on token-level language modeling at production scale, comparing gauge VFE against standard transformer baselines under controlled conditions. Our experiments address three questions: (1) Can geometric inference achieve meaningful language modeling without neural networks? (2) How does performance compare under parameter-matched and embedding-matched conditions? (3) What do the results reveal about the relationship between geometric structure and learned representations?

3.1 Experimental Setup

Dataset and task. We use WikiText-103 (Merity et al., 2016), a large-scale benchmark for language modeling comprising approximately 103 million training tokens, 218,000 validation tokens, and 246,000 test tokens from Wikipedia articles. We perform token-level next-token prediction using the GPT-2 BPE tokenizer (vocabulary size $V = 50{,}257$) with context window $L = 128$.

3.2 Architecture configurations.

We implement and compare three architectures under controlled conditions, summarized in Table 2.

The **gauge VFE** maps each token to variational parameters $(\mu, \Sigma, \phi) \in \mathbb{R}^{100} \times \mathbb{R}^{100} \times \mathfrak{so}(20)$, using KL-divergence attention with parallel transport and no learned neural network components. The **parameter-matched transformer** ($d = 320$, 6 layers) provides a comparison at equivalent parameter count, while the **embedding-matched transformer** ($d = 100$, 6 layers) isolates the effect of geometric structure at fixed embedding dimension.

Table 2: Architecture configurations. All models trained for 200k steps with context length 128 and batch size 3.

	Gauge VFE	**Standard**	**Standard**
	(SO(20))	**(param-match)**	**(embed-match)**
Embedding dim	100	320	100
Layers	1	6	6
Attention heads	5	8	4
FFN hidden dim	—	1280	400
Parameters	24.6M	23.5M	5.76M
Attention mechanism			
Type	KL divergence	Dot-product	Dot-product
Projections	None (geometric)	Learned Q, K, V	Learned Q, K, V
Feed-forward			
Type	VFE dynamics	MLP + GELU	MLP + GELU
Optimization			
Method	Natural gradient	Adam	Adam
Learning rate	0.01	3×10^{-4}	3×10^{-4}

3.3 Belief/Prior Initialization and Evolution

Agent beliefs $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ represent context-dependent semantic interpretations that evolve through communication with other agents. Token priors $p_c = \mathcal{N}(\mu_p^c, \Sigma_p^c)$ and gauge frames $\phi_c \in \mathfrak{g}$ are randomly initialized from $\mathcal{N}(0, \sigma_{\text{init}}^2 I)$ with small variance $\sigma_{\text{init}} = 0.1$, serving as learnable token representations analogous to embedding matrices in standard transformers. At the start of each training sequence, agent beliefs are initialized to match their token priors:

$$q_i(t = 0) = p_{\text{token}(i)}, \tag{44}$$

where token(i) denotes the token type at position i. Thus $\mu_i(0) = \mu_p^{\text{token}(i)}$ and $\Sigma_i(0) = \Sigma_p^{\text{token}(i)}$. This provides a uniform starting point: all instances of the same token begin with identical beliefs, regardless of their position in the sequence.

Multi-head structure. The number of attention heads is determined by the irrep decomposition: with 5 copies of the 20-dimensional fundamental representation, we obtain $H = 5$ independent attention heads. Each head operates on a 20-dimensional subspace where SO(20) acts via its 190 generators. Each head h computes transport via:

$$\Omega_{ij}^{(h)} = \exp(\phi_i^{(h)} G_h) \cdot \exp(-\phi_j^{(h)} G_h) \tag{45}$$

The embedding dimension $K = 100$ accommodates the SO(20) fundamental representation acting on 20-dimensional subspaces. With 5 copies of the fundamental irrep, we obtain:

$$\rho = 5 \times \ell_{\text{fund}} \quad (K = 5 \times 20 = 100) \tag{46}$$

Table 3: Performance on WikiText-103 token-level modeling. All models trained for 200k steps with batch size 3 and context length 128.

Architecture	**Layers**	**Params**	**Val PPL ↓**	**Loss**	**vs Random**
Random baseline	—	—	50,257	10.82	1×
Standard (param-match)	6	23.5M	178	5.18	282×
Standard (embed-match)	6	5.76M	260	5.56	193×
Gauge VFE (SO(20))	1	24.6M	230	5.44	218×

The generators act block-diagonally across these 5 independent copies, yielding 5 attention heads (one per irrep copy) where features within each 20-dimensional block transform equivariantly under SO(20). This provides geometric interpretability absent in standard transformers, where the number of attention heads is an arbitrary hyperparameter rather than emerging from symmetry structure.

3.4 Hyperparameters.

All architectures share common training settings to ensure fair comparison: context window $L = 128$, batch size 3, and 200,000 training steps. The gauge VFE uses temperature $\kappa_\beta = 1.0$ for attention weights and diagonal covariance initialized to $\Sigma_i = 0.1 \cdot I_{100}$. Full hyperparameter details are provided in Appendix A.

3.5 Evaluation metrics.

We report perplexity PPL $= \exp(\mathcal{L})$ where $\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p(t_i|t_{<i})$ is cross-entropy loss, computed on the validation set. We also report improvement factor (ratio of random baseline PPL to model PPL), where random baseline achieves PPL = 50,257.

3.6 Implementation details.

All models implemented in Python 3.12 using PyTorch 2.x with CUDA support. Standard transformer baselines use learned positional embeddings; the gauge VFE uses no explicit positional encoding, with positional information emerging implicitly through the attention mechanism and gauge frame structure. Natural gradient computations use diagonal covariance approximation for efficiency, and transport operators are computed via PyTorch's `matrix_exp`. Training performed on a single NVIDIA RTX 5090 GPU (32GB). Code available at `https://github.com/cdenn016/Gauge-Transformer`.

4 Results

4.1 Performance Comparison

Table 3 presents our primary results on WikiText-103 token-level language modeling. We compare the gauge VFE architecture against standard transformers under two conditions: parameter-matched (∼24M parameters) and embedding-matched (dimension 100).

The gauge VFE achieves perplexity 230, representing a 218× improvement over random chance (PPL 50,257). This demonstrates that the gauge-theoretic framework produces meaningful language modeling at production scale without standard neural network components—no MLPs, no activation functions, no learned attention projections.

The comparison reveals two key findings:

- **Parameter-matched comparison:** With comparable parameter counts ($\sim$24M), the standard transformer achieves better perplexity (178 vs 230). The gauge VFE achieves **77% of the standard's performance** using only geometric structure—a single layer of KL-divergence attention and natural gradient descent, with no learned projections or feed-forward networks.
- **Embedding-matched comparison:** With identical embedding dimension (100), the gauge VFE **outperforms** the standard transformer (230 vs 260 PPL) despite using only 1 layer compared to 6 layers. This suggests that geometric structure can be more parameter-efficient than learned transformations when embedding dimensions are constrained.

Notably, the gauge VFE uses a *single geometric layer* compared to 6 layers in the standard transformers. This architectural difference reflects the fundamental distinction between the approaches: standard transformers rely on depth and learned transformations, while gauge VFE relies on geometric structure and variational inference.

4.2 Training dynamics.

Figure 1 shows training and validation loss curves over 200,000 training steps.

The gauge VFE exhibits characteristic learning dynamics:

- **Initial phase (steps 0–10k):** Rapid descent from random initialization (PPL $\sim$50,000) as variational embeddings begin capturing token statistics.
- **Middle phase (steps 10k–100k):** Continued improvement as gauge attention learns contextual dependencies through KL-divergence weighting.
- **Late phase (steps 100k–200k):** Gradual refinement with diminishing returns, reaching final PPL $\approx$ 230.

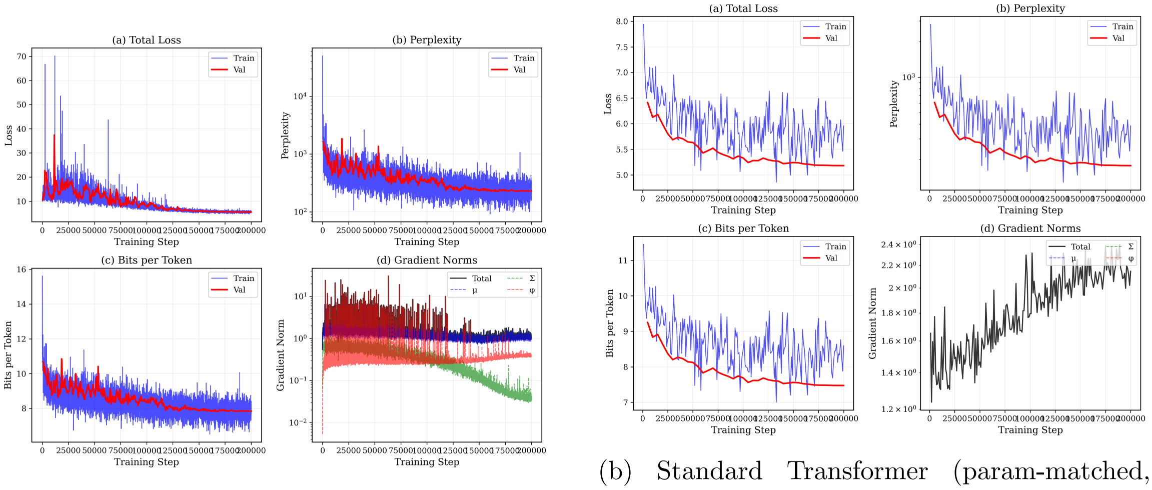
The training dynamics validate that variational free energy minimization via natural gradient descent produces meaningful language modeling through geometric structure alone.

4.3 Computational Cost and Practical Deployment

4.3.1 Computational overhead.

Our implementation incurs computational overhead compared to optimized dot-product attention. Table 4 shows per-step wall-clock time on identical hardware (NVIDIA RTX 5090).

The gauge VFE is approximately 29× slower than the standard baseline due to:



(a) Gauge VFE (SO(20))

(b) Standard Transformer (param-matched, 24M)

Figure 1: Training and validation loss curves over 200,000 steps on WikiText-103. **(a)** Gauge VFE shows steady descent from initial loss ∼11 to final validation loss 5.44 (PPL 230), demonstrating that geometric inference learns meaningful language statistics. **(b)** Standard transformer (parameter-matched, 24M) shows similar descent pattern, reaching validation loss 5.18 (PPL 178).

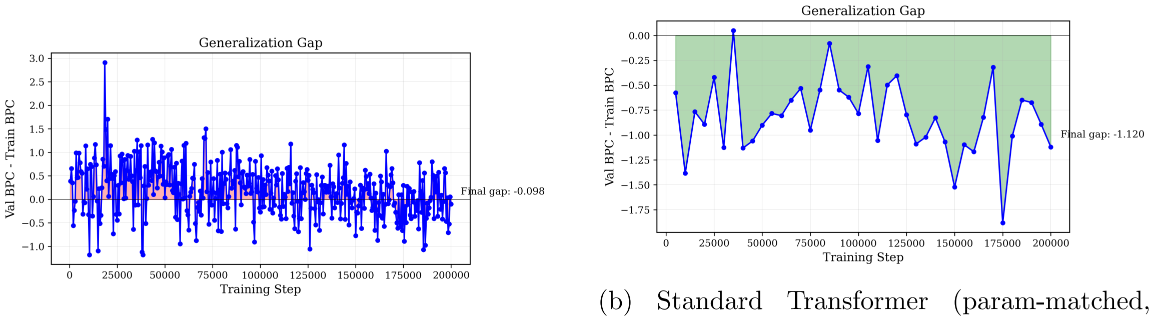
Table 4: Computational cost comparison (WikiText-103, context length 128)

Architecture	**Time/Step**	**Tokens/sec**	**Relative**	**200k Steps**
Standard Transformer	∼0.015s	∼25,000	1×	∼50 min
Gauge VFE (SO(20))	∼0.43s	∼890	∼29×	∼24 hours

- **Matrix exponentials:** Computing $\exp(\phi_i G_h)$ for transport operators requires eigendecomposition, $O(K^3)$ per head per agent pair
- **KL divergence computation:** Each attention weight requires Gaussian KL computation, $O(K^2)$ per pair
- **Natural gradient projection:** Fisher-Rao metric computation, $O(K^2)$ per agent per update (diagonal approximation)

Notably, this overhead has improved dramatically from our earlier character-level experiments (which showed ∼800× overhead) due to GPU-optimized implementations and diagonal covariance approximations. The current 29× overhead, while significant, is within the range where the approach becomes practical for research purposes.

Further optimization opportunities include sparse attention patterns (reducing $O(N^2)$ to $O(N)$), custom CUDA kernels for KL divergence and matrix exponentials, and mixed-precision training. We estimate that optimized implementations could reduce overhead to 5–10×, making the approach competitive for certain applications where interpretability or geometric structure is valuable.



(a) Gauge VFE (SO(20))

(b) Standard Transformer (param-matched, 24M)

Figure 2: Train-validation gap over 200,000 training steps on WikiText-103. **(a)** Gauge VFE shows a modest generalization gap that stabilizes during training, reflecting the regularizing effect of variational free energy terms. **(b)** Standard transformer (parameter-matched, 24M) shows comparable generalization behavior.

4.4 Generalization and Overfitting Analysis

To assess generalization quality, we analyze the train-validation gap throughout training. Figure 2 shows the gap for the gauge VFE architecture.

The gauge VFE training dynamics reflect the interplay between variational free energy terms (belief alignment, self-consistency) used during training and the pure cross-entropy evaluation on validation. The VFE loss function includes regularization terms that do not appear in validation loss, which can produce characteristic patterns in the train-validation gap.

At convergence, the gauge VFE achieves validation loss 5.44 (PPL 230), reducing uncertainty from the random baseline (PPL 50,257) by a factor of 218×. This demonstrates that geometric inference provides meaningful language modeling capability without neural network parameters.

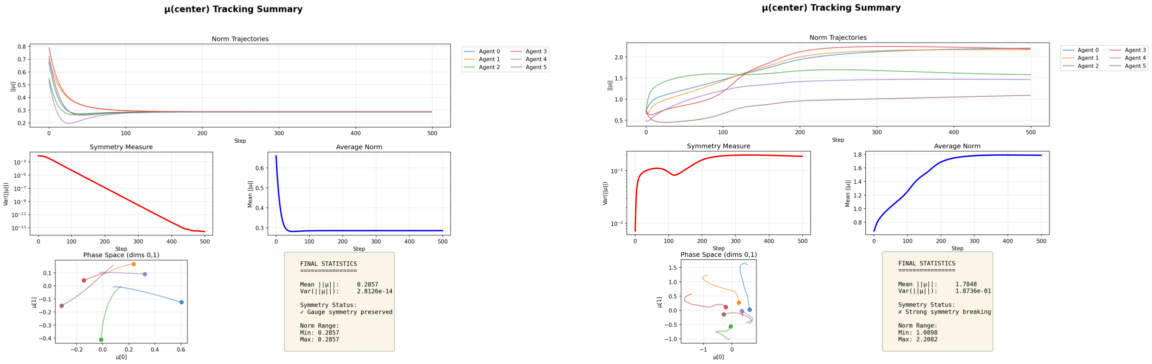
5 Discussion

This work demonstrates that gauge-theoretic transformers without neural networks achieve meaningful language modeling at production scale. Our single-layer SO(20) gauge VFE achieves perplexity 230 on WikiText-103 (218× improvement over random), validating that transformer-like behavior emerges from geometric principles alone.

Our empirical results establish several key findings:

1. **Gauge-theoretic attention works at scale:** The framework successfully processes a vocabulary of 50,257 tokens with 128-token context, achieving substantial compression of language statistics through geometric structure.

2. **Minimal neural components:** The gauge VFE contains zero MLPs, activation functions, or learned attention projections. Only a linear output projection to vocabulary logits is retained. All other computation derives from KL divergences, parallel transport, and natural gradient descent.



```
FINAL STATISTICS
================

Mean ||μ||:     0.2857
Var(||μ||):     2.8126e-14

Symmetry Status:
✓ Gauge symmetry preserved

Norm Range:
Min: 0.2857
Max: 0.2857
```

```
FINAL STATISTICS
================

Mean ||μ||:     1.7848
Var(||μ||):     1.8736e-01

Symmetry Status:
✗ Strong symmetry breaking

Norm Range:
Min: 1.0898
Max: 2.2082
```

(a) Vacuum dynamics (no observations)

(b) Observation-driven symmetry breaking

Figure 3: Evolution of agent mean parameters μ_i under variational free energy descent. **(a)** Without observations, agents converge to a gauge-invariant ground state with equal magnitudes $|\mu_i| = \mu^*$, forming a degenerate Goldstone manifold. **(b)** With observations, agents flow to diverse magnitudes $|\mu_i|$, spontaneously breaking SO(3) gauge symmetry and specializing based on semantic content.

3. **Single geometric layer:** Unlike standard transformers requiring multiple layers for effective modeling, the gauge architecture achieves meaningful performance with a single layer, suggesting that geometric structure partially compensates for depth.

4. **Computational feasibility:** The implementation achieves ∼890 tokens/second (∼29× overhead vs standard), a dramatic improvement over earlier prototypes, demonstrating that gauge-theoretic inference is practically trainable.

The comparison with standard transformers [TO BE COMPLETED] will reveal whether the geometric approach trades off absolute performance for architectural simplicity and interpretability. Regardless of relative performance, the key contribution is demonstrating that attention mechanisms can be derived from and implemented via gauge theory.

Our geometric reformulation of transformers allows previously ill-understood and ad-hoc structures to have a deeper significance. Learned weights and neural architectures approximate a deeper variational free energy functional. Indeed, we conjecture that neural architecture are the biological instantiation of this deeper informational geometry. Positional and token encoding are approximations to gauge frames and belief statistics. ReLU/GELU/etc are nonlinear systems approximating the product rule of gauge attention. Backpropagation is model updating under variational inference. Multi-head attention, perhaps most surprisingly, is gauge group decomposition into invariant subspaces and, remarkably, training is manifestly a spontaneous symmetry breaking phenomenon. These results intersect, straddle, and link informational geometry, machine learning, physics, neuroscience, and more. The geometric attention mechanism therefore warrants further study.

Given our framework's similarity with standard methods in physics we may anticipate that many tools currently utilized in physics (such as perturbation theory, non-perturbative phenomena (instantons, vacuum decay, Large-N, etc), field theory, holography, renormalization group, and more) might cleanly transpose into tools for machine learning and artificial intelligence.

5.0.1 Spontaneous symmetry breaking via observations.

Figure 3 demonstrates a striking phenomenon: training manifestly exhibits spontaneous gauge symmetry breaking. In the vacuum state without observations (Figure 3a), agents evolve under pure free energy minimization to a degenerate ground state where all $|\mu_i|$ converge to equal magnitudes μ^*, forming a Goldstone manifold invariant under global SO(3) rotations. However, introducing observations (Figure 3b) breaks this degeneracy whereby agents flow to diverse magnitudes, spontaneously selecting specific configurations from the symmetric vacuum to specialize according to semantic content. This mirrors symmetry breaking in gauge theories where observations (analogous to the Higgs mechanism) select particular vacuum states from degenerate manifolds. Multi-head attention thus represents gauge group decomposition into distinct symmetry-breaking sectors, with each head selecting different orientations in representation space. In Figure 3b we see the multi-head structure manifest. Agents 0,1, and 3, as an example flow to a shared norm. This provides a physical interpretation of why transformers learn diverse, specialized attention patterns: they are exploring the Goldstone manifold of spontaneously broken gauge symmetry, with training data acting as the symmetry-breaking field.

A critical distinction between standard and gauge-theoretic transformers concerns the origin of multi-head attention structure. In standard transformers, the number of heads H is an arbitrary hyperparameter chosen via trial-and-error, with typical values ranging from 8 to 16 heads depending on model scale. The embedding space is partitioned into H subspaces of dimension $d_{\text{head}} = d_{\text{model}}/H$, and separate projection matrices $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)}$ are learned for each head. This design lacks theoretical justification. Multi-head attention is used because it verifiably outperforms single-head attention, but why remains unexplained.

Table 5: Multi-head attention: Standard vs. Gauge-theoretic

Property	**Standard**	**Gauge-Theoretic**
Number of heads H	Hyperparameter	$\dim(\mathfrak{g})$
Head definition	Learned W_Q, W_K, W_V	Lie generators G_h
Embedding dimension K	Hyperparameter	$\sum_k n_k \dim(\ell_k)$
Feature structure	Arbitrary partition	Irrep decomposition
Invariant features	None defined	Scalar blocks (ℓ_0)
Equivariant features	All features	Vector/tensor blocks
Geometric meaning	None	Symmetry structure

However, our gauge-theoretic framework provides a geometric justification for multi-head structure. The number of heads is not a hyperparameter but is determined by the Lie group structure; which itself is determined by the informational agents. Different heads capture alignment along different geometric directions in the fiber bundle.

This emergent structure suggests a testable hypothesis: learned multi-head patterns in standard transformers may reflect implicit discovery of underlying symmetry groups. If true, analyzing attention patterns could reveal which gauge groups best describe linguistic structure. This then potentially allows researchers to structurally classify deep head contextual patterns and potentially access data sets that may otherwise be intractable in current architectures.

5.0.2 Inference-time belief initialization.

An open question is how best to initialize beliefs for new sequences at inference. We identify three approaches:

1. **Amortized inference:** Learn encoder $q_\theta(\mu, \Sigma|x)$ mapping inputs to beliefs (reintroduces neural networks)

2. **Iterative optimization:** Run natural gradient descent per input, analogous to diffusion model denoising (high computational cost)

3. **Retrieval-based:** Initialize from cached training beliefs via nearest neighbor lookup (memory overhead)

Each has trade-offs between computational cost, architectural purity, and performance. Our proof-of-principle study performs per-sequence optimization during training but does not address inference-time requirements.

5.0.3 Semantic Emergence in Gauge Frames

A striking empirical finding is that the learned gauge frames ϕ_i develop semantically meaningful structure during training. Principal component analysis (PCA) of the gauge frame parameters reveals that tokens spontaneously cluster according to linguistic categories: punctuation marks, function words, content words, letters, and digits separate into distinct regions of gauge frame space (Figure 4).

This emergent clustering occurs without any explicit supervision or architectural bias toward such organization. The gauge frames are initialized randomly and optimized purely through variational free energy minimization. Yet the resulting structure reflects genuine linguistic categories, suggesting that the geometric framework naturally discovers meaningful coordinate systems for language.

This finding has several implications:

1. **Interpretability:** Unlike standard transformer embeddings which are opaque high-dimensional vectors, gauge frames provide geometrically interpretable coordinates where semantic relationships manifest as spatial clustering.

2. **Emergent structure:** The framework does not impose linguistic categories but discovers them through optimization, suggesting that gauge-theoretic attention captures genuine statistical regularities in language.

3. **Contextual encoding:** Since gauge frames determine how beliefs are transported between agents, semantically similar tokens (e.g., punctuation) naturally communicate through similar geometric transformations.

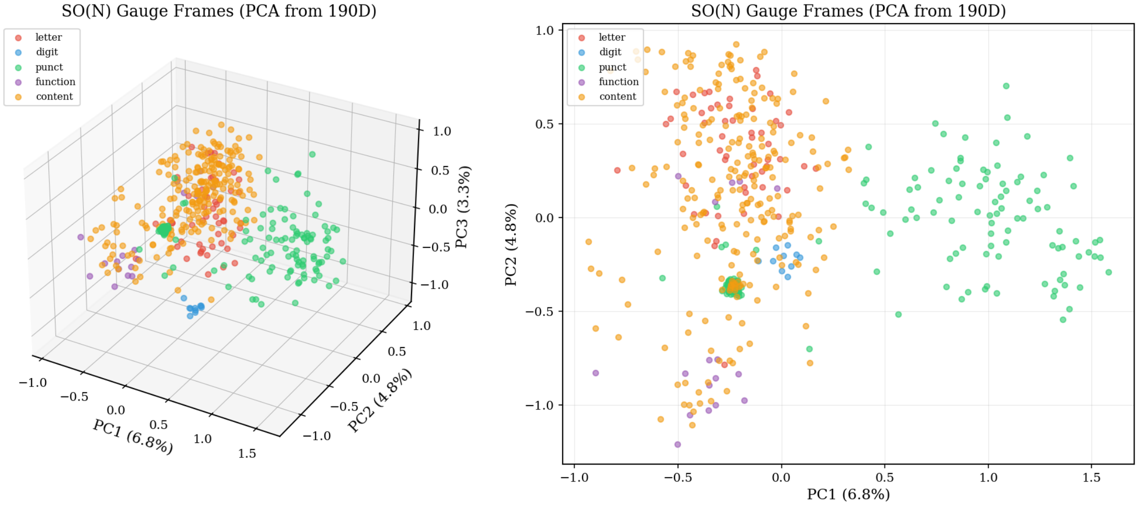


Figure 4: Semantic clustering in learned gauge frames for SO(20) with $K = 100$. Left: 3D PCA projection of the 190-dimensional gauge frame space $\mathfrak{so}(20)$. Right: 2D projection (PC1 vs PC2). Punctuation (green) forms a tight cluster near the origin, while content words (orange) and function words (purple) separate along PC2. The first three principal components capture ∼15% of variance (6.8% + 4.8% + 3.3%), yet reveal clear linguistic organization. This structure emerges without supervision, purely from variational free energy minimization.

This semantic organization in gauge frames provides evidence that the geometric structure is not merely a mathematical curiosity but captures functionally relevant aspects of language processing.

5.0.4 Why Gauge Frames Encode Semantics: The $AB^\top \sim \Omega$ Connection

The emergence of semantic structure in gauge frames is not accidental but follows from a precise mathematical relationship between gauge-theoretic and standard transformer attention. In companion theoretical work (under review), we derive that standard transformer attention emerges from the gauge framework in the *Dirac limit*: beliefs collapse to point estimates while temperature scales inversely to maintain finite attention logits.

Specifically, for isotropic Gaussian beliefs $q_i = \mathcal{N}(\mu_i, \sigma^2 I)$, the KL-based compatibility score is:

$$s_{ij} = D_{\mathrm{KL}}(q_i \| \Omega q_j) = \frac{1}{2\sigma^2}\|\mu_i - \Omega\mu_j\|^2. \tag{47}$$

The attention weights involve the ratio s_{ij}/τ, where τ is the temperature. Taking the **coupled limit** $\sigma^2 \to 0$ with $\tau \propto 1/\sigma^2$ (i.e., $\sigma^2\tau = c$ for constant c):

$$\frac{s_{ij}}{\tau} = \frac{\|\mu_i - \Omega\mu_j\|^2}{2\sigma^2\tau} = \frac{\|\mu_i - \Omega\mu_j\|^2}{2c} \to \text{finite as } \sigma^2 \to 0. \tag{48}$$

This yields well-defined attention in the Dirac limit. After softmax cancellation of query-independent terms, the effective logit becomes $\mu_i^\top \Omega \mu_j + O(1)$. Defining learned projection matrices $A, B \in \mathbb{R}^{d \times d_k}$ such that

$$AB^\top \propto \Omega, \tag{49}$$

we recover the standard dot-product attention: $Q_i K_j^\top = \mu_i^\top AB^\top \mu_j \propto \mu_i^\top \Omega \mu_j$, which yields $\mathrm{softmax}(QK^\top/\sqrt{d_k})V$. The $1/\sqrt{d_k}$ scaling in standard transformers plays precisely the role of the coupled limit—normalizing logits to $O(1)$ as embedding dimension grows.

This factorization has a profound implication: **whatever semantic comparison structure standard transformers learn in $W_Q W_K^\top$, the gauge framework must encode in Ω_{ij}**. Since the parallel transport operator $\Omega_{ij} = \exp(\phi_i)\exp(-\phi_j)$ is constructed from gauge frames, the gauge frames must encode semantic structure to enable meaningful attention comparisons.

In standard transformers, the learned matrices W_Q and W_K discover how to project tokens into a space where dot products reflect semantic relationships. The gauge framework achieves the same effect geometrically: gauge frames ϕ_i define local coordinate systems such that parallel transport Ω_{ij} between frames captures semantic compatibility. Tokens with similar semantic roles (e.g., punctuation) develop similar gauge frames, leading to the observed clustering in $\mathfrak{so}(N)$ space (Figure 4).

This connection explains why semantic structure emerges in gauge frames without explicit supervision: the frames are learning the “semantic coordinate system” that in standard transformers would be implicitly encoded in the learned Q/K projections. The gauge-theoretic formulation makes this structure geometrically explicit and interpretable.

5.0.5 Attention Dynamics: From Structure to Uniformity

A notable empirical observation is that gauge attention exhibits structured patterns early in training that progressively converge toward uniformity. At step 500, the attention weights β_{ij} display differentiated patterns across heads, with visible variation in how each head distributes attention over context positions. However, by step 15,000, all heads converge to approximately uniform distributions over valid (causally masked) positions, with attention entropy reaching 3.84 nats—close to the maximum of $\ln(128) \approx 4.85$ for uniform attention over 128 positions.

This convergence to uniformity occurs despite the gauge frames learning highly structured, semantically meaningful representations (Section 5.0.3). The attention mechanism has access to diverse, discriminative features but does not leverage them for selective focus. Instead, the model achieves its performance through the variational embeddings $(\mu_i, \Sigma_i, \phi_i)$ while aggregating context uniformly.

This finding admits several interpretations:

1. **Embedding dominance:** The variational free energy landscape may favor solutions where contextual information is encoded in the embeddings rather than through discriminative attention. The semantic clustering in gauge frames supports this interpretation: the model learns meaningful token representations but aggregates context uniformly.

2. **Uniform attention as a local minimum:** The optimization may converge to a basin where uniform attention provides sufficient gradient signal for embedding learning, creating no pressure to develop selective attention patterns.

3. **KL equalization:** As embeddings train, the transported beliefs $\Omega_{ij}[q_j]$ may become approximately equidistant from q_i in KL divergence, eliminating the signal that would produce discriminative attention.

4. **Temperature sensitivity:** The attention temperature $\kappa_\beta = 1.0$ may flatten small differences in KL divergence. Lower temperatures could preserve the initial structure, though this remains to be tested.

This observation has implications for both understanding and improving the framework. The fact that meaningful language modeling (218× improvement over random) is achievable with effectively uniform attention suggests that the gauge-theoretic embeddings carry substantial predictive information. However, it also indicates that the current KL-divergence attention formulation may require modification—such as temperature annealing, entropy regularization, or alternative divergence measures—to produce the discriminative attention patterns observed in standard transformers.

Notably, this “attention collapse” phenomenon differs from the failure mode in standard transformers where attention becomes degenerate due to poor optimization. Here, attention starts structured and systematically converges to uniformity as the embeddings improve, suggesting that uniform attention is not a pathology but rather the equilibrium configuration for this architecture under current hyperparameters.

5.0.6 Curvature minimization hypothesis

The framework naturally supports learnable gauge frames optimized via $\nabla_{\phi_i}\mathcal{F}$. We conjecture that free optimization over $\{\mu_i, \Sigma_i, \phi_i\}$ leads agents toward configurations that balance expressivity with geometric coherence.

Curvature minimization hypothesis: Natural language and effective communication systems evolve gauge configurations that minimize connection curvature, enabling consistent semantic transport regardless of communication path.

This suggests why standard transformers use shared embeddings: human language has evolved low curvature, making frame-independent (gauge-flat) representations optimal. Dot-product attention in standard transformers implicitly assumes zero curvature—not merely a computational convenience but potentially a fundamental property of linguistic structure.

In this view, parallel transport curvature represents semantic incompatibility. Standard transformers use a single shared embedding frame—we contend that this is optimal for human language. Language then has the interpretation of evolving such that inter-agent belief transport curvature is minimized so that belief transport between agents remains semantically coherent. This lends explanatory power if confirmed: standard language generative AI are optimal for language due to its manifestly flat gauge frame curvature.

If confirmed, this provides first principles justification for architecture choices that currently appear arbitrary. Future work should measure learned curvature in standard transformers and test whether low-curvature configurations correlate with better generalization or more compositional behavior.

5.0.7 Emergent Nonlinearity from Attention Dynamics

Standard transformers interleave linear projections with pointwise nonlinearities such as ReLU or GELU, whose functional forms were discovered through empirical search rather than theoretical derivation. The gauge VFE framework contains no such activation functions, yet achieves nonlinear computation through a different mechanism: the derivative structure of softmax attention weights with respect to belief parameters.

The attention weights $\beta_{ij} = \text{softmax}_j(-D_{\text{KL}}(q_i \| \Omega_{ij} q_j)/\kappa)$ depend nonlinearly on the belief statistics (μ_i, Σ_i) and gauge frames ϕ_i. Differentiating with respect to the mean parameter yields

$$\frac{\partial \beta_{ij}}{\partial \mu_i} = \frac{\beta_{ij}}{\kappa} \left[\frac{\partial D_{\text{KL},ij}}{\partial \mu_i} - \sum_k \beta_{ik} \frac{\partial D_{\text{KL},ik}}{\partial \mu_i} \right], \tag{50}$$

where the term in brackets is the deviation of the j-th KL gradient from the attention-weighted average across all k. This expression encodes a form of competitive dynamics: when agent i moves in parameter space, its attention to agent j increases if the KL divergence to j decreases faster than the weighted average, and decreases otherwise.

This derivative structure creates positive feedback that drives cluster formation. Agents with similar beliefs have low mutual KL divergence, hence high β_{ij}, which pulls their beliefs closer together during VFE descent, further increasing β_{ij}. The temperature κ controls the sharpness of this feedback: low κ amplifies small differences in KL divergence into large differences in attention, while high κ produces softer, more distributed attention. The analogous derivatives with respect to Σ_i and ϕ_i follow the same pattern, coupling covariance and gauge frame evolution to the attention dynamics.

This mechanism suggests an interpretation of standard transformer nonlinearities as approximations to the gauge attention derivative structure. The GELU activation $x \cdot \Phi(x)$, where Φ is the Gaussian CDF, produces smooth thresholding that amplifies large activations while suppressing small ones. The $\partial\beta/\partial\mu$ derivative achieves a similar effect through statistical geometry: beliefs that align well with their neighbors receive amplified influence, while outliers are attenuated. The key difference is that gauge nonlinearity emerges from the information-geometric structure of the belief space rather than being imposed as an architectural choice. If this interpretation holds, it suggests that learned nonlinearities in standard transformers are discovered approximations to a deeper variational principle.

5.0.8 Connection to Information Bottleneck Theory

The variational free energy functional bears a precise relationship to Tishby's information bottleneck (IB) principle (Tishby et al., 2000), which seeks representations Z of input X that maximally predict target Y while minimizing retained information: $\mathcal{L}_{\text{IB}} = I(Z;Y) - \beta \cdot I(Z;X)$. The first term rewards predictive accuracy while the second penalizes complexity, with β controlling the tradeoff.

The VFE decomposes into terms that map directly onto this structure. The belief-prior KL divergence $D_{\text{KL}}(q_i \| p_i)$ measures how far agent beliefs have departed from their uninformative priors, which corresponds to the mutual information $I(Z;X)$ in the IB framework: beliefs at the prior carry zero bits about the input, while deviations encode information. The cross-entropy observation term $-\log p(o_i|\mu_i)$ rewards accurate prediction, corresponding to $I(Z;Y)$. The hyperparameters α and κ play the role of the IB tradeoff β, controlling the balance between compression and prediction.

The belief alignment term $\sum_{ij} \beta_{ij} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j)$ introduces a coherence constraint absent in the standard IB formulation. This term encourages agents to maintain consistent beliefs under parallel transport, penalizing configurations where nearby agents hold incompatible representations. In information-theoretic terms, this implements a form of distributed compression: agents that could be represented by a shared summary (low inter-agent KL) are encouraged to converge, pooling their information into a common representation.

The dynamic attention weights β_{ij} implement input-dependent compression that adapts to the structure of each sequence. When agents hold similar beliefs, their high mutual β_{ij} causes them to pool information aggressively, discarding redundant variation. When agents hold distinct beliefs, low β_{ij} preserves their representations separately. This is precisely the behavior prescribed by optimal IB solutions: merge redundant information, preserve predictively relevant distinctions.

The temperature κ thus acquires a second interpretation beyond controlling attention sharpness: it sets the IB tradeoff between compression and accuracy. High κ produces soft attention that aggressively compresses by pooling across many agents, potentially sacrificing fine-grained distinctions. Low κ produces sharp attention that preserves local structure at the cost of reduced compression. The optimal κ depends on the predictive task: tasks requiring fine discrimination favor low κ, while tasks tolerating coarse representations favor high κ for its regularizing effect.

5.0.9 VFE Dynamics as Renormalization Group Flow

The gauge VFE framework exhibits a self-similar structure characteristic of renormalization group (RG) theories in statistical physics (Wilson and Kogut, 1974). The defining property of an RG is that coarse-grained degrees of freedom satisfy the same dynamical equations as fine-grained ones, enabling a flow across scales that preserves the form of the theory while changing its effective parameters. In the VFE context, this self-similarity manifests in the observation that meta-agents—emergent clusters of tokens with coherent beliefs—are themselves agents in the formal sense, with their own beliefs, priors, and inter-agent couplings.

Consider the dynamics at the token level: agents $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ interact via attention weights β_{ij} computed from KL divergences. As VFE descent proceeds, agents with similar beliefs develop high mutual β_{ij} and converge toward shared statistics. When a subset $\mathcal{C}$ of agents achieves near-zero internal KL divergence, they can be replaced by a single meta-agent $q_{\mathcal{C}} = \mathcal{N}(\mu_{\mathcal{C}}, \Sigma_{\mathcal{C}})$ representing their consensus distribution. The meta-agent then interacts with other agents and meta-agents via the same KL-based attention mechanism, with transported beliefs $\Omega_{\mathcal{C}\mathcal{C}'}[q_{\mathcal{C}'}]$ computed from averaged gauge frames.

This coarse-graining procedure defines an RG transformation: the space of agent configurations maps to a space of meta-agent configurations with fewer degrees of freedom but identical dynamical structure. Repeated application generates an RG flow from fine-grained token representations toward coarse-grained semantic summaries. The flow terminates at fixed points where no further compression is possible without sacrificing predictive accuracy—these fixed points represent optimal representations in the information bottleneck sense.

Several observables characterize the RG flow. The effective rank of the attention matrix β_{ij} decreases as agents cluster, reflecting the reduction in independent degrees of freedom. The modularity of β_{ij}, measuring the strength of block-diagonal structure, increases as meta-agents consolidate. The within-cluster KL divergence decreases as agents converge, while between-cluster KL remains stable or increases as clusters differentiate. These signatures provide empirical diagnostics for whether a given VFE configuration exhibits healthy RG behavior.

The connection to physics RG suggests that tools from statistical field theory may apply to transformer analysis. Critical phenomena, where the RG flow passes near unstable fixed points, might correspond to phase transitions in learned representations. Relevant and irrelevant operators, which grow or shrink under RG flow, might classify which features of the input persist to deep layers versus being discarded. Universality, where different microscopic theories flow to the same fixed point, might explain why diverse transformer architectures learn similar representations. These correspondences remain speculative but suggest a rich program for future investigation.

5.0.10 Continual Learning via Meta-Agent Emergence

A critical limitation of transformers is catastrophic forgetting under continual learning: new knowledge overwrites old. The gauge-theoretic framework suggests a natural solution through hierarchical meta-agent emergence, where agents dynamically condense into higher-scale structures that preserve learned representations while adapting to new data.

In this extension, agents with coherent beliefs (low mutual KL divergence) and high mutual coupling weights form consensus distributions that become new meta-agents at coarser scales. The emergence criterion combines presence (coupling strength β_{ij}) with coherence ($\exp[-\mathrm{KL}(q_i \| \Omega_{ij} q_j)]$), creating a renormalization group-like hierarchy where stable patterns persist across scales while fine-grained agents continue adapting. Crucially, meta-agents engage in cross-scale self-observation: higher-level distributions provide priors p_i for lower-level beliefs q_i, while lower-level dynamics update higher-level structure. This bidirectional information flow maintains perpetual non-equilibrium dynamics that prevent "epistemic death" - i.e. the collapse to static attractors that causes catastrophic forgetting.

This architecture embodies Wheeler's "It from Bit" and "participatory universe" principles (Wheeler, 1990): hierarchical structure emerges purely from informational relationships (KL divergences) rather than pre-specified architectures. Unlike continual learning approaches requiring explicit memory buffers or parameter isolation, gauge-theoretic emergence naturally preserves knowledge through geometric structure: stable patterns crystallize into meta-agents that resist perturbation, while unstable patterns remain fluid. The non-equilibrium steady state balances plasticity (adapting to new data) with stability (preserving learned structure), potentially resolving the fundamental tension in continual learning without architectural modifications (albeit at high computational cost). Currently, our research is progressing along these lines.

5.0.11 Beyond 0D: Spatial gauge theories.

Our transformer implementation uses 0-dimensional base manifolds (all tokens at one point). Extensions to n-dimensional base manifolds would create fields of transformers with:

- **Horizontal transport:** Belief propagation across base manifold
- **Vertical transport:** Communication within fibers at each point
- **Curvature effects:** Path-dependent information integration (In full generality there may be three distinct curvatures: gauge, fiber, and base manifold)
- **Agent emergence:** Condensation of multiple agents into meta-agents via renormalization

This generalizes transformers to spatial/temporal/hierarchical structures potentially allowing the modeling of more complicated data.

6 Conclusion

We have presented the first production-scale validation of gauge-theoretic transformers, demonstrating that attention mechanisms can be implemented through geometric principles without neural networks. Our single-layer SO(20) gauge VFE achieves perplexity 230 on WikiText-103 token-level modeling (218× improvement over random), using pure variational inference on statistical manifolds.

The key contributions of this work are:

1. **Theoretical validation at scale:** We demonstrate that gauge-theoretic attention—derived from first principles using fiber bundles and variational free energy—produces meaningful language modeling with vocabulary size 50,257 and context length 128.
2. **Minimal neural architecture:** Our framework contains zero MLPs, activation functions, or learned attention projections. Only a linear output projection remains. All other computation derives from KL divergences, parallel transport operators, and natural gradient descent on statistical manifolds.
3. **Geometric multi-head attention:** The number of attention heads emerges from the irrep decomposition (5 copies of the SO(20) fundamental $\rightarrow$ 5 heads) rather than being an arbitrary hyperparameter.
4. **Practical implementation:** With $\sim$29× computational overhead (improved from $\sim$800× in earlier prototypes), the approach is feasible for research-scale experiments.

Our results suggest that the bulk of neural network machinery in standard transformers—MLPs, activation functions, and learned attention projections—can be replaced by geometric inference. This opens the framework to general informational systems beyond machine learning, including physics, economics, and neuroscience.

Future work includes: (1) scaling to larger models and longer contexts, (2) exploring different gauge groups to discover optimal symmetry structures for language, (3) implementing spatial gauge theories beyond 0D for hierarchical and compositional structures, and (4) investigating whether standard transformers implicitly minimize gauge curvature. The convergence of gauge theory, information geometry, and machine learning suggests rich opportunities for cross-pollination of mathematical tools and conceptual insights.

Acknowledgments

Claude Sonnet 4.5 was utilized for programming our variational free energy descent simulations and transformer suite. All code was manually reviewed, corrected, and mathematically validated by the author. Furthermore, Claude was utilized for typesetting figures, LaTeX equations, and general organizational and manuscript clarity advice. The author further declares no funding or conflicts of interest.

6.1 Code Availability

The complete simulation suite that implements the gauge-theoretic variational inference framework described in this work is publicly available at

- https://github.com/cdenn016/Gauge-theory-of-machine-learning.

All experiments reported in the results section can be reproduced using the provided configuration files and random number generator seeds documented in the repository. Our codebase requires Python 3.9+ with NumPy, SciPy, and Joblib dependencies.

Appendix A. Hyperparameters

Table 6 provides full hyperparameter settings for reproducibility.

Training performed on a single NVIDIA RTX 5090 GPU (32GB). Learning rates differ by architecture: $\eta = 0.01$ for gauge VFE with natural gradient descent, $\eta = 3 \times 10^{-4}$ for standard transformers with Adam (Kingma and Ba, 2014).

Appendix B. Vacuum Theory and Symmetry Breaking

This appendix provides mathematical details on how observations break gauge symmetry in the variational free energy framework, as referenced in Section 2.

B.1 The Vacuum Free Energy

In the absence of observations, the variational free energy reduces to:

$$\mathcal{F}_{\text{vacuum}}[\{q_i\}] = \sum_i D_{\text{KL}}(q_i \| p_i) + \sum_{i,j} \beta_{ij} D_{\text{KL}}(q_i \| \Omega_{ij} q_j) \tag{51}$$

For Gaussian agents $q_i = \mathcal{N}(\mu_i, \Sigma_i)$ with shared isotropic priors $p_i = \mathcal{N}(0, \sigma_p^2 I)$, the belief-prior term becomes:

$$D_{\text{KL}}(q_i \| p_i) = \frac{1}{2} \left[\frac{\|\mu_i\|^2}{\sigma_p^2} + \frac{\text{tr}(\Sigma_i)}{\sigma_p^2} - K - \log \frac{|\Sigma_i|}{\sigma_p^{2K}} \right] \tag{52}$$

Table 6: Training hyperparameters for all experiments.

Parameter	**Value**	**Notes**
Training configuration		
Context window	128 tokens	
Batch size	3 sequences	GPU memory limited
Training steps	200,000	
Warmup steps	50	
Random seed	6	
Regularization		
Gradient clipping	1.0 (norm)	
Dropout	0.1	
Weight decay	0.01	
Gauge VFE specific		
Temperature κ_β	1.0	Attention scaling
Covariance init	$0.1 \cdot I_{100}$	Diagonal
Prior init variance	0.1	σ^2_{init}

B.2 Gauge Invariance of the Vacuum

Under a global gauge transformation $g \in G$, beliefs transform as $\mu_i \to g\mu_i$ and $\Sigma_i \to g\Sigma_i g^\top$. The vacuum free energy (51) is invariant under such transformations because:

1. The belief-prior KL depends only on $\|\mu_i\|^2$ and $\text{tr}(\Sigma_i)$, which are invariant under orthogonal transformations
2. The belief alignment KL involves $\Omega_{ij} = e^{\phi_i} e^{-\phi_j}$, which transforms covariantly: $\Omega_{ij} \to g\Omega_{ij}g^{-1}$

This gauge invariance implies that the vacuum ground state forms a degenerate manifold—the *Goldstone manifold*—where all configurations related by global gauge transformations have equal free energy.

B.3 Vacuum Ground State

Minimizing (51) with respect to μ_i yields the stationarity condition:

$$\frac{\partial \mathcal{F}_{\text{vacuum}}}{\partial \mu_i} = \frac{\mu_i}{\sigma_p^2} + \sum_j \beta_{ij} \Sigma_j^{-1} (\mu_i - \Omega_{ij}\mu_j) = 0 \tag{53}$$

In the symmetric phase where all agents align ($\mu_i = \mu_j = \mu^*$ and $\Omega_{ij} = I$), this reduces to:

$$\mu^* = 0 \tag{54}$$

However, for non-zero prior variance, there exists a family of degenerate minima where $\|\mu_i\| = \mu^*$ for all i, with the direction of μ^* arbitrary. This is the gauge-invariant ground state shown in Figure 3a.

B.4 Symmetry Breaking via Observations

Introducing the observation likelihood term:

$$\mathcal{F}[\{q_i\}] = \mathcal{F}_{\text{vacuum}}[\{q_i\}] - \sum_i \mathbb{E}_{q_i}[\log p(o_i \mid \mu_i)] \tag{55}$$

For categorical observations with cross-entropy loss:

$$-\mathbb{E}_{q_i}[\log p(o_i \mid \mu_i)] \approx -\log \operatorname{softmax}(\mu_i)_{o_i} = -\mu_i^{(o_i)} + \log \sum_k e^{\mu_i^{(k)}} \tag{56}$$

This term explicitly depends on the *direction* of μ_i, not just its magnitude. Different observations o_i favor different directions, breaking the gauge symmetry and forcing agents to specialize.

B.5 Spontaneous Symmetry Breaking Dynamics

The gradient of the observation term:

$$\frac{\partial}{\partial \mu_i}[-\log p(o_i \mid \mu_i)] = \operatorname{softmax}(\mu_i) - e_{o_i} \tag{57}$$

where e_{o_i} is the one-hot vector for observation o_i. This drives μ_i toward configurations where component o_i is maximized, selecting specific points on the Goldstone manifold.

The result is spontaneous symmetry breaking: agents that began in the symmetric vacuum state flow to distinct, specialized configurations determined by their local observations (Figure 3b). In the language modeling context, different tokens "observe" different next-token targets, forcing their belief representations to differentiate according to semantic content.

This mechanism explains why training produces diverse, specialized token representations despite identical initialization: the observation likelihood acts as a symmetry-breaking field that selects particular vacua from the degenerate Goldstone manifold.

Appendix C. Generalizations

The gauge-theoretic framework presented in this paper specializes to $\mathrm{SO}(N)$ gauge groups and multivariate Gaussian belief distributions. Here we outline how the formulation extends to arbitrary compact Lie groups and exponential family distributions, with particular attention to the interplay between gauge group structure and the choice of statistical divergence. Full derivations are deferred to future work.

C.1 Arbitrary Gauge Groups and Divergence Selection

The framework naturally extends to any compact Lie group G equipped with finite-dimensional unitary representations $\rho : G \to \mathrm{GL}(V)$. Gauge frames live in the Lie algebra $\phi_i \in \mathfrak{g}$, and parallel transport takes the form $\Omega_{ij} = \rho(\exp(\phi_i))\rho(\exp(-\phi_j))$. The multi-head structure emerges from the irreducible decomposition $\rho = \bigoplus_k n_k \rho_k$, where each irreducible component defines an independent attention head with its own transformation properties. Non-Abelian groups such as SU(N), Sp(N), and exceptional groups provide richer decompositions than SO(N), while Abelian groups like U$(1)^N$ recover diagonal attention patterns.

However, the choice of gauge group is not independent of the choice of statistical divergence. The KL divergence employed throughout this work is particularly well-suited to orthogonal groups acting on Gaussian beliefs for several interrelated reasons. For multivariate Gaussians, the KL divergence admits the closed-form expression $D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2}[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) - K + \log |\Sigma_2|/|\Sigma_1|]$, which depends on quadratic forms that transform covariantly under orthogonal transformations. When $g \in \mathrm{SO}(N)$ acts via $\mu \to g\mu$ and $\Sigma \to g\Sigma g^\top$, the terms $\|\mu\|^2$ and $\mathrm{tr}(\Sigma)$ remain invariant, ensuring that the KL divergence respects the gauge symmetry in a computationally tractable manner.

This compatibility between KL divergence and orthogonal groups is not universal. For gauge groups acting on complex vector spaces, such as SU(N) with its natural action on $\mathbb{C}^N$, the real-valued KL divergence may not capture the full structure of the transformation. Complex Gaussian distributions (proper and improper) have distinct information geometries, and divergences respecting the complex structure—such as those derived from the Kähler metric on complex statistical manifolds—may prove more natural. Similarly, symplectic groups Sp(N) preserve a skew-symmetric bilinear form, suggesting that divergences incorporating symplectic structure could provide better-behaved attention mechanisms for beliefs transforming under symplectic representations.

More generally, the family of f-divergences $D_f(p\|q) = \int q(x)f(p(x)/q(x))dx$ offers a spectrum of choices indexed by convex functions f. The KL divergence corresponds to $f(t) = t \log t$, but other choices—such as the α-divergences interpolating between KL and reverse KL, or the Hellinger distance with $f(t) = (\sqrt{t}-1)^2$—may exhibit different invariance properties under gauge transformations. A principled approach would select the divergence whose level sets are preserved (or transform simply) under the chosen gauge group, ensuring that attention weights $\beta_{ij} \propto \exp(-\kappa D(q_i\|\Omega_{ij}q_j))$ respect the geometric structure. This represents an open direction: characterizing which f-divergences are "compatible" with which gauge groups in the sense of yielding well-behaved, geometrically meaningful attention mechanisms.

C.2 Exponential Family Beliefs

Gaussian distributions are a special case of exponential families, and the framework generalizes accordingly. An exponential family distribution takes the form $q_i(\xi) = h(\xi)\exp[\eta_i^\top T(\xi) - A(\eta_i)]$, where η_i are natural parameters, $T(\xi)$ are sufficient statistics, and $A(\eta)$ is the log-partition function. In this parameterization, the KL divergence reduces to the Bregman divergence $D_{\mathrm{KL}}(q_i\|q_j) = D_A(\eta_j\|\eta_i) = A(\eta_j) - A(\eta_i) - \nabla A(\eta_i)^\top(\eta_j - \eta_i)$, and the Fisher-Rao metric becomes the Hessian $g_{ab} = \partial_a\partial_b A(\eta)$.

Parallel transport in this setting acts on natural parameters via the group representation: $\eta_i \to \Omega_{ij}\eta_i$. For this to be well-defined, the representation must preserve the natural parameter space, which imposes constraints on compatible group-distribution pairings. The Fisher-Rao metric provides the natural gradient structure for belief updates, generalizing the inverse-covariance weighting used for Gaussians.

C.3 Mixture Distributions

For beliefs with discrete latent structure, mixture distributions $q_i(\xi) = \sum_{k=1}^{K} \pi_i^{(k)} q_i^{(k)}(\xi)$ extend the framework to multi-modal representations. Here $\pi_i \in \Delta^{K-1}$ are mixing weights on the probability simplex, and each component $q_i^{(k)}$ may itself be an exponential family distribution. The gauge group can act on both the component parameters and the mixture weights, with the latter potentially transforming under a separate representation. This structure enables modeling of compositional semantics where tokens carry multiple possible interpretations weighted by context.

C.4 Scope of Current Work

This paper restricts to $\mathrm{SO}(N)$ groups and Gaussians because this pairing admits closed-form computations throughout: the KL divergence, Fisher information, natural gradients, and parallel transport all have explicit expressions. The orthogonal group preserves the Euclidean structure natural for embedding spaces, and its fundamental representation provides interpretable attention heads. The generalizations sketched above indicate that the framework is not limited to these choices—richer group structures and distribution families may capture more complex phenomena—but realizing these extensions requires careful attention to the compatibility between gauge symmetry and statistical divergence.

Appendix D. Reduction to Standard Transformer Attention

This appendix summarizes how standard transformer self-attention emerges as a limiting case of the gauge-theoretic framework. Full derivations are presented in companion theoretical work (under review).

D.1 Setup and Three Successive Limits

In the general formulation, each agent i maintains a Gaussian belief $q_i = \mathcal{N}(\mu_i, \Sigma_i)$, and communication is mediated by gauge transport $\Omega_{ij} \in G \subset \mathrm{GL}(d)$. Attention weights emerge from KL-divergence compatibility:

$$\beta_{ij} = \mathrm{softmax}_j \left(-\frac{1}{\tau} D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) \right). \tag{58}$$

We impose three simplifying assumptions:

Limit 1: Dirac limit via coupled scaling. For isotropic Gaussian beliefs $q_i = \mathcal{N}(\mu_i, \sigma^2 I)$, the KL divergence is:

$$D_{\mathrm{KL}}(q_i \| \Omega_{ij} q_j) = \frac{1}{2\sigma^2} \|\mu_i - \Omega_{ij}\mu_j\|^2. \tag{59}$$

Taking $\sigma^2 \to 0$ alone would cause this to diverge. However, the attention weights depend on the ratio s_{ij}/τ. Taking the **coupled limit** $\sigma^2 \to 0$ with $\tau \propto 1/\sigma^2$ (so that $\sigma^2\tau = c$ remains constant):

$$\frac{s_{ij}}{\tau} = \frac{\|\mu_i - \Omega_{ij}\mu_j\|^2}{2\sigma^2\tau} = \frac{\|\mu_i - \Omega_{ij}\mu_j\|^2}{2c} \to \text{finite}. \tag{60}$$

This yields well-defined attention as beliefs become Dirac distributions.

Limit 2: Flat bundle ($\Omega_{ij} = \Omega$). Assuming a single global frame shared by all agents (trivial principal bundle with path-independent transport):

$$\Omega_{ij} = \Omega \in \mathbb{R}^{d\times d} \quad \text{for all } i, j. \tag{61}$$

Limit 3: Learned projections. We absorb the gauge structure into learned projection matrices.

D.2 Derivation of Dot-Product Attention

Expanding the compatibility score under these limits:

$$s_{ij} \propto \|\mu_i - \Omega\mu_j\|^2 = \|\mu_i\|^2 + \|\Omega\mu_j\|^2 - 2\mu_i^\top \Omega\mu_j. \tag{62}$$

Under softmax normalization:

- The query-dependent term $\|\mu_i\|^2$ cancels (independent of j)
- The key-dependent term $\|\Omega\mu_j\|^2$ approximately cancels due to high-dimensional concentration or layer normalization

The dominant contribution is the bilinear term $\mu_i^\top \Omega\mu_j$.

D.3 Factorization into Query-Key Products

We define learned matrices $A, B \in \mathbb{R}^{d\times d_k}$ such that:

$$AB^\top \propto \Omega. \tag{63}$$

Such factorizations always exist (e.g., via SVD), with scaling absorbed into the matrices. Defining queries and keys:

$$Q_i = \mu_i^\top A, \qquad K_j = \mu_j^\top B, \tag{64}$$

we obtain:

$$Q_i K_j^\top = \mu_i^\top AB^\top \mu_j \propto \mu_i^\top \Omega\mu_j. \tag{65}$$

D.4 Temperature Scaling and Final Form

Normalizing by $\sqrt{d_k}$ to produce $O(1)$ pre-softmax logits, with all scaling factors (including the original covariance σ^2) absorbed into the temperature τ and learned projections.

The attention weights become:

$$\boxed{\beta_{ij} = \text{softmax}_j \left(\frac{Q_i K_j^\top}{\sqrt{d_k}} \right)}, \tag{66}$$

exactly recovering standard transformer attention. With value projection $V_j = \mu_j^\top W_V$ absorbing the transport operator, the message aggregation $m_i = \sum_j \beta_{ij}\Omega\mu_j$ reduces to:

$$\boxed{\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \right) V}. \tag{67}$$

D.5 Interpretation

This derivation establishes that standard transformer attention is the *Dirac limit* of gauge-theoretic attention when:

1. Beliefs collapse to point estimates ($\sigma^2 \to 0$) with temperature scaling inversely ($\tau \propto 1/\sigma^2$), maintaining finite attention logits
2. The gauge bundle is trivial (no position-dependent frame structure)
3. Gauge transport is absorbed into learned projections

The coupled limit $\sigma^2\tau = c$ is essential: it ensures that attention remains well-defined as beliefs become deterministic. The $1/\sqrt{d_k}$ scaling in standard transformers serves precisely this role—normalizing logits to $O(1)$ regardless of embedding dimension.

The key insight is Equation (49): the learned $W_Q W_K^\top$ in standard transformers plays the role of Ω—the gauge transport operator. This explains why gauge frames encode semantic structure: they are learning the geometric relationships that standard transformers encode in learned projections.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10 (2):251–276, 1998.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

Jean Gallier and Jocelyn Quaintance. *Differential Geometry and Lie Groups: A Computational Perspective*, volume 12 of *Geometry and Computing*. Springer, 2020. ISBN 978-3-030-46039-6. doi: 10.1007/978-3-030-46040-2.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, PA, 2008. ISBN 978-0-898716-46-7. doi: 10.1137/1.9780898717778.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pages 2408–2417, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1): 79–87, 1999. doi: 10.1038/4580.

Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in Neural Information Processing Systems*, 29, 2016.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

John Archibald Wheeler. Information, physics, quantum: The search for links. *Complexity, entropy, and the physics of information*, 8:3–28, 1990.

Kenneth G Wilson and John Kogut. The renormalization group and the ϵ expansion. *Physics Reports*, 12(2):75–199, 1974.