Article

Not peer-reviewed version

Toward a Possibilistic Language Model: Grounding Large Language Model Uncertainty in Epistemic Support-Point Theory and Possibility Theory

Moriba Kemessia Jah *

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0920.v1

Keywords: possibility theory; epistemic uncertainty; large language model; possibilistic attention; Epistemic Support-Point Filter; non-Bayesian inference; vocabulary falsification; epistemic width; Possibilistic Cramér–Rao Bound; transformer architecture

 doi:10.20944/preprints202604.0920.v1

Article

Toward a Possibilistic Language Model: Grounding Large Language Model Uncertainty in Epistemic Support-Point Theory and Possibility Theory

Moriba Kemessia Jah [1,2,3]

[1] Black Swan Research Group, GaiaVerse, Ltd.; moriba@utexas.edu
[2] Jah Decision Intelligence Group
[3] Aerospace Engineering & Engineering Mechanics, University of Texas at Austin

Abstract

Contemporary large language models (LLMs) generate text by sampling from probability distributions over vocabulary tokens, implicitly asserting that uncertainty about language is well-modeled by calibrated stochastic processes. This paper argues that such probabilistic closure is epistemically unjustified in the regime of language: meaning is bounded by admissible interpretation, not governed by frequentist statistics. The precise claim is not that probability cannot model language but that probabilistic closure is an epistemically overcommitted representation for token uncertainty under bounded admissibility—one that prevents the architecture from representing categorical inadmissibility and compels the assignment of positive probability mass to tokens that should not exist in the support at all. We propose the Possibilistic Language Model (PLM), a novel architecture grounding language generation in possibility theory and the Epistemic Support-Point Filter (ESPF) framework of Jah and Haslett (2025). The PLM replaces softmax probability distributions over tokens with possibilistic compatibility fields over vocabulary support sets; it replaces maximum-likelihood training with a Possibilistic Cramér–Rao regularized entropy-minimization objective; and it replaces standard scaled dot-product attention with Epistemic Possibilistic Attention (EPA), a falsification-driven attention operator that gates keys by admissible innovation geometry rather than by likelihood weighting. The PLM is not merely compatible with these foundations: it is the unique architecture forced by the TEAG axioms when instantiated over a discrete vocabulary manifold. The governing equation of token generation is a tropical Hamilton–Jacobi equation in the max-plus semiring, with the vocabulary surprisal field as Hamiltonian and the PCRB as minimum action per generation step. Epistemic Possibilistic Attention is the discretized Lax–Oleinik operator of this system. The minimax medoid commitment is proved to be the geodesic attractor of the surviving vocabulary well, not a heuristic selection rule. We define (not merely propose) the vocabulary possibility distribution and EPA operator; we derive the PCRB regularization from established ESPF theory inherited from Jah (2026a); we prove, conditionally on VFI maintenance, that the PLM produces non-degenerate generation from actual vocabulary tokens; and we prove that the PLM recovers a standard transformer in the Gaussian epistemic limit, which is the zero-temperature limit of the tropical Hamilton–Jacobi framework established in Jah (2026d). We are explicit throughout about what is fully self-contained here, what is inherited from prior TEAG works, and what remains as open proof obligations. The architecture naturally produces diagnostics—necessity, epistemic width, and surprisal—that quantify when a generation step is epistemically supported versus epistemically strained. We discuss initialization, training, inference, and multi-modal extensions, and identify open theoretical obligations alongside a concrete research agenda. The PLM is not presented as a drop-in replacement for probabilistic LLMs, but as a principled alternative for language tasks where epistemic humility, interpretability, and bounded admissibility matter more than distributional calibration.

 doi:10.20944/preprints202604.0920.v1

Keywords: possibility theory; epistemic uncertainty; large language model; possibilistic attention; Epistemic Support-Point Filter; non-Bayesian inference; vocabulary falsification; epistemic width; Possibilistic Cramér–Rao Bound; transformer architecture

---

1. Introduction

A unified epistemic program.

This paper is one instantiation of a broader research program called the Theory of Epistemic Abductive Geometry (TEAG)—a general theory of how evidence contracts a nested geometry of admissible hypotheses without forcing premature probabilistic closure. The name reflects three constitutive ideas: *epistemic* (the framework operates on admissibility under evidence, not probability under assumed statistics), *abductive* (inference proceeds by eliminating what cannot be, committing to the best surviving explanation), and *geometric* (the admissible support has shape, volume, nesting, and curvature that are primary objects of analysis). The unifying object across the TEAG program is the TEAG object:

$$\mathcal{E} = \big(H,\ \pi,\ \{H_\alpha\}_{\alpha\in[0,1]},\ C,\ A\big), \tag{1}$$

where H is a hypothesis space, $\pi : H \to [0,1]$ is a possibility field encoding ordinal admissibility, $H_\alpha = \{h \in H : \pi(h) \geq \alpha\}$ is the nested α-cut family, C is an evidence-driven contraction operator satisfying Popperian monotonicity ($C(\pi, y)(h) \leq \pi(h)$ always), and A is a minimax commitment rule.

Three prior works instantiate this object in different domains. Jah and Haslett (2025) develops the algorithmic realization of C and A as a recursive state estimator—the Epistemic Support-Point Filter (ESPF)—in which H is a finite support set of state hypotheses and C is geometric falsification through residual compatibility. Jah (2026b) lifts the same object to the measure-theoretic continuum, proving the conditions under which the credal envelope $\mathcal{P}_\pi$ induced by π contracts to a singleton probability law via Choquet-to-Lebesgue convergence—establishing probability as the collapse limit of the TEAG object rather than its primitive. Jah (2026a) characterizes which contraction operator C is uniquely justified by a minimax-entropy criterion over the α-cut volume family, proving the ESPF the unique optimal evidence-only filter in its class and establishing the Possibilistic Cramér–Rao Bound as a universal floor on admissible epistemic contraction per update. Jah (2026d) establishes the dynamical geometric foundation: proves that any inference system satisfying the TEAG axioms must obey a tropical Hamilton–Jacobi equation in the max-plus semiring (with the surprisal field as Hamiltonian), identifies the ESPF predict–update recursion as the Lax–Oleinik operator of this system, shows the minimax medioid is the geodesic attractor of the surviving well, and establishes the contact geometry of epistemic phase space. The present paper inherits this dynamical structure directly.

The present paper instantiates the TEAG object over token hypothesis spaces: H becomes the vocabulary V, π becomes token admissibility, $\{H_\alpha\}$ becomes nested admissible token clouds, C becomes compatibility-driven vocabulary falsification, and A becomes minimax medoid token commitment. The result is the Possibilistic Language Model (PLM).

PLM is forced by the dynamics, not merely inspired by them.

The relationship between the PLM and the wavefront dynamics of Jah (2026d) is stronger than compatibility. Jah (2026d) proves that any inference system satisfying the TEAG axioms must satisfy a tropical Hamilton–Jacobi equation

$$\Phi_{k+1}(v) = \Phi_k(v) \oplus \Phi_S(v) = \max\big(\Phi_k(v),\ \Phi_S(v)\big), \tag{2}$$

where $\Phi_k(v) = -\log \pi_k(v)$ is the impossibility field over the vocabulary, $\Phi_S(v) = \frac{1}{2}\|L_e^{-1}(q_k - e_v)\|^2$ is the surprisal field generated by the current context embedding, and $\oplus = \max$ is the tropical addition. This is not a modeling choice: Popperian contraction forces max-plus algebra (evidence can only

 doi:10.20944/preprints202604.0920.v1

increase impossibility, never decrease it), and the evidence-referencing axiom forces the Hamiltonian to depend only on the hypothesis position in embedding space, not on the impossibility gradient. No alternative update structure consistent with the TEAG axioms exists. The PLM is therefore not *one possibilistic architecture among many*—it is the *necessary consequence* of the tropical Hamilton–Jacobi dynamics when instantiated on a discrete vocabulary manifold. Every section of this paper should be read in that light.

The epistemic problem with probabilistic closure.

Every modern LLM answers the question "which token comes next?" by producing a probability distribution over a vocabulary V. The chain rule of probability decomposes a sequence $t_1, \ldots, t_n$ as

$$P(t_1, \ldots, t_n) = \prod_{k=1}^{n} P(t_k \mid t_1, \ldots, t_{k-1}), \tag{3}$$

and the model is trained to maximize the log-likelihood of a corpus. This is an elegant formulation with well-understood optimization theory.

The critique we advance is not that modern LLMs naively assume language is a stationary stochastic process—they do not; they model conditional sequence distributions over contexts. The critique is more precise: *probabilistic closure is an epistemically overcommitted representation for token uncertainty under bounded admissibility*. The softmax output head assigns strictly positive probability to every token in V at every step—including tokens that are semantically incoherent, syntactically impossible, or factually ruled out by local context. This is not a calibration failure; it is a structural feature of the representation. A model operating under softmax cannot represent categorical inadmissibility; it can only assign low (but always nonzero) probability. It is, as Kalman warned in the context of estimation, an architecture that "displays the prejudices of its creator"—in this case, the prejudice that all token uncertainty is probabilistic (Kalman, 1982).

Possibility theory (Dubois and Prade, 1988; Zadeh, 1978) provides an alternative: an ordinal, maxitive calculus in which evidence eliminates inadmissible hypotheses rather than redistributing probability mass. A token with possibility zero is categorically excluded, not merely improbable. This is the epistemic posture the PLM adopts.

Overview.

We propose a Possibilistic Language Model that replaces probabilistic token distributions with possibilistic compatibility fields—ordinal rankings of tokens by their admissibility under bounded epistemic constraints—and replaces maximum-likelihood generation with a falsification-driven support contraction that eliminates inadmissible tokens rather than redistributing probability mass.

The theoretical foundations are those of the ESPF (Jah and Haslett, 2025), which maintains a finite set of explicit hypotheses ordered by geometric compatibility with evidence. In the language domain, "state" becomes latent semantic embedding, "measurement" becomes the observed token or context, and "evidence-driven pruning" becomes vocabulary falsification.

1.1. Contributions

The principal contributions of this paper are organized by their epistemic status, to avoid conflating what is defined here, what is derived from prior TEAG results, what is proved under stated conditions, and what remains open:

Definitions (self-contained constructions):

1. **Possibilistic vocabulary representation.** We define a vocabulary support set $\mathcal{X}_k \subset V$ at each generation step and associate with it a possibility distribution $\pi^{(k)} : V \to [0, 1]$ encoding admissibility rather than probability (Section 4).

 doi:10.20944/preprints202604.0920.v1

2. **Epistemic Possibilistic Attention (EPA).** We define an attention operator that replaces softmax likelihood weighting with a whitened innovation compatibility gate, falsifying keys whose residual energy exceeds an admissible innovation bound (Section 5).
3. **Epistemic diagnostics for generation.** We define necessity, surprisal, and epistemic tension for autoregressive generation (Section 9).

Derivations (inherited from prior TEAG theory):

4. **PCRB-regularized training objective.** We derive a training loss grounded in possibilistic entropy $H_\pi = \int_0^1 \log V_\alpha \, d\alpha$, with a PCRB penalty preventing inadmissible vocabulary collapse per generation step. The justification of the PCRB as a universal floor is inherited from Theorem 5.2 of Jah (2026a); it is not proved independently here (Section 7).

Conditional theorems (proved here, under stated assumptions):

5. **VFI for token support sets.** We extend the Volumetric Faithfulness Invariant of Jah and Haslett (2025) to vocabulary hypothesis spaces and prove that, *assuming the VFI is maintained through depth* (an open proof obligation), non-degenerate generation from actual vocabulary tokens is guaranteed (Section 8).
6. **Gaussian limit recovery.** We prove that, under conditions (G1)–(G3) on support cloud geometry and token co-occurrence statistics, the PLM recovers a standard transformer. These conditions are not trivially satisfied by all language domains; their scope is discussed explicitly (Section 11).

1.2. Relation to Existing Work

The present work sits at the intersection of three traditions: (i) possibility theory and ordinal uncertainty (Dubois and Prade, 1988; Zadeh, 1978), (ii) non-Bayesian filtering and the ESPF (Jah and Haslett, 2025; Jah, 2026a,2), and (iii) language model uncertainty quantification (Guo et al., 2017; Maddox et al., 2019; Malinin and Gales, 2021).

1.3. Conceptual Overview: Probabilistic Versus Possibilistic Generation

In a standard LLM, softmax assigns strictly positive probability to every token in V, including tokens that are semantically, syntactically, or factually inadmissible. The committed token is the argmax of this distribution. In the PLM, inadmissible tokens are falsified before any commitment is made; the surviving support cloud Ξ_k spans only admissible hypotheses; and commitment is the minimax medoid of that cloud—the most geometrically central surviving token—rather than the argmax of a probability distribution.

2. Background

2.1. Standard Transformer Language Models

A standard autoregressive transformer (Vaswani et al., 2017) represents context as a sequence of token embeddings $\{e_1, \ldots, e_k\} \subset \mathbb{R}^d$ and produces, at each position k, a probability distribution over V via scaled dot-product attention followed by a linear projection and softmax:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \quad P(t_k \mid t_{<k}) = \mathrm{softmax}(W_o h_k + b_o), \tag{4}$$

where h_k is the final hidden state at position k. The model is trained by minimizing

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{k=1}^{N} \log P(t_k \mid t_{<k}). \tag{5}$$

The softmax in both the attention mechanism and the output head enforces distributional closure: every token receives a strictly positive probability and the weights sum to one. This is the probabilistic commitment we seek to relax.

 doi:10.20944/preprints202604.0920.v1

2.2. Possibility Theory

Possibility theory (Dubois and Prade, 1988; Zadeh, 1978) provides a maxitive, ordinal calculus for reasoning under epistemic uncertainty. A possibility distribution $\pi : \Omega \to [0, 1]$ satisfies

$$\sup_{\omega \in \Omega} \pi(\omega) = 1, \tag{6}$$

and induces a possibility measure $\Pi(A) = \sup_{\omega \in A} \pi(\omega)$ and a dual necessity measure $N(A) = 1 - \Pi(\bar{A})$.

Conditioning is conjunctive: given evidence e, posterior admissibility is

$$\pi(\omega \mid e) = \min\big(\pi(\omega),\, \kappa(e \mid \omega)\big), \tag{7}$$

where $\kappa(e \mid \omega) \in (0, 1]$ is a compatibility score. Evidence contracts admissibility but cannot amplify it above prior support. This Popperian falsification semantic is central to the ESPF and to the PLM.

2.3. The Epistemic Support-Point Filter

The ESPF (Jah and Haslett, 2025) maintains a finite set of support points $\mathcal{X}_k = \{\chi_k^{(i)}\}_{i=1}^{M} \subset \mathbb{R}^n$ representing hypotheses not yet falsified by evidence. Prediction propagates each point deterministically through system dynamics and expands the set under bounded perturbation; measurement update evaluates each point via a whitened innovation metric and prunes those whose geometric surprisal exceeds an evidence-calibrated threshold.

Key quantities governing ESPF stability include:

- Geometric surprisal: $\mathcal{S}_k^{(i)} = \frac{1}{2}\|L_e^{-1} e_k^{(i)}\|_2^2$, where $e_k^{(i)} = y_k - h(\chi_{k|k-1}^{(i)})$ and L_e is the Cholesky factor of the innovation shape matrix Π_e.
- Compatibility: $\kappa_k^{(i)} = \exp(-\mathcal{S}_k^{(i)}) \in (0, 1]$.
- Possibilistic entropy: $H_\pi = \int_0^1 \log V_\alpha \, d\alpha$, where $V_\alpha = \mathrm{Vol}(\{\chi : \pi(\chi) \geq \alpha\})$.
- Possibilistic Cramér–Rao Bound (PCRB): $\kappa_k \geq \kappa_{\min}(I_k, n) = (1 - I_k)^{n/2}$.
- Volumetric Faithfulness Invariant (VFI): a joint condition on support geometry, spread parameter, and survivor count ensuring well-posedness.

Two companion papers are directly load-bearing for the PLM. Jah (2026a) proves that the ESPF is the unique optimal recursive estimator within the class of epistemically admissible evidence-only filters under a possibilistic minimax-entropy criterion. The PCRB is established as a universal lower bound on entropy reduction per measurement—not a heuristic penalty but a provable floor on admissible epistemic contraction. This uniqueness result carries into the PLM: the possibilistic cross-entropy objective and PCRB regularization of Section 7 inherit their justification from this theorem.

Jah (2026b) develops the measure-theoretic framework governing the transition from possibilistic to probabilistic inference, proving (via Choquet-to-Lebesgue convergence) the conditions under which a possibility distribution contracts to a probability density.

2.4. The Tropical Hamilton–Jacobi Foundation

Jah (2026d) establishes the result that carries most directly into PLM architecture: any TEAG-consistent inference system obeys a tropical Hamilton–Jacobi equation. We state the key results here because they transform PLM from an architecture paper into a derivation.

The two scalar fields.

The ESPF—and by inheritance the PLM—operates through two scalar fields on hypothesis space. The *impossibility field* $\Phi_\varnothing(h) = -\log \pi(h)$ encodes accumulated epistemic history: zero where a hypothesis enjoys full prior support, growing without bound as evidence withdraws that support. The *surprisal field* $\Phi_S(h) = \frac{1}{2}\|L_e^{-1}(y - g(h))\|^2$ encodes the tension between each hypothesis and the current observation in MVEE-whitened measurement space. In the language domain: $\Phi_\varnothing(v) = -\log \pi_k(v)$ is

 doi:10.20944/preprints202604.0920.v1

the token impossibility field, and $\Phi_S(v) = \frac{1}{2}\|L_e^{-1}(q_k - e_v)\|^2$ is the vocabulary surprisal field generated by the current context query q_k.

The update is an algebraic identity.

The possibilistic conjunctive update $\pi'(h) = \min(\pi(h), \kappa(h))$ is equivalent in field coordinates to:

$$\Phi^+_\varnothing(h) = \Phi^-_\varnothing(h) \oplus \Phi_S(h) = \max\big(\Phi^-_\varnothing(h),\, \Phi_S(h)\big). \tag{8}$$

This equality follows from $-\log\min(a,b) = \max(-\log a, -\log b)$: it is an algebraic identity, not a modeling choice or an analogy. The posterior impossibility field is the pointwise max-plus upper envelope of the prior and surprisal fields. Whichever source of impossibility is stronger at a given hypothesis governs its posterior admissibility.

The axioms force this structure.

Jah (2026d), Proposition 5.7, proves that three TEAG axioms jointly force Eq. (8) with no degrees of freedom. (i) Popperian contraction requires $\Phi^+(h) \geq \Phi^-(h)$ for all h; the unique binary operation on $[0,\infty)$ that is non-decreasing in both arguments, associative, and reduces to the identity under no evidence is the max operation. (ii) The evidence-referencing axiom requires that the surprisal depends only on each hypothesis's position in measurement space, not on the impossibility gradient at that point; this forces the Hamiltonian to be momentum-independent. (iii) A momentum-independent tropical Hamiltonian forces the Lax–Oleinik reduction to a pointwise max. No alternative update structure consistent with these axioms exists.

Epistemic time in generation.

Jah (2026d) introduces a discrete epistemic time that advances only upon registered evidence, and becomes continuous in the limit of constant information flux. In the PLM, this structure maps directly: each autoregressive generation step k is one *epistemic update cycle*. The sequence $k = 1, \ldots, T$ is an information trajectory through vocabulary space. Generation depth equals the number of registered epistemic events. A generation step with zero effective surprisal—where the context provides no vocabulary discrimination—does not advance the epistemic state in any meaningful sense; the model is in epistemic free fall between genuinely informative positions.

3. Related Work

Probabilistic LLM uncertainty.

Calibration (Guo et al., 2017), Bayesian neural networks (Maddox et al., 2019), conformal prediction (Angelopoulos and Bates, 2022), and Monte Carlo dropout (Gal and Ghahramani, 2016) all attempt to quantify uncertainty within the probabilistic framework. They improve distributional calibration but do not question the fundamental appropriateness of probability as the uncertainty representation for language—which is the object of the PLM's critique.

Imprecise and set-based uncertainty.

Credal sets (Walley, 1991) and outer probability measures maintain sets of distributions. These relax single-distribution commitment but do not provide a recursive, geometry-driven contraction mechanism suited to autoregressive generation.

Structured decoding and constrained generation.

A line of work imposes hard constraints on generation through vocabulary masking, grammar constraints, and logit post-processing (Hokamp and Liu, 2017; Hu et al., 2017; Post and Vilar, 2018). Top-k truncation (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020) perform post-hoc vocabulary pruning by discarding low-probability tokens. The PLM differs from all of these in a fundamental respect: pruning in the PLM is grounded in geometric admissibility in whitened innovation space and is

 doi:10.20944/preprints202604.0920.v1

applied *before* any commitment is made, not as a post-processing heuristic on a probability distribution. Constrained decoding methods retain the probabilistic output head and impose external constraints afterward; the PLM replaces the probabilistic output head entirely. Furthermore, constrained decoding provides no epistemic diagnostics—no signal of when the constraint is being violated or strained. The PLM's necessity, surprisal, and epistemic tension are continuously available at every generation step.

Retrieval-grounded and symbolic-neural hybrids.

Retrieval-augmented generation (Lewis et al., 2020) grounds generation in retrieved evidence but maintains probabilistic output distributions over retrieved tokens. Neuro-symbolic approaches (Marcus, 2020) impose symbolic constraints on neural generation. Neither provides a recursive, non-probabilistic falsification architecture operating over the full vocabulary embedding space at every generation step.

Token masking.

Masked language models (Devlin et al., 2019) use token masking during training to force context-based prediction but still train against a probabilistic objective. The PLM's vocabulary falsification operates at inference time under a possibilistic objective, not as a training-time masking heuristic.

Possibilistic reasoning in NLP.

Fuzzy logic (Zadeh, 1965) and possibility theory have been applied to lexical ambiguity (Dubois and Prade, 2000) and fuzzy information retrieval (Benferhat et al., 2002), but have not been developed into a full sequence-generation architecture with a recursive falsification mechanism.

Epistemic neural networks.

Prior-network (Malinin and Gales, 2021) and evidential deep learning (Senséoy et al., 2018) approaches treat distributional uncertainty using Dirichlet priors. These are second-order probabilistic methods, distinct from the ordinal, non-additive framework of possibility theory. They answer "how uncertain is the distribution?" rather than "which tokens are categorically inadmissible?"

Summary.

The PLM fills a gap occupied by none of these traditions: a full recursive, non-probabilistic architecture for autoregressive language generation grounded in ordinal possibility theory and evidence-driven geometric falsification, with formally grounded epistemic diagnostics and a proved Gaussian limit connecting it to standard LLMs.

4. Possibilistic Vocabulary Representation

4.1. The Vocabulary as a Hypothesis Space

Let $V = \{v_1, \ldots, v_{|V|}\}$ be a finite vocabulary of tokens. In the PLM, the vocabulary is treated as an epistemic hypothesis space.

Definition 1 (Vocabulary possibility distribution)**.** *At generation step k given context $t_{<k}$, the vocabulary possibility distribution is*

$$\pi_k : V \to [0,1], \quad \sup_{v \in V} \pi_k(v) = 1, \tag{9}$$

where $\pi_k(v)$ encodes the ordinal admissibility of v as the next token. The value is purely ordinal: it expresses relative plausibility under bounded epistemic constraints, not frequency, likelihood, or Bayesian degree of belief.

Definition 2 (Vocabulary support set)**.** *The active vocabulary support at step k is*

$$\mathcal{X}_k = \{v \in V : \pi_k(v) > 0\}, \tag{10}$$

 doi:10.20944/preprints202604.0920.v1

the set of tokens not yet falsified by the context. Tokens in $V \setminus \mathcal{X}_k$ are categorically excluded.

4.2. Embedding-Space Representation and the Geometry Assumption

Vocabulary tokens are represented via an embedding matrix $E \in \mathbb{R}^{|V| \times d}$, mapping each token v_j to a dense vector $e_j \in \mathbb{R}^d$. The PLM maintains a vocabulary support cloud

$$\Xi_k = \{e_j : v_j \in \mathcal{X}_k\} \subset \mathbb{R}^d. \tag{11}$$

Definition 3 (Vocabulary MVEE)**.** *The Minimum-Volume Enclosing Ellipsoid of the vocabulary support cloud,*

$$\Sigma_{\Xi_k} = \mathrm{MVEE}(\Xi_k), \tag{12}$$

defines a shape matrix $\Sigma_{\Xi_k} \succ 0$ characterizing the admissible spread of meaning at step k. Its determinant $\det(\Sigma_{\Xi_k})^{1/2}$ is the vocabulary epistemic volume $\mathcal{V}_k$.

Remark 1 (The embedding geometry assumption and its limitations)**.** *The PLM assumes that geometric proximity in the learned embedding space $\mathbb{R}^d$ is a meaningful proxy for semantic admissibility: tokens that are geometrically compatible with the context under the MVEE geometry are treated as epistemically admissible. This assumption deserves explicit scrutiny.*

Embedding geometry is itself learned and can exhibit pathologies—anisotropy, dimensional collapse, and nonlinear clustering that do not correspond to intuitive semantic structure (Ethayarajh, 2019). The claim that MVEE over token embeddings tracks semantic admissibility is a substantive empirical hypothesis, not a mathematical theorem. Several considerations support its plausibility: (a) embedding training objectives (e.g., masked language modeling, contrastive objectives) explicitly encourage semantically similar tokens to occupy nearby regions; (b) the PLM's falsification threshold is context-sensitive—it is computed relative to the whitened innovation geometry of the current query, not in an absolute embedding metric; and (c) approximation strategies (Smolyak sampling, hierarchical super-tokens) are applied in the embedding manifold, not in token-index space, and can leverage the manifold's known low-dimensional structure.

Nevertheless, whether MVEE geometry over learned embeddings faithfully tracks admissibility in practice is an open empirical question. We list it explicitly as Open Problem 6 in Section 13. The theoretical framework is stated over this geometry; empirical validation of the geometry assumption is a first-priority experimental obligation.

Remark 2 (Computational feasibility)**.** *For large vocabularies ($|V| \sim$ 50,000–200,000) and embedding dimensions $d \sim$ 768–4096, exact MVEE computation over the full vocabulary cloud at every generation step is computationally prohibitive. Three approximation strategies make the approach tractable: (1) Smolyak sparse sampling of $M \ll |\mathcal{X}_k|$ geometry-preserving representatives; (2) hierarchical super-tokens with offline vocabulary clustering and per-cluster refinement; (3) approximate MVEE via Johnson–Lindenstrauss projections to dimension $d' = O(\log |\mathcal{X}_k|)$. Formal coverage guarantees for these approximations under the VFI are listed as Open Problem 2 in Section 13.*

4.3. Possibilistic α-Cuts and Epistemic Volume

For the vocabulary cloud equipped with possibility values $\{\pi_k^{(j)}\}$, the α-cut at level $\alpha \in [0,1]$ is

$$\Xi_k^\alpha = \{e_j \in \Xi_k : \pi_k(v_j) \geq \alpha\}. \tag{13}$$

The α-cut volume $V_\alpha = \mathrm{Vol}(\mathrm{MVEE}(\Xi_k^\alpha))$ gives rise to the possibilistic entropy

$$H_{\pi_k} = \int_0^1 \log V_\alpha \, d\alpha, \tag{14}$$

the PLM's primary measure of epistemic uncertainty at generation step k. Unlike Shannon entropy, H_{π_k} has direct geometric interpretability: it is the log-volume integral of the nested α-cut ellipsoids.

 doi:10.20944/preprints202604.0920.v1

5. Epistemic Possibilistic Attention (EPA)

The standard scaled dot-product attention (Vaswani et al., 2017) computes query–key similarity via inner products and normalizes via softmax. This is the probabilistic closure we replace. From the tropical Hamilton–Jacobi foundation of Section 2.4, the correct replacement is not arbitrary: the update operator on the vocabulary impossibility field must be the Lax–Oleinik operator of the tropical Hamilton–Jacobi equation (Jah, 2026d, Theorem 5.5). EPA is that operator, discretized over the key cloud.

5.1. Whitened Innovation Geometry in Attention

Let $Q \in \mathbb{R}^{n_q \times d_k}$ be the query matrix, $K \in \mathbb{R}^{n_k \times d_k}$ the key matrix, and $V \in \mathbb{R}^{n_k \times d_v}$ the value matrix. Each key k_i is treated as a context hypothesis: a candidate semantic context that may or may not be admissible for the current query q.

Definition 4 (Attention innovation residual). *For query $q \in \mathbb{R}^{d_k}$ and key $k_i \in \mathbb{R}^{d_k}$, the attention innovation residual is*

$$r_i = q - k_i, \tag{15}$$

the discrepancy between the query's semantic target and the context hypothesis k_i.

Definition 5 (Attention innovation shape matrix). *Let $\Sigma_K = \mathrm{MVEE}(\{k_i\}_{i=1}^{n_k})$ be the MVEE of the key cloud, and let $\Sigma_Q \succ 0$ encode bounded query uncertainty. The attention innovation shape matrix is*

$$\Sigma_e = (S_K + S_Q)(S_K + S_Q)^\top, \quad S_K S_K^\top = \Sigma_K,\ S_Q S_Q^\top = \Sigma_Q, \tag{16}$$

a conservative Minkowski outer bound on the admissible innovation set.

5.2. Compatibility and Falsification

Let L_e be the Cholesky factor of Σ_e. The whitened attention innovation is

$$z_i = L_e^{-1} r_i, \quad q_i = \|z_i\|_2^2. \tag{17}$$

The attention geometric surprisal and attention compatibility are

$$S_i = \tfrac{1}{2} q_i, \quad \kappa_i = \exp(-S_i) \in (0, 1]. \tag{18}$$

Unlike softmax weights, κ_i is not a probability; it is an ordinal compatibility score encoding geometric consistency between key and query in whitened innovation space.

5.3. Possibilistic Attention Operator

Definition 6 (Epistemic Possibilistic Attention (EPA)). *Given a prior key possibility field $\pi_K^{(i)} \in [0, 1]$ initialized to 1 for all keys, the EPA operator computes:*

1. *Conjunctive possibility update: $\tilde{\pi}^{(i)} = \min\big(\pi_K^{(i)}, \kappa_i\big)$.*
2. *Max-rescaling for ordinal integrity: $\hat{\pi}^{(i)} = \tilde{\pi}^{(i)} / (\max_j \tilde{\pi}^{(j)} + \varepsilon)$.*
3. *Falsification gate: keys with $\hat{\pi}^{(i)} < \theta_{\min}$ are excluded ($\hat{\pi}^{(i)} \leftarrow 0$), where $\theta_{\min}$ ensures at least $N_{\min} = 2d_k + 1$ keys survive.*
4. *Output aggregation:*

$$\mathrm{EPA}(Q, K, V) = \hat{W} V, \quad \hat{W}_{ij} = \hat{\pi}^{(i)} \Big/ \sum_\ell \hat{\pi}^{(\ell)}. \tag{19}$$

Remark 3 (On row normalization and the probabilistic objection). *A natural objection is that the row normalization in step 4 above—dividing possibility weights by their sum to aggregate values—is structurally*

 doi:10.20944/preprints202604.0920.v1

identical to probabilistic normalization, and therefore the EPA operator has not genuinely escaped probabilistic semantics.

This objection conflates the mathematical operation of normalization with the epistemic semantics of the resulting weights. The distinction is as follows. In softmax attention, the weights are derived as $a_i = \exp(q \cdot k_i/\sqrt{d_k})/Z$*, where* Z *forces full distributional closure: every key receives positive mass, and the weights are* the probability distribution *from which the attention output is computed. Crucially, no key is ever falsified; the weights are positive for all keys regardless of admissibility.*

In EPA, normalization occurs after *categorical falsification. Keys with compatibility below* $\theta_{\min}$ *are assigned* $\hat{\pi}^{(i)} = 0$ *and excluded from the sum. The normalization that follows is therefore a value-aggregation device over the surviving admissible set—analogous to computing a weighted centroid of surviving hypotheses—not a statement that the weights constitute a probability distribution over all keys. The epistemic distinction is that the set of participating keys is itself determined by falsification, not by distributional weighting.*

Formally: let $\mathcal{S} \subseteq \{1, \ldots, n_k\}$ *denote the surviving key set after falsification. Then* $\mathrm{EPA}(Q, K, V) = \sum_{i \in \mathcal{S}} \hat{\pi}^{(i)} v_i / \sum_{j \in \mathcal{S}} \hat{\pi}^{(j)}$*. This is a minimax-medoid approximation over* $\mathcal{S}$*, not a probability-weighted sum over all keys. The denominator is a normalization for aggregation, not for distributional closure. When* $|\mathcal{S}| = n_k$ *(no falsification), EPA approaches standard softmax attention in the appropriate limit. When* $|\mathcal{S}| < n_k$*, EPA produces an output that a probabilistic attention mechanism cannot produce: it simply does not attend to falsified keys at all.*

Remark 4 (EPA as Lax–Oleinik operator)**.** *The EPA operator is the discretized Lax–Oleinik operator of the tropical Hamilton–Jacobi equation on vocabulary space. To see this: the Lax–Oleinik operator maps the prior impossibility field* $\Phi_\varnothing$ *forward under the action of the measurement Hamiltonian* Φ_S *via the pointwise tropical superposition* $\Phi_\varnothing^+ = \Phi_\varnothing \oplus \Phi_S = \max(\Phi_\varnothing, \Phi_S)$*. The EPA update implements precisely this: the conjunctive update step* $\tilde{\pi}^{(i)} = \min(\pi_K^{(i)}, \kappa_i)$ *is the possibility-space statement of the same identity (since* $-\log\min(a, b) = \max(-\log a, -\log b)$*). The falsification gate is the PCRB-admissible basin threshold: keys expelled by EPA are those outside the basin of the posterior impossibility field. The surviving keys are those inside the basin. Row-normalized aggregation over survivors is the discrete approximation to the geodesic attractor computation over the surviving well. The EPA operator is therefore not designed to mimic the ESPF; it is the ESPF's Lax–Oleinik recursion instantiated over a key cloud.*

Remark 5 (Comparison with softmax attention)**.** *Standard softmax attention assigns weights* $a_i = \exp(q \cdot k_i/\sqrt{d_k})/\sum_j \exp(q \cdot k_j/\sqrt{d_k})$*, which are always strictly positive and do not falsify any key. The EPA operator falsifies keys whose innovation geometry is inadmissible, preserving ordinal ordering rather than redistributing probability mass. When* $\Sigma_e \propto I$*, EPA recovers a Gaussian likelihood-weighted attention, approaching softmax in the high-temperature limit.*

5.4. Multi-Head Possibilistic Attention

Multi-head possibilistic attention applies H independent EPA heads with different projection matrices, then fuses compatibility fields conjunctively across heads:

$$\kappa_i^{\text{fused}} = \min_{h=1,\ldots,H} \kappa_i^{(h)}. \tag{20}$$

A token context hypothesis must be admissible across all semantic subspaces represented by the H heads in order to survive. This is a strictly stronger falsification criterion than per-head softmax averaging.

6. PLM Architecture

6.1. Overview

The PLM is a transformer-like encoder-decoder architecture in which all probabilistic operations are replaced by possibilistic ones. The architecture proceeds as follows:

 doi:10.20944/preprints202604.0920.v1

1. **Input encoding.** Token embeddings $e_k \in \mathbb{R}^d$ are computed as in a standard transformer, with positional encodings.
2. **Possibilistic encoder.** N_E layers of EPA with position-wise feed-forward networks compute a sequence of hidden representations $H = (h_1, \ldots, h_n)$.
3. **Possibilistic decoder.** N_D layers of masked self-EPA and cross-EPA compute decoder hidden states.
4. **Vocabulary falsification.** The final hidden state h_k is projected to the embedding space via $W_V \in \mathbb{R}^{d \times d}$, yielding a query vector $q_k = W_V h_k$ into the embedding matrix E.
5. **Possibilistic output head.** Compatibility scores between q_k and all vocabulary embeddings $\{e_j\}$ are computed via the EPA residual geometry, yielding $\pi_k(v_j) = \kappa_j$ after max-rescaling.
6. **Generation commitment.** The minimax medoid $\hat{v}_k = \arg\min_{v \in \mathcal{X}_k} \max_{v' \in \mathcal{X}_k} \|e_v - e_{v'}\|_{\Sigma_{\Xi_k}^{-1}}$ commits the filter to the surviving vocabulary hypothesis most geometrically central in admissible embedding space.

6.2. Spreader and Regeneration in Embedding Space

By analogy with the ESPF prediction step, the PLM requires a mechanism to regenerate a vocabulary support cloud Ξ_{k+1} for the next generation step given the survivor set $\mathcal{X}_k^{\text{surv}}$:

$$\Xi_{k+1} = \left\{ e_j + \sigma_k L_{\Xi_k} \zeta^{(\ell)} : v_j \in \mathcal{X}_k^{\text{surv}},\ \zeta^{(\ell)} \in \mathcal{A}(d, \ell) \right\} \cap E, \tag{21}$$

where $\mathcal{A}(d, \ell)$ is the Smolyak sparse grid at level ℓ, L_{Ξ_k} is the Cholesky factor of Σ_{Ξ_k}, and σ_k is a spread parameter.

Remark 6 (Limits of the ESPF analogy for regeneration)**.** *The regeneration rule above is constructed by direct analogy with the ESPF prediction step, in which support points evolve under known physical dynamics and are expanded under bounded perturbation. That analogy carries intuitive force in the language setting—the surviving vocabulary embeddings are perturbed in the direction of their geometric spread to sample nearby admissible tokens—but it rests on a substantive disanalogy that should be stated plainly.*

In the ESPF, support points are continuous state vectors evolving under a known dynamics function f: propagation is faithful to a physical model. In the PLM, tokens are discrete symbolic units embedded in a learned manifold with no known dynamics governing their sequential evolution. The perturbed vectors $e_j + \sigma_k L_{\Xi_k} \zeta^{(\ell)}$ are not the embeddings of physically predicted future states; they are geometry-preserving samples in the neighborhood of surviving tokens, retained only if they correspond to actual vocabulary tokens via nearest-neighbor projection.

This nearest-neighbor projection is clever and geometrically principled, but it is also where the ESPF analogy is most strained. Whether perturbing surviving token embeddings and projecting back to vocabulary space produces a meaningfully richer candidate set—versus simply recovering a subset of tokens already geometrically proximate to survivors—is an open empirical question. We flag this as Open Problem 5 in Section 13 and do not claim more for the regeneration step than that it is a principled geometric heuristic grounded in the ESPF architecture.

6.3. Epistemic Width Monitor

The aggregate epistemic width of the vocabulary at step k is

$$W_k = \int_V \left[\pi_k(v) - \nu_k(v) \right] d\mu(v), \tag{22}$$

 doi:10.20944/preprints202604.0920.v1

where $\nu_k(v) = 1 - \Pi_k(V \setminus \{v\})$ is the necessity of v. A large W_k signals broad uncertainty; $W_k \to 0$ signals epistemic collapse. A computationally tractable proxy based on NEES of the innovation sequence is:

$$\hat{W}_k = 1 - \exp(-\kappa_W \cdot \mathrm{NEES}_k), \quad \mathrm{NEES}_k = \frac{1}{T_w} \sum_{\tau=k-T_w}^{k} \left\| L_{e,\tau}^{-1}(q_\tau - \hat{e}_\tau) \right\|_2^2. \tag{23}$$

7. Training Objective

7.1. Possibilistic Cross-Entropy

We replace the standard cross-entropy loss with a possibilistic cross-entropy that penalizes the failure to include the ground-truth token in the admissible support:

$$\mathcal{L}_{\mathrm{PCE}} = -\frac{1}{N} \sum_{k=1}^{N} \log\big(1 - \exp(-\pi_k(t_k^*))\big), \tag{24}$$

where t_k^* is the ground-truth token at position k. This loss is zero when $\pi_k(t_k^*) = 1$ and increases as $\pi_k(t_k^*) \to 0$. Crucially, the loss does not penalize high possibility for other admissible tokens: epistemic breadth is preserved wherever the evidence justifies it.

Remark 7 (Epistemic posture of PCE). *The distinction from standard cross-entropy is fundamental. Standard cross-entropy $-\log P(t_k^* \mid t_{<k})$ drives the model to concentrate all probability mass onto the ground-truth token. The PCE loss instead asks only that the ground-truth token not be falsified: "I know this token is admissible; I do not claim it is the only admissible token."*

Remark 8 (Open questions for PCE training). *The PCE objective raises several training questions that are not resolved in the current work. First, gradient behavior: standard cross-entropy gradients are well-characterized; the gradient of $\log(1 - \exp(-\pi_k(t_k^*)))$ with respect to network parameters has not been analyzed in the language setting. Second, trivial high-possibility clouds: without sufficient regularization, the model may learn to set $\pi_k(v) > 0$ for all v at every step, trivially satisfying the PCE loss while providing no epistemic discrimination. The PCRB and entropy regularization terms are designed to counteract this, but their interaction in a gradient-based optimizer has not been validated. Third, the supervised learning signal: natural language corpora provide one observed continuation per context, not the full admissible set. Learning to assign high possibility to multiple admissible continuations from single-continuation supervision is a non-trivial generalization challenge. These questions are listed as Open Problem 5 in Section 13.*

7.2. PCRB Regularization

Training without additional regularization may drive the PLM to collapse the vocabulary support set $\mathcal{X}_k$ to a singleton at every step. The PCRB prevents this. Define the per-step vocabulary contraction ratio:

$$\kappa_k = \frac{|\mathcal{X}_k^{\mathrm{post}}|}{|\mathcal{X}_k^{\mathrm{prior}}|}. \tag{25}$$

The possibilistic information content of the k-th token is $I_k = 1 - \exp(-\bar{S}_k)$, $\bar{S}_k = \mathrm{median}_j S_k^{(j)}$, and the PCRB asserts $\kappa_k \geq (1 - I_k)^{1/2}$. We add a penalty:

$$\mathcal{L}_{\mathrm{PCRB}} = \frac{\lambda}{N} \sum_{k=1}^{N} \max\big(0,\ (1 - I_k)^{1/2} - \kappa_k\big)^2. \tag{26}$$

The justification of the PCRB as a universal floor is inherited from Theorem 5.2 of Jah (2026a) and is not independently proved here for the language setting.

 doi:10.20944/preprints202604.0920.v1

7.3. Possibilistic Entropy Regularization

A second regularization term encourages maintenance of non-trivial vocabulary support:

$$\mathcal{L}_{H_\pi} = -\frac{\gamma}{N} \sum_{k=1}^{N} H_{\pi_k}, \tag{27}$$

where $\gamma > 0$ penalizes low possibilistic entropy.

7.4. Total Loss

$$\mathcal{L}_{\text{PLM}} = \mathcal{L}_{\text{PCE}} + \mathcal{L}_{\text{PCRB}} + \mathcal{L}_{H_\pi}. \tag{28}$$

8. Volumetric Faithfulness Invariant for Token Support Sets

The VFI of Jah and Haslett (2025) governs well-posedness of the ESPF recursion. We extend it to the PLM's vocabulary generation loop.

Definition 7 (Vocabulary VFI)**.** *The PLM satisfies the Vocabulary VFI at generation step k if:*

(i) ***Geometric coverage.*** *Ξ_k contains vocabulary embeddings at multiple radial scales and in all principal directions under Σ_{Ξ_k}.*
(ii) ***Admissibility.*** *All $e_j \in \Xi_k$ correspond to tokens $v_j \in V$ with $\pi_k(v_j) > 0$.*
(iii) ***Anchor faithfulness.*** *The minimax medoid $\hat{e}_k \in \Xi_k$ is an actual surviving vocabulary embedding.*
(iv) ***Non-degeneracy.*** $\text{MVEE}(\Xi_k) \succ 0$.
(v) ***Survivor floor.*** $|\mathcal{X}_k^{\text{surv}}| \geq N_{\min} = 2d + 1$.

Theorem 1 (VFI ensures non-degenerate generation)**.** *Suppose the PLM maintains the Vocabulary VFI at every generation step $k \in \{1, \ldots, T\}$. Then:*

(a) The epistemic vocabulary volume satisfies $\mathcal{V}_k \geq \mathcal{V}_{\min} > 0$ for all k.
(b) The possibilistic entropy H_{π_k} remains finite and bounded from below.
(c) The minimax medoid commitment $\hat{v}_k$ corresponds to an actual vocabulary token at every step; no hallucinated or interpolated token is committed.

Proof sketch. The proof follows directly from Theorem A.1 of Jah and Haslett (2025) applied to the discrete embedding-space setting. Condition (i) provides geometric non-collapse via the Smolyak grid minimum inter-point spacing δ_ℓ; conditions (iii)–(iv) provide anchor faithfulness and MVEE non-degeneracy; condition (v) bounds H_{π_k} from below via the α-cut volume analysis of Jah (2026a), Lemma 3.1. The discrete vocabulary intersection in the regeneration equation preserves condition (ii) by construction. □

Remark 9 (Minimax medioid as geodesic attractor)**.** *Theorem 1(c) states that the minimax medoid commitment $\hat{v}_k$ corresponds to an actual vocabulary token. This is stronger than it first appears: Jah (2026d), Proposition 11.3, proves that the whitened minimax medioid is the* geodesic attractor *of the surviving well in the MVEE-whitened metric—the surviving support point that minimizes worst-case whitened distance to all other survivors, maximally insulated from the PCRB equipotential boundary in all directions. It is the last point to be falsified under any sequence of admissible contractions, and the natural reinitialization point for the next prediction step. The PLM's commitment rule is therefore not a heuristic selection from the surviving cloud: it is the fixed point of the Lax–Oleinik dynamics on the vocabulary manifold. The minimax medoid is to the PLM what the geoid center is to gravitational potential theory—the point of maximum epistemic protection within the admissible basin.*

This also clarifies the Remark 1 concern about embedding geometry. The geodesic attractor property holds in whatever metric the MVEE defines over the surviving cloud. If embedding geometry faithfully tracks semantic admissibility (the open empirical question), the geodesic attractor is semantically the most central surviving

 doi:10.20944/preprints202604.0920.v1

token. If embedding geometry is imperfect, the geodesic attractor is still geometrically optimal within the available metric—which is the best any commitment rule can do given the available information.

Remark 10 (Conditionality of Theorem 1)**.** *Theorem 1 is conditional on the VFI being maintained at every generation step k. We have not proved that the PLM's layers jointly maintain the VFI through $N_E + N_D$ depth—this is Open Problem 1 in Section 13. The force of the theorem is therefore of the form:* if *the architecture preserves the invariant,* then *these properties follow. This conditionality must be kept front and center when interpreting the anti-hallucination results below. In particular, the phrase "formally eliminated" for Class 1 hallucinations (below) should be read as "formally eliminated conditional on VFI maintenance and the embedding geometry assumption of Remark 1."*

Remark 11 (TEAG hallucination taxonomy)**.** *Theorem 1 and the TEAG framework collectively address a specific, formally characterizable subset of the LLM hallucination problem. We state precisely what is and is not claimed.*

Class 1: Probabilistic closure hallucinations—formally eliminated (conditional). *A standard LLM assigns strictly positive softmax probability to every token in V, including tokens that are geometrically incompatible with the current context embedding under any admissible innovation geometry. The PLM eliminates this failure mode* by construction and conditional on VFI maintenance and the embedding geometry assumption*: the committed token $\hat{v}_k$ is guaranteed to have survived the TEAG contraction and the VFI survivor floor. No token outside $\mathcal{X}_k$ can be committed. The precise guarantee is: under Theorem 1 and assuming embedding geometry faithfully tracks semantic admissibility (Remark 1), the PLM cannot commit a token that is geometrically inadmissible. It can still commit a token that is wrong—if the embedding geometry does not correctly capture admissibility, or if the VFI is not maintained.*

Class 2: Overconfident generation hallucinations—detected, not eliminated. *When ν_k is low, $\bar{S}_k$ is elevated, or $\Delta_k \gg 0$, the model's high-confidence generation is internally flagged as unsupported. These diagnostics do not prevent a wrong commitment, but they make overconfident hallucination visible in a way absent from any standard LLM.*

Class 3: Distributional shift hallucinations—detected geometrically. *Out-of-distribution context shifts the query embedding outside the admissible innovation ellipsoid, producing elevated whitened residuals and a sustained rise in $\bar{S}_k$.*

Outside scope. *Knowledge hallucinations (incorrect world knowledge in weights), multi-step reasoning hallucinations (failures above the token level), and underspecified context hallucinations (genuine ambiguity) lie outside the current framework.*

9. Epistemic Diagnostics for Generation

Definition 8 (Vocabulary necessity)**.** *The necessity of vocabulary support $\mathcal{X}_k$ at step k is*

$$\nu_k = N(\mathcal{X}_k) = 1 - \Pi(V \setminus \mathcal{X}_k) = 1 - \sup_{v \notin \mathcal{X}_k} \pi_k(v). \tag{29}$$

Definition 9 (Generation surprisal and epistemic tension)**.** *The generation surprisal at step k is the median whitened innovation of the surviving vocabulary embeddings:*

$$\bar{S}_k = \text{median}_{v \in \mathcal{X}_k} S_k(v). \tag{30}$$

The epistemic tension is $\Delta_k = \log(\bar{S}_k / S_{\text{ref}})$.

These diagnostics quantify: factual grounding (high ν_k), semantic strain (high $\bar{S}_k$), and epistemic balance ($|\Delta_k|$).

 doi:10.20944/preprints202604.0920.v1

9.1. Field-Theoretic Interpretation of Diagnostics

The diagnostics of this section have a precise interpretation in the impossibility field language of Section 2.4 and Jah (2026d). The vocabulary impossibility field at step k is $\Phi_{\emptyset,k}(v) = -\log \pi_k(v)$, and the vocabulary surprisal field is $\Phi_{S,k}(v) = \frac{1}{2}\|L_e^{-1}(q_k - e_v)\|^2$.

The *active deformation front* at step k is the tropical variety of the posterior impossibility field:

$$\mathcal{F}_k = \{v \in V : \Phi_{\emptyset,k}(v) = \Phi_{S,k}(v)\}, \tag{31}$$

the locus where prior impossibility and current surprisal are in exact balance. A token at the front is in epistemic equilibrium between its history and the current context; tokens inside the surviving well (below the PCRB basin threshold) survive; tokens outside are falsified. Generation surprisal $\bar{S}_k$ measures how far the surviving cloud sits from the active deformation front. Epistemic tension $\Delta_k \gg 0$ signals that the front is advancing rapidly through the vocabulary—high discrimination; $\Delta_k \ll 0$ signals that the context provides little falsifying information—the front has stalled. Necessity ν_k measures how deep the surviving well is: high necessity means the prior impossibility field strongly confines the admissible vocabulary; low necessity means many tokens remain near-equally admissible, the well is shallow, and commitment will be weakly supported.

The *confirmation bias index* of Jah (2026d), $B_k = \bar{S}_k/(L_k + \epsilon)$ where L_k is the Hausdorff deformation of the admissible vocabulary between consecutive steps, is a diagnostic available from the PLM's generation loop without additional computation. When $B_k \gg 1$, evidence is arriving (surprisal is nonzero) but the vocabulary support is not contracting commensurately—the generation is in inertial shielding, which the PLM can flag as a high-risk commitment step.

10. Spread Control and PCRB Enforcement

The spread parameter σ_k is controlled by:

$$\log \sigma_k = \log \sigma_{k-1} + \alpha_\sigma d_k, \tag{32}$$

with per-step rate limits $\sigma_{k-1} r^- \le \sigma_k \le \sigma_{k-1} r^+$, $r^+ = 1.15$, $r^- = 0.97$, and hard bounds $\sigma_{\min} \le \sigma_k \le \sigma_{\max}$.

Proposition 1. *The contraction rate limit $r^- = 0.97$ ensures that in any single generation step, the vocabulary cloud volume $\mathcal{V}_k$ cannot decrease by more than approximately $1 - (r^-)^d \approx 20\%$ per step (for $d = 7$), providing a per-step lower bound on κ_k independent of context geometry.*

11. Recovery of the Standard LLM in the Gaussian Limit

Theorem 2 (PLM Gaussian limit)**.** *Suppose:*

(G1) The vocabulary support cloud Ξ_k contracts to a Gaussian in $L^1(\mu)$: $\pi_k(v) \to \mathcal{N}(e_v; \hat{e}_k, \Sigma_k)$ as $W_k \to 0$.
(G2) The MVEE shape matrix $\Sigma_{\Xi_k} \to \Sigma_k$.
(G3) Token co-occurrence statistics are Gaussian with covariance $R_k \succ 0$.

Then: (a) $H_{\pi_k} \to \frac{1}{2}\log\det\Sigma_k + \text{const}(d)$, so minimizing H_{π_k} is asymptotically equivalent to minimizing $\log\det\Sigma_k$; (b) the EPA operator converges to scaled dot-product softmax attention; (c) the PCE loss converges to standard cross-entropy; (d) the PCRB penalty vanishes. Consequently, in the limit $W_k \to 0$, the PLM recovers a standard transformer LLM.

Remark 12 (Scope of the Gaussian limit)**.** *Conditions (G1)–(G3) are substantive assumptions that will not hold in general language domains. Gaussian contraction of the vocabulary support cloud requires unimodal, low-ambiguity contexts in which the admissible token set collapses to a well-defined neighborhood. Gaussian token co-occurrence statistics are known to fail for rare tokens, long-tail distributions, and syntactically complex constructions. The Gaussian limit theorem is therefore not a claim that PLM always recovers standard LLM behavior, nor that the limit is the natural operating regime for language. It is a theoretical bridge: it shows that*

 doi:10.20944/preprints202604.0920.v1

the PLM can *recover standard LLM behavior when the Gaussian assumptions happen to hold, establishing strict generalization rather than replacement.*

Conditions (G1)–(G3) may be chosen to be sufficient for the desired recovery result; they are not claimed to be necessary. We view the recovery as a theoretical coherence result grounding the PLM in the existing landscape of LLM architectures, not as evidence that PLM and standard LLM are interchangeable in practice.

Remark 13 (Gaussian limit as zero-temperature limit)**.** *The convergence of Theorem 1 has a precise thermodynamic interpretation established in Jah (2026d), Theorem 10.1. The Bayesian update corresponds to finite epistemic temperature $T > 0$, where the posterior impossibility field is a smooth log-sum-exp:*

$$\Phi^{+}_{\varnothing,T}(v) = -T \log\left(e^{-\Phi_{\varnothing}(v)/T} + e^{-\Phi_S(v)/T}\right). \tag{33}$$

As $T \to 0$, the log-sum-exp degenerates to the max: $\Phi^{+}_{\varnothing,T}(v) \xrightarrow{T\to 0} \max(\Phi_{\varnothing}(v), \Phi_S(v))$, recovering the possibilistic posterior. The PLM's Gaussian limit theorem (conditions G1–G3) is the language-domain instantiation of this zero-temperature limit: where Gaussian assumptions hold, the vocabulary cloud collapses to a unimodal distribution and the tropical dynamics converge to the probabilistic dynamics. The standard LLM is the $T \to T_{\text{Gauss}} > 0$ limit of the PLM; the PLM is the $T \to 0$ ground state of the same tropical Hamilton–Jacobi system. The two are not competing architectures—they are different operating temperatures of a single underlying framework.

Corollary 1 (Strict generalization)**.** *The PLM is a strict generalization of the standard transformer LLM: where Gaussian assumptions hold, the PLM recovers the optimal maximum-likelihood solution. Where Gaussian assumptions fail, the PLM remains the minimum-H_π optimal generator in its class while the standard LLM's optimality claims no longer apply.*

Remark 14 (Hölder mean structure and convergent optimality)**.** *The three principal generation optimality criteria emerge as evaluations of a single log-power-mean functional $\log M_p(\{V_\alpha\})$ at different Hölder orders:*

Order p	*Functional*	*Criterion*
$p \to +\infty$	$\log\det(\text{MVEE})$	*Popperian minimax (falsification)*
$p = 0$	$H_\pi = \int_0^1 \log V_\alpha \, d\alpha$	*PLM minimax-entropy optimality*
$p = 0$, *Gaussian*	$\frac{1}{2}\log\det\Sigma_k$	*Standard LLM cross-entropy*

Possibility theory and probability theory evaluate the same ignorance functional under different geometries of α-cuts. The standard LLM is the Gaussian specialization of the PLM's optimality criterion, not a separate invention. This is convergent optimality: two optimal solutions to categorically different problems that agree precisely when their domains of applicability coincide.

12. Multi-Modal and Multi-Sensor Extensions

For S modalities with associated encoders $f^{(s)}$ mapping to a shared embedding space $\mathbb{R}^d$, define modality-specific vocabulary compatibility:

$$\kappa_k^{(j,s)} = \exp\left(-\tfrac{1}{2}\left\|L_e^{(s)-1}(q_k - e_j)\right\|_2^2\right), \tag{34}$$

with fused compatibility computed conjunctively:

$$\kappa_k^{(j)} = \min_{s=1,\ldots,S} \kappa_k^{(j,s)}. \tag{35}$$

A token hypothesis must be simultaneously admissible under all modalities to survive. Cross-modal contradiction is exposed rather than averaged. The Choquet integral provides an optional diagnostic aggregation over modalities; it does not override conjunctive falsification.

 doi:10.20944/preprints202604.0920.v1

13. Limitations and Open Proof Obligations

The PLM architecture is presented as a theoretical framework and research blueprint. We identify the following open obligations, organized to make their impact on current claims clear:

Open 1: Vocabulary VFI maintenance under depth. Theorem 1 proves non-degenerate generation *assuming* the VFI is maintained. We have not proved that the PLM's EPA layers jointly maintain the VFI through $N_E + N_D$ depth. Until this is established, the Class 1 anti-hallucination guarantee is conditional. This is the most consequential open obligation.

Open 2: Scalability of vocabulary support geometry. For large vocabularies ($|V| \sim$ 50,000–100,000), exact MVEE computation is computationally prohibitive. Efficient approximations require formal coverage guarantees under the VFI.

Open 3: Formal PAC-possibilistic learnability. A formal learnability theory for PLM training under the PCE loss is not yet established.

Open 4: Reliability damping sufficiency. The formal proof that reliability damping prevents VFI violation when the survivor count is near-minimal carries over from ESPF Open 3 (Jah and Haslett, 2025) but has not been proved in the discrete vocabulary setting.

Open 5: Training stability and regeneration validity. The interaction between $\mathcal{L}_{\text{PCE}}$, $\mathcal{L}_{\text{PCRB}}$, and $\mathcal{L}_{H_\pi}$ in a gradient-based optimizer has not been validated. Gradient behavior of the PCE loss, sensitivity to the trivial-cloud failure mode, and the validity of the vocabulary regeneration step (Remark 6) are all open engineering and theoretical questions.

Open 6: Empirical validation of the embedding geometry assumption. As noted in Remark 1, whether MVEE geometry over learned token embeddings faithfully tracks semantic admissibility is an open empirical question. This is a first-priority experimental obligation before the PLM can be claimed to solve Class 1 hallucinations in a meaningful sense.

14. Discussion and Research Agenda

14.1. Adaptive Switching Between PLM and Standard LLM

The PLM need not operate in possibilistic mode at every generation step. When $\hat{W}_k < \bar{W}_{\text{crit}}$ (epistemic collapse approaching), the system transitions to standard LLM generation; when $\hat{W}_k \geq \bar{W}_{\text{crit}}$, the full PLM machinery activates:

$$\text{mode}_k = \begin{cases} \text{PLM} & \hat{W}_k \geq \bar{W}_{\text{on}}, \\ \text{LLM} & \hat{W}_k < \bar{W}_{\text{off}}, \\ \text{mode}_{k-1} & \text{otherwise.} \end{cases} \tag{36}$$

This switching is principled: Jah (2026b), Theorem 4.5, establishes that the possibilistic and probabilistic regimes are separated by the Choquet-to-Lebesgue convergence boundary, and $\hat{W}_k$ is a computable proxy for that boundary.

14.2. What TEAG Solves, Detects, and Does Not Address in the Hallucination Problem

The hallucination problem in LLMs is not a single phenomenon. The PLM makes precise, bounded contributions. As established in Section 8, these are: formal elimination (conditional) of Class 1 probabilistic closure hallucinations; formal detection of Class 2 overconfident and Class 3 distributional shift hallucinations; and explicit non-claims for knowledge, reasoning, and ambiguity hallucinations. A precisely bounded guarantee is more credible and more useful than an overbroad one, because it survives the counterexamples that defeat overbroad claims.

14.3. Connection to Kalman's Vision

The PLM inherits the ESPF's alignment with Kalman's call for prejudice-free modeling (Kalman, 1982). Standard LLMs embed the designer's prejudice that language uncertainty is stochastic, additive,

 doi:10.20944/preprints202604.0920.v1

and Gaussian in the limit. The PLM instead maintains only what the evidence supports and removes only what the evidence contradicts.

14.4. Research Agenda

1. **Empirical validation of embedding geometry.** Before larger architectural claims, establish whether compatibility scores derived from token embedding MVEE geometry correlate with human judgments of semantic admissibility. Even a small-scale study would significantly strengthen the foundation of Section 4.
2. **Efficient vocabulary geometry.** Develop $O(|V| \log |V|)$ approximations to MVEE computation for large vocabularies.
3. **Pre-training from scratch.** Train a small PLM ($\sim$125M parameters) on a standard benchmark corpus and compare epistemic diagnostics against a comparably sized standard LLM.
4. **Fine-tuning from probabilistic checkpoints.** Develop a curriculum for converting a pretrained probabilistic LLM into a PLM by gradually introducing the EPA operator and PCRB regularization.
5. **Hallucination detection benchmarks.** Evaluate whether ν_k, $\bar{S}_k$, and Δ_k correlate with factual errors on knowledge-intensive tasks (TriviaQA, NaturalQuestions, HaluEval).
6. **Adaptive switching calibration.** Systematic benchmarking of $\bar{W}_{\text{crit}}$, $\bar{W}_{\text{on}}$, $\bar{W}_{\text{off}}$, κ_W, and T_w across language tasks and model scales.
7. **GaiaVerse integration.** Deploy the PLM as the epistemic inference engine within the GaiaVerse planetary stewardship knowledge graph.

15. Conclusion

We have proposed the Possibilistic Language Model (PLM), a transformer-like sequence generation architecture grounded in possibility theory and the Epistemic Support-Point Filter. The PLM replaces probabilistic token distributions with possibilistic compatibility fields; replaces softmax attention with Epistemic Possibilistic Attention; replaces maximum-likelihood training with a possibilistic entropy objective regularized by the PCRB; and extends the Volumetric Faithfulness Invariant to the vocabulary generation setting.

We have proved that the PLM is a strict generalization of the standard transformer LLM, recovering it in the Gaussian epistemic limit under stated conditions. We have identified formal anti-hallucination properties conditional on VFI maintenance, and described epistemic diagnostics providing interpretable real-time signals of generation confidence and model strain.

Throughout, we have been explicit about the epistemic status of each contribution: what is defined, what is derived from prior TEAG works, what is proved under stated assumptions, and what remains open. This specificity is not a weakness of the program; it is its epistemic posture applied to itself.

The PLM is not offered as a drop-in replacement for existing LLMs. It is a distinct inferential framework with different assumptions, strengths, and failure modes. Its primary claim is philosophical as much as technical: language generation under epistemic uncertainty does not require probabilistic closure. A disciplined, Popperian, falsification-driven architecture can be made theoretically principled, computationally tractable, and empirically grounded. The first and most consequential next step is empirical: to validate whether token embedding geometry faithfully tracks semantic admissibility. The path from theory to implementation is open; this paper lays its first stones.

Acknowledgments: The development of the Possibilistic Language Model builds on the theoretical foundations laid by the Epistemic Support-Point Filter. I am grateful to Van Haslett for his contribution to the ESPF architecture that makes this extension possible. I am also grateful to the broader community whose work on possibility theory, non-Bayesian filtering, and language model uncertainty has shaped the intellectual landscape within which this work sits.

 doi:10.20944/preprints202604.0920.v1

References

Hokamp, C. and Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1535–1546.

Angelopoulos, A. N. and Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*.

Benferhat, S., Dubois, D., Garcia, L., and Prade, H. (2002). On the transformation between possibilistic logic bases and possibilistic causal networks. *International Journal of Approximate Reasoning*, 29(2):135–173.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*, pages 4171–4186.

Dubois, D. and Prade, H. (1988). *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press.

Dubois, D. and Prade, H. (2000). Possibility theory in information fusion. *International Journal of Intelligent Systems*, 15(7):621–640.

Dubois, D., Prade, H., and Sabbadin, R. (1997). A possibilistic logic machinery for qualitative decision. In *Proceedings of the AAAI Spring Symposium*, pages 47–54.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *Proceedings of EMNLP*, pages 55–65.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv:1805.04833*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of ICML*, pages 1050–1059.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of ICML*, pages 1321–1330.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. *Proceedings of ICLR*.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2017). Toward controlled generation of text. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1587–1596.

Jah, M. K. and Haslett, V. (2025). The Epistemic Support-Point Filter (ESPF): A bounded possibilistic framework for ordinal state estimation. *arXiv:2508.20806*.

Jah, M. K. (2026a). The Epistemic Support-Point Filter: Jaynesian maximum entropy meets Popperian falsification. A possibilistic minimax-entropy optimality proof. *arXiv:2603.10065*.

Jah, M. K. (2026b). The geometry of knowing: From possibilistic ignorance to probabilistic certainty. A measure-theoretic framework for epistemic convergence. *arXiv:submit/7362363*.

Jah, M. K. (2026c). Theory of Epistemic Abductive Geometry (TEAG): A unified theory of admissibility-driven inference across dynamical systems, measure theory, and language. Manuscript in preparation.

Jah, M. K. (2026d). The Epistemic Support-Point Filter as a Tropical Hamilton–Jacobi System: Wavefront Propagation and Possibilistic Inference. Preprints.org, version 2, posted 31 March 2026. DOI: 10.20944/preprints202603.2110.v2.

Kalman, R. E. (1982). System identification from noisy data. In A. Bednarek and L. Cesari (eds.), *Dynamic Systems II: A University of Florida International Symposium*, pages 135–164. Academic Press, New York.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32.

Malinin, A. and Gales, M. (2021). Uncertainty estimation in autoregressive structured prediction. *Proceedings of ICLR*.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv:2002.06177*.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *Proceedings of NAACL*.

Senséoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.

 doi:10.20944/preprints202604.0920.v1

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.
Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28.