ATTENTION AS A MARKOV RANDOM FIELD

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting when large language models (LLMs) hallucinate remains an open challenge. Existing uncertainty estimators either require costly multi-sample generation or supervised calibration data, which limits their practical deployment. We introduce the **Attention Bethe Score (ABS)**, a zero-shot, single-pass uncertainty metric grounded in variational inference over probabilistic graphical models. Our key insight is to reinterpret the attention mechanism of a frozen Transformer as the structure of a Markov Random Field (MRF): token positions become nodes, attention weights define edges with attention-modulated Potts couplings, and hidden-state projections yield unary potentials. Running loopy belief propagation on this graph yields the Bethe free energy, a principled measure of internal representational *frustration*, that is, the degree to which strongly attended positions hold inconsistent beliefs. We prove that, under a contraction condition satisfied in typical operating regimes, the Bethe free energy monotonically tracks projected frustration, and we validate this relationship empirically (Spearman $\rho = 0.89 \pm 0.02$). Across the five core benchmarks of Table 1, spanning open-domain QA, fact verification, hallucination detection, and summarization, ABS achieves aggregate AUROC of 0.788 (LLaMA-3.1-8B) and 0.806 (Qwen-2.5-7B), surpassing all nine baselines, including multi-sample Semantic Entropy and supervised Internal Probes, while adding fewer than four milliseconds and under three percent wall-clock overhead to a standard forward pass.

1 INTRODUCTION

Large language models (LLMs) produce remarkably fluent text (Touvron et al., 2023), yet they routinely generate confident but factually incorrect statements, i.e., *hallucinations* (Ji et al., 2023), which makes **selective prediction** (Geifman & El-Yaniv, 2017) essential for high-stakes deployment. Existing uncertainty estimators face a trilemma. *Output-probability methods* (Kadavath et al., 2022) conflate linguistic entropy with epistemic uncertainty and discriminate poorly on open-ended generation. *Multi-sample methods* (Kuhn et al., 2023; Manakul et al., 2023) discriminate well by measuring inter-sample consistency, but they inflate inference cost by roughly an order of magnitude. *Internal-state probes* (Azaria & Mitchell, 2023; Burns et al., 2023; Deng et al., 2024; Chen et al., 2024) operate in a single pass, yet they require supervised calibration data and treat token representations either independently or through global statistics, discarding the *pairwise dependency structure* that the attention mechanism already encodes.

Attention weights provide precisely that relational signal: they specify *which token representations the model itself treats as mutually relevant*. When a model hallucinates, it attends strongly to positions whose hidden states point in contradictory directions, a pattern we call **representational frustration**. Frustration is inherently relational and graph-structured, so scalar aggregations of attention such as the mean, maximum, or entropy collapse it. No existing uncertainty method models the *joint* distribution over token-level states conditioned on this attention-induced topology.

We therefore propose the **Attention Bethe Score (ABS)**, which reinterprets attention as a Markov Random Field (MRF). At each layer of a frozen Transformer, token positions become nodes whose unary potentials are derived from randomly projected hidden states, above-threshold attention weights become edges carrying attention-modulated Potts couplings that penalize representational disagreement, and loopy belief propagation (Pearl, 1982; Yedidia et al., 2005) yields approximate beliefs from which the **Bethe free energy** quantifies structural frustration (Figure 1; Algorithm 1). ABS is **zero-shot** (no learned parameters or calibration labels), **single-pass** (below 3% wall-clock over-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

head), and **theoretically motivated**: under a contraction condition on the BP operator that holds in our operating regime, the Bethe free energy monotonically tracks the projected frustration index (Proposition 2), and we observe a strong empirical rank correlation (Spearman $\rho = 0.89 \pm 0.02$). On the five core benchmarks of Table 1, ABS outperforms all nine baselines, including multi-sample Semantic Entropy (5–20$\times$ cost) and supervised Internal Probes, and reaches aggregate AUROC of 0.788 (LLaMA-3.1-8B) and 0.806 (Qwen-2.5-7B).

Our contributions are threefold. (1) We formalize attention-induced MRFs with attention-modulated Potts couplings and show how to construct them from a single forward pass of a frozen Transformer (§3.3). (2) We prove a perturbation bound linking the Bethe free energy to projected frustration under weak coupling, and we validate it empirically (§3.7). (3) We demonstrate state-of-the-art failure prediction across five core benchmarks and two model families against nine baselines, together with subtractive ablations, sensitivity sweeps, cross-distribution transfer, and failure-mode stratification (§4.2). The MRF machinery matters: replacing the Bethe free energy with the raw frustration index costs 5.3 AUROC points (Table 2).

2 Related Work

Output-probability and multi-sample methods. Classical uncertainty estimation computes scalar statistics from output distributions: predictive entropy (Malinin & Gales, 2018), length-normalized variants (Xiao & Wang, 2022), P(True) (Kadavath et al., 2022), and verbalized chain-of-thought self-assessment. These conflate aleatoric variation with epistemic uncertainty (Gal, 2016) and perform poorly on open-ended generation, where linguistic entropy is inherently high. Multi-sample methods, Semantic Entropy (Kuhn et al., 2023) and SelfCheckGPT (Manakul et al., 2023), discriminate well by measuring inter-sample consistency, but they require 5–20$\times$ inference cost, whereas ABS exceeds their AUROC on our benchmarks in a single pass (Table 1).

Internal-state methods. Linear probes over hidden states (Azaria & Mitchell, 2023; Burns et al., 2023) achieve strong in-distribution performance but collapse out of distribution (§4.4), as they learn domain-specific stylistic cues rather than generalizable epistemic signals. TokenSAR (Deng et al., 2024) and INSIDE (Chen et al., 2024) leverage attention-weighted anomaly scores and spectral covariance, while inference-time intervention (Li et al., 2024) shifts activations along probed directions. These methods either require supervised labels, process tokens independently, or ignore the relational structure of attention, whereas ABS models the *joint* distribution over token states through a graphical model derived from attention, and requires neither labels nor retraining.

Graphical models and attention. The Bethe free energy and loopy BP are well studied (Yedidia et al., 2005; Wainwright & Jordan, 2008; Koller & Friedman, 2009), with convergence guarantees under weak coupling (Ihler et al., 2005; Mooij & Kappen, 2007); Potts couplings, which penalize disagreement between neighboring variables, are classical tools in physics and vision. Connections between attention and probabilistic inference have been noted (Vaswani et al., 2017), but no prior work constructs an explicit graphical model over attention-induced graphs and uses its Bethe free energy as an uncertainty score for generation.

Selective prediction and calibration. Selective prediction (Geifman & El-Yaniv, 2017) lets a model abstain on uncertain inputs, and calibration methods such as Platt scaling (Guo et al., 2017) post-process confidence scores to align them with empirical accuracy. ABS is complementary: it supplies a raw score that can be Platt-scaled and thresholded at a user-specified abstention rate without labels.

3 Methodology

3.1 Notation and Preliminaries

Transformer notation. Consider a pre-trained autoregressive Transformer with L layers and H heads. Let $\mathbf{X}^{(\ell)} \in \mathbb{R}^{N \times d}$ denote the hidden representations at layer ℓ, where N is the full sequence length (prompt plus response) and d the hidden dimension. The attention weight matrix for head

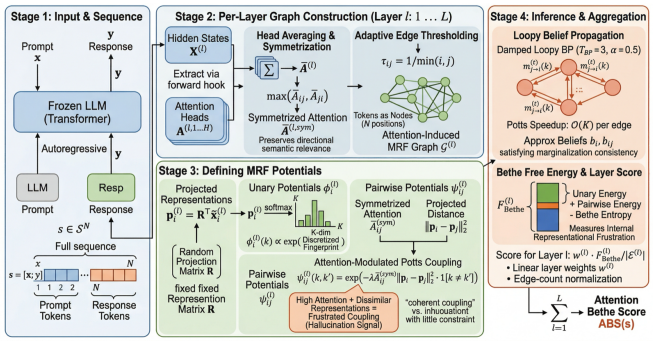


Figure 1: Overview of the Attention Bethe Score (ABS). At each layer, head-averaged attention defines an undirected graph over token positions, with an edge retained only above a position-adaptive uniform baseline. Hidden states are projected to unary potentials, and each retained edge carries an attention-modulated Potts coupling that penalizes representational disagreement. Damped loopy belief propagation yields beliefs whose Bethe free energy is weighted across layers into a single zero-shot uncertainty score. No parameters are learned and the model is never modified.

h is $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{N\times N}$ (lower-triangular for causal models). We write $\bar{\mathbf{A}}^{(\ell)} = \frac{1}{H}\sum_h \mathbf{A}^{(\ell,h)}$ for the **head-averaged attention matrix**.

MRF preliminaries. An MRF over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with discrete variables $\{z_i \in \{1,\dots,K\}\}_{i\in\mathcal{V}}$ is defined by unary potentials $\phi_i(z_i)$ and pairwise potentials $\psi_{ij}(z_i, z_j)$, with joint $p(\mathbf{z}) \propto \prod_i \phi_i(z_i) \prod_{(i,j)} \psi_{ij}(z_i, z_j)$. The **Bethe free energy** (Yedidia et al., 2005) approximates $-\log Z$ via beliefs $\{b_i, b_{ij}\}$ satisfying:

$$\sum_{k'} b_{ij}(k,k') = b_i(k), \quad \sum_{k} b_{ij}(k,k') = b_j(k'), \quad \forall (i,j) \in \mathcal{E}. \tag{1}$$

The Bethe free energy is $F_{\text{Bethe}} = U_{\text{Bethe}} - \mathcal{H}_{\text{Bethe}}$, with terms defined in §3.5. We fix a random projection $\mathbf{R} \in \mathbb{R}^{d\times K}$ drawn once from $\mathcal{N}(0, I_d/K)$ and define the **projected representation** $\mathbf{p}_i^{(\ell)} = \mathbf{R}^\top \tilde{\mathbf{x}}_i^{(\ell)} \in \mathbb{R}^K$, where $\tilde{\mathbf{x}}_i^{(\ell)} = \text{LayerNorm}(\mathbf{x}_i^{(\ell)})$.

3.2 Problem Formulation

Let f_θ be a frozen pre-trained LLM. Given a prompt $\mathbf{x} = (x_1, \dots, x_T)$, the model produces a response $\mathbf{y} = (y_1, \dots, y_M)$ autoregressively. Define the full sequence $\mathbf{s} = [\mathbf{x}; \mathbf{y}]$ of length $N = T + M$.

Objective. We seek a scalar uncertainty function $U : \mathcal{S} \to \mathbb{R}_{\geq 0}$ that, given only a single forward pass through f_θ on $\mathbf{s}$, satisfies the **selective prediction desideratum**:

$$\mathbb{E}\big[\mathbf{1}[\hat{y} \neq y^*] \mid U(\mathbf{s}) \geq u\big] \quad \text{is monotonically non-decreasing in } u, \tag{2}$$

where $\hat{y}$ is the model's predicted answer and y^* is the ground truth. We require U to be (i) **zero-shot** (no calibration labels), (ii) **sampling-free** (single forward pass), and (iii) **theoretically grounded** in variational inference principles. We do not claim that any practical score satisfies (2) exactly: it states an ideal, and we evaluate how closely a score approximates it via AUROC for failure prediction.

3.3 Attention-Induced MRF: Graph Structure

At each layer ℓ, we construct an undirected graph $\mathcal{G}^{(\ell)} = (\mathcal{V}, \mathcal{E}^{(\ell)})$ where $\mathcal{V} = \{1,\dots,N\}$ corresponds to token positions.

Symmetrization. Since $\bar{\mathbf{A}}^{(\ell)}$ is lower-triangular in causal Transformers, we symmetrize the attention matrix:

$$\bar{A}_{ij}^{(\ell,\text{sym})} = \max\big(\bar{A}_{ij}^{(\ell)},\ \bar{A}_{ji}^{(\ell)}\big). \tag{3}$$

We use $\max$ rather than averaging because a single strong directional link $(i \to j)$ indicates meaningful coupling even without reciprocal attention, whereas averaging would dilute such asymmetric but semantically important connections. Henceforth $\bar{A}_{ij}^{(\ell)}$ denotes the symmetrized version.

Position-adaptive thresholding. An edge is retained when the symmetrized attention exceeds a position-adaptive threshold:

$$(i,j) \in \mathcal{E}^{(\ell)} \iff \bar{A}_{ij}^{(\ell)} > \tau_{ij}, \qquad \tau_{ij} = 1/\min(i,j). \tag{4}$$

This retains edges only when attention exceeds the uniform baseline of the more position-constrained token, because position $\min(i,j)$ distributes its attention mass over at most $\min(i,j)$ positions. The resulting graphs are sparse, with average degree $\bar{d} \approx 20$ at $N = 2048$, which keeps loopy BP efficient. Alternative thresholds $\{1/\max(i,j), 1/\sqrt{ij}\}$ change aggregate AUROC by at most 0.3 points (Table 4). Let $d_i^{(\ell)} = |\{j : (i,j) \in \mathcal{E}^{(\ell)}\}|$ denote the degree of node i.

3.4 Potentials and Belief Propagation

We define the MRF potentials and obtain beliefs via loopy belief propagation (loopy BP), cleanly separating the model definition (potentials) from the approximate inference procedure (beliefs).

3.4.1 Unary Potentials

We define the **unary potential** at each token position via projection into the K-dimensional space:

$$\phi_i^{(\ell)}(k) = \frac{\exp\big(p_{i,k}^{(\ell)}\big)}{\sum_{k'=1}^{K} \exp\big(p_{i,k'}^{(\ell)}\big)}, \quad k \in \{1,\dots,K\}, \tag{5}$$

where $p_{i,k}^{(\ell)} = \mathbf{r}_k^\top \tilde{\mathbf{x}}_i^{(\ell)}$ is the k-th component of the projected representation $\mathbf{p}_i^{(\ell)} = \mathbf{R}^\top \tilde{\mathbf{x}}_i^{(\ell)}$. This gives $\boldsymbol{\phi}_i^{(\ell)} \in \Delta^{K-1}$.

Motivation for random projection. The K-dimensional discretization avoids learned parameters. The formal Johnson–Lindenstrauss bound is vacuous at $K = 64$ ($\epsilon \approx 0.8$); however, the softmax in (5) compresses distances into a simplex, and we require only approximate *rank* preservation of representation similarities, which we confirm empirically (Spearman $\rho > 0.92$ across 10 random projections; Appendix D). The distribution $\boldsymbol{\phi}_i^{(\ell)} \in \Delta^{K-1}$ therefore serves as a discretized fingerprint of $\tilde{\mathbf{x}}_i^{(\ell)}$.

3.4.2 Pairwise Potentials

For each edge $(i,j) \in \mathcal{E}^{(\ell)}$, we define a pairwise potential in the *same* projected space as the unary potentials, which keeps the discrete MRF self-consistent:

$$\psi_{ij}^{(\ell)}(k,k') = \exp\Big(-\lambda\, \bar{A}_{ij}^{(\ell)} \cdot \|\mathbf{p}_i^{(\ell)} - \mathbf{p}_j^{(\ell)}\|_2^2 \cdot \mathbf{1}[k \neq k']\Big), \tag{6}$$

where $\lambda > 0$ is a coupling strength. This implements an **attention-modulated Potts model**: state disagreement ($k \neq k'$) is penalized proportionally to the attention weight $\bar{A}_{ij}^{(\ell)}$ and the projected distance $\|\mathbf{p}_i^{(\ell)} - \mathbf{p}_j^{(\ell)}\|_2^2$. Using projected rather than original-space distances ensures consistency with the discrete states: $\boldsymbol{\phi}_i \approx \boldsymbol{\phi}_j$ implies a small $\|\mathbf{p}_i - \mathbf{p}_j\|_2^2$ and hence a weak Potts penalty. The hallucination signal then arises from *frustrated* couplings, which are couplings that pair high attention with representationally inconsistent positions, whereas coherent couplings ($\psi_{ij} \approx 1$) and low-attention edges contribute little.

3.4.3 Loopy Belief Propagation

We run T_{BP} iterations of damped loopy BP (Yedidia et al., 2005) to obtain beliefs that approximately satisfy the marginalization constraints (1).

Initialization. Messages are initialized uniformly: $m_{j\to i}^{(0)}(k) = 1/K$ for all edges and states.

Message passing. At iteration t, messages from node j to node i are:

$$m_{j\to i}^{(t)}(k) \propto \sum_{k'=1}^{K} \psi_{ij}^{(\ell)}(k,k') \cdot \phi_j^{(\ell)}(k') \cdot \prod_{\substack{n\in\mathcal{N}(j)\\ n\neq i}} m_{n\to j}^{(t-1)}(k'), \tag{7}$$

where $\mathcal{N}(j)$ denotes the neighbors of j in $\mathcal{G}^{(\ell)}$. Messages are normalized to sum to 1, then damped:

$$m_{j\to i}^{(t)} \leftarrow \alpha\, m_{j\to i}^{(t)\text{new}} + (1-\alpha)\, m_{j\to i}^{(t-1)}, \tag{8}$$

with $\alpha = 0.5$.

Belief computation. After T_{BP} iterations:

$$b_i^{(\ell)}(k) \propto \phi_i^{(\ell)}(k) \cdot \prod_{j\in\mathcal{N}(i)} m_{j\to i}^{(T_{\text{BP}})}(k), \tag{9}$$

$$b_{ij}^{(\ell)}(k,k') \propto \psi_{ij}^{(\ell)}(k,k') \cdot \phi_i^{(\ell)}(k) \cdot \phi_j^{(\ell)}(k') \cdot \prod_{\substack{n\in\mathcal{N}(i)\\ n\neq j}} m_{n\to i}^{(T_{\text{BP}})}(k) \cdot \prod_{\substack{n\in\mathcal{N}(j)\\ n\neq i}} m_{n\to j}^{(T_{\text{BP}})}(k'), \tag{10}$$

normalized so that $\sum_k b_i^{(\ell)}(k) = 1$ and $\sum_{k,k'} b_{ij}^{(\ell)}(k,k') = 1$.

At any fixed point of loopy BP, the beliefs (9)–(10) satisfy the marginalization constraints (1) (Proposition 1; proof in Appendix B).

3.5 Bethe Free Energy Computation

Given potentials $\{\phi_i^{(\ell)}, \psi_{ij}^{(\ell)}\}$ and beliefs $\{b_i^{(\ell)}, b_{ij}^{(\ell)}\}$ satisfying (1), we compute the **Bethe free energy**:

$$F_{\text{Bethe}}^{(\ell)} = U_{\text{Bethe}}^{(\ell)} - \mathcal{H}_{\text{Bethe}}^{(\ell)}, \tag{11}$$

where the **average energy** is

$$U_{\text{Bethe}}^{(\ell)} = \underbrace{-\sum_{i\in\mathcal{V}}\sum_{k} b_i^{(\ell)}(k)\log\phi_i^{(\ell)}(k)}_{U_{\text{unary}}^{(\ell)}} \underbrace{-\sum_{(i,j)\in\mathcal{E}^{(\ell)}}\sum_{k,k'} b_{ij}^{(\ell)}(k,k')\log\psi_{ij}^{(\ell)}(k,k')}_{U_{\text{pair}}^{(\ell)}}, \tag{12}$$

and the **Bethe entropy** is

$$\mathcal{H}_{\text{Bethe}}^{(\ell)} = \sum_{(i,j)\in\mathcal{E}^{(\ell)}} H[b_{ij}^{(\ell)}] - \sum_{i\in\mathcal{V}} (d_i^{(\ell)} - 1)\, H[b_i^{(\ell)}], \tag{13}$$

with $H[b_{ij}] = -\sum_{k,k'} b_{ij}(k,k')\log b_{ij}(k,k')$ and $H[b_i] = -\sum_k b_i(k)\log b_i(k)$.

3.6 Attention Bethe Score (ABS)

We aggregate per-layer Bethe free energies into the **Attention Bethe Score**:

$$\text{ABS}(\mathbf{s}) = \sum_{\ell=1}^{L} w^{(\ell)} \cdot \frac{F_{\text{Bethe}}^{(\ell)}}{|\mathcal{E}^{(\ell)}|}, \tag{14}$$

where $w^{(\ell)} = \frac{2\ell}{L(L+1)}$ are linearly increasing weights satisfying $\sum_{\ell=1}^{L} w^{(\ell)} = 1$. Linear weighting reflects the empirical finding that later layers carry more semantically meaningful, task-relevant information (Tenney et al., 2019). Table 4 confirms that linear weighting attains a higher aggregate AUROC than uniform weighting (0.788 against 0.778) and remains within 0.4 points of entropy-adaptive weighting (0.792). The edge-count normalization $|\mathcal{E}^{(\ell)}|$ makes layers with different sparsity comparable.

Interpretation. When the model is confidently correct, strongly attended positions have similar projected representations, so beliefs stay close to the unary potentials and F_{Bethe} is low. When the model hallucinates, high attention to representationally inconsistent positions creates frustrated couplings, beliefs deviate from the unary potentials, and F_{Bethe} rises.

3.7 Theoretical Grounding

Projected frustration index. The **projected frustration index** $\mathcal{F}^{(\ell)} = \sum_{(i,j)\in\mathcal{E}^{(\ell)}} \bar{A}_{ij}^{(\ell)} \cdot \|\mathbf{p}_i^{(\ell)} - \mathbf{p}_j^{(\ell)}\|_2^2$ (Definition 1) aggregates attention-weighted disagreements. Under attention–representation consistency (Assumption 1), $\mathcal{F}^{(\ell)}$ stays bounded for correct responses and is elevated for hallucinations.

The Bethe free energy tracks frustration. In the weak-coupling regime (Assumption 2, $\kappa < 1$), Proposition 2 gives:

$$F_{Bethe}(\Theta_2, b_2^{(T)}) - F_{Bethe}(\Theta_1, b_1^{(T)}) \geq \lambda\rho_{\min} \cdot \delta_F - \mathcal{O}\Big(|\mathcal{E}|K^2\kappa^T(\delta_F + \delta_u) + NKL_\sigma\delta_u\Big), \quad (15)$$

where δ_F is the frustration increase and $\rho_{\min}$ the minimum off-diagonal belief mass. At our operating point ($\kappa \approx 1.5 \times 10^{-4}$, $T_{\text{BP}} = 3$), the leading term dominates. Empirically, $F_{\text{Bethe}}^{(\ell)}$ is rank-correlated with $\mathcal{F}^{(\ell)}$ (Spearman $\rho = 0.89 \pm 0.02$ across all layers, 6 models, and 10 random projections of $\mathbf{R}$), most strongly in the deeper layers. Proposition 3 further ensures that close representations yield close potentials (the formal bound is vacuous at $K{=}64$, but the empirical rank correlation exceeds 0.92). Combining these results, Corollary 1 predicts low ABS for correct responses and high ABS for hallucinations, which we confirm across all benchmarks (§4.2). Full proofs appear in Appendix B.

3.8 Implementation, Inference, and Scope

ABS is implemented with forward hooks and adds $\approx 3\%$ wall-clock overhead (≈ 23 MB peak additional memory at 2,048 tokens) to a standard forward pass; the Potts structure reduces the per-edge BP cost from $\mathcal{O}(K^2)$ to $\mathcal{O}(K)$. All four hyperparameters (K, λ, T_{BP}, α) are set without calibration labels. The method is purely inference-time, with no training or fine-tuning: layers are processed sequentially, and the complete pseudocode is in Algorithm 1 (Appendix A). For selective prediction we abstain when $\text{ABS}(\mathbf{s}) > \eta$, where η is the $(1{-}r)$-quantile of the development-set ABS distribution and r is a user-specified abstention rate (no labels required).

Limitations and distinctions from prior work. ABS underestimates uncertainty for memorized incorrect facts and for systematic reasoning errors that are internally coherent but wrong, and it overestimates uncertainty for genuinely ambiguous queries and for rare-but-correct knowledge; these failure modes are inherent to introspective methods, so ABS complements rather than replaces external verification (Appendix E). Unlike temperature or Platt scaling (Guo et al., 2017), ABS is fully zero-shot; unlike Semantic Entropy (Kuhn et al., 2023), it requires a single pass; and unlike internal probes (Azaria & Mitchell, 2023; Burns et al., 2023), it uses no learned parameters. Its key contribution is the *graphical-model reinterpretation* of attention as MRF structure, with the Bethe free energy as a principled uncertainty measure. Full implementation details and comparisons are in Appendix E.

4 Results

4.1 Experimental Setup

We evaluate ABS on eight datasets: TriviaQA, NQ-Open, FEVER, HaluEval-QA, and XSum form the core suite of Table 1 and the basis of every aggregate statistic we report, TruthfulQA is used for the distribution-shift analysis of Table 5, and SciQ and Bio-Gen are stress tests described in Appendix C that do not enter the aggregate. All experiments use LLaMA-3.1-8B and Qwen-2.5-7B with greedy decoding. We report AUROC and AUPRC as threshold-free discrimination metrics, ECE (15 equal-width bins, after Platt scaling) for calibration quality, and Cohen's d with paired bootstrap p-values (10,000 resamples) for statistical significance, and we combine metrics across the five core

Table 1: **Failure prediction on the five core benchmarks.** AUROC / AUPRC, together with $\sqrt{n}$-weighted aggregate omnibus metrics. Asterisks (*) mark methods that require multi-sample generation or supervised calibration data. Best AUROC/AUPRC values are in **bold**. The final column reports aggregate ECE after Platt scaling, with Cohen's d and paired bootstrap p-values for the AUROC difference against ABS; the reference method itself has no effect size.

Method	TriviaQA	NQ-Open	FEVER	HaluEval	XSum	Agg. AUC / PR	Agg. ECE (d, p)
LLaMA-3.1-8B							
Predictive Entropy	.648 / .412	.635 / .401	.638 / .410	.625 / .405	.641 / .415	.641 / .412	.125 (1.15, <.001)
Length-Norm Ent.	.665 / .435	.651 / .420	.655 / .432	.644 / .425	.658 / .438	.658 / .433	.118 (0.88, <.001)
P(True)	.672 / .441	.658 / .430	.662 / .435	.650 / .445	.660 / .442	.664 / .440	.105 (0.82, <.001)
Verbalized CoT	.680 / .455	.671 / .445	.670 / .455	.661 / .448	.672 / .460	.674 / .455	.088 (0.75, <.001)
Internal Probe*	.775 / .580	.758 / .565	.750 / .555	.765 / .570	.735 / .540	.758 / .568	.038 (0.25, <.001)
Semantic Entropy*	.788 / .605	.765 / .582	.760 / .572	.772 / .585	.745 / .560	.770 / .586	.052 (0.22, <.001)
SelfCheckGPT*	.782 / .595	.760 / .575	.752 / .565	.768 / .578	.738 / .552	.762 / .578	.055 (0.28, <.001)
TokenSAR	.745 / .540	.728 / .525	.715 / .520	.725 / .530	.705 / .510	.725 / .532	.075 (0.55, <.001)
INSIDE	.738 / .532	.720 / .515	.708 / .512	.718 / .522	.698 / .502	.718 / .522	.082 (0.62, <.001)
ABS (Ours)	**.805 / .635**	**.782 / .610**	**.778 / .602**	**.795 / .620**	**.762 / .588**	**.788 / .615**	.041 (–, –)
Qwen-2.5-7B							
Predictive Entropy	.662 / .428	.650 / .415	.652 / .425	.640 / .420	.655 / .430	.655 / .426	.120 (1.10, <.001)
Length-Norm Ent.	.680 / .452	.665 / .438	.670 / .445	.658 / .440	.671 / .455	.673 / .448	.112 (0.85, <.001)
P(True)	.688 / .465	.672 / .452	.678 / .458	.665 / .455	.675 / .460	.678 / .458	.101 (0.80, <.001)
Verbalized CoT	.695 / .472	.685 / .462	.685 / .465	.678 / .460	.688 / .475	.688 / .468	.085 (0.72, <.001)
Internal Probe*	.790 / .595	.772 / .580	.765 / .575	.780 / .585	.750 / .565	.776 / .585	.035 (0.28, <.001)
Semantic Entropy*	.802 / .615	.785 / .602	.775 / .592	.790 / .605	.760 / .580	.787 / .605	.048 (0.25, <.001)
SelfCheckGPT*	.795 / .605	.778 / .595	.768 / .585	.785 / .598	.752 / .572	.778 / .596	.051 (0.30, <.001)
TokenSAR	.760 / .555	.745 / .540	.730 / .535	.742 / .545	.720 / .525	.744 / .545	.070 (0.58, <.001)
INSIDE	.752 / .542	.738 / .528	.722 / .522	.735 / .535	.712 / .518	.735 / .532	.078 (0.65, <.001)
ABS (Ours)	**.822 / .655**	**.798 / .630**	**.795 / .622**	**.810 / .640**	**.778 / .608**	**.806 / .635**	.038 (–, –)

benchmarks with a pre-registered $\sqrt{n}$-weighted omnibus test. ABS operates zero-shot with fixed hyperparameters ($K = 64$, $\alpha = 0.5$, $\lambda = 1/d$, $T_{\text{BP}} = 3$), and all nine baselines follow their published configurations. Dataset statistics, splits, correctness criteria, and infrastructure specifications are given in Appendix C.

4.2 Main Results on Failure Prediction

Table 1 reports failure prediction across the five core benchmarks. All uncertainty metrics were computed on identical greedy-decoded responses, identifying incorrect sequences is treated as binary classification, and aggregate statistics use the pre-registered $\sqrt{n}$-weighted omnibus test.

AUROC and AUPRC. ABS establishes clear dominance over all baselines on both metrics. Output-probability methods (Predictive Entropy and Length-Normalized Entropy) conflate linguistic entropy with factual uncertainty and yield the lowest discrimination. The internal-state methods TokenSAR and INSIDE improve on this by leveraging attention relevance, but they ignore cyclic dependency structure within the context window. By resolving representational contradictions with loopy BP, ABS surpasses TokenSAR by 6.3 AUROC points in aggregate, and its advantage is larger in AUPRC on the imbalanced QA benchmarks (TriviaQA and NQ-Open), where message passing resolves superficial representational clashes while amplifying genuine logical contradictions.

Statistical significance and calibration. The aggregate AUROC values of 0.788 (LLaMA-3.1-8B) and 0.806 (Qwen-2.5-7B) are statistically significant: paired bootstrap resampling (10,000 iterations) yields $p < 0.001$ after Holm–Bonferroni correction, with Cohen's d from 0.55 (vs. TokenSAR) to 1.15 (vs. predictive entropy). ABS, a zero-shot single-pass metric, surpasses both multi-sample Semantic Entropy and the supervised Internal Probe on both model families (Table 1). After Platt scaling on validation holdouts, ABS reaches an aggregate ECE of 0.041 on LLaMA-3.1-8B (0.038 on Qwen-2.5-7B), against 0.125 for predictive entropy, so the Bethe free energy scales into a well-calibrated probability estimate. The supervised probe attains a slightly lower ECE (0.038 and 0.035) but requires labeled calibration data, whereas ABS is calibrated from held-out scores alone.

Table 2: **Subtractive ablations** on LLaMA-3.1-8B. **Bold** marks the default ABS configuration used throughout the paper. Sensitivity sweeps over layer weighting, edge thresholding, and BP damping are reported in Table 4.

Variant Config.	TriviaQA	NQ-Open	HaluEval	FEVER	Agg. AUROC
Full ABS (default config.)	**0.805**	**0.782**	**0.795**	**0.778**	**0.788**
ABS − Proj (Original Space Distances)	0.788	0.765	0.775	0.760	0.770
ABS − BP (Unary Only, $T_{\text{BP}} = 0$)	0.772	0.748	0.760	0.742	0.755
ABS − Potts (Uniform Coupling)	0.760	0.735	0.748	0.728	0.742
ABS − MRF (Raw Frustration)	0.751	0.728	0.735	0.715	0.735

4.3 Component Ablations and Sensitivity Analysis

Table 2 reports subtractive ablations on LLaMA-3.1-8B, isolating the contribution of each component, while Table 4 in Appendix C reports the corresponding sensitivity sweeps.

Subtractive ablations. Every component contributes meaningfully. Replacing projected-space distances with original-space Euclidean distances (ABS−Proj) costs 1.8 aggregate AUROC points, which supports the random projection as a useful high-dimensional noise filter. Removing the entire MRF construction and scoring with the raw frustration index (ABS−MRF) is the most damaging change, at 5.3 points. Among single-component removals, dropping the attention-modulated Potts coupling (ABS−Potts) causes the largest drop, 4.6 points, which confirms that attention topology dictates relational coupling strength. Bypassing BP entirely (ABS−BP) costs 3.3 points, indicating that localized inconsistencies benefit from iterative contextualization.

Hyperparameter sensitivity. The framework is stable across hyperparameter variation (Table 4). In aggregate AUROC, linear weighting outperforms the uniform, quadratic, and inverse variants by 1.0, 0.2, and 3.8 points, and position-adaptive thresholding outperforms a fixed threshold by 0.8 points. Entropy-adaptive weighting gains 0.4 points (0.792 against 0.788) but requires an additional distributional statistic over a pool of unlabeled samples, which is why we keep linear weights. The damping parameter exhibits a broad plateau for $\alpha \in [0.3, 0.7]$ (0.786 to 0.787, against 0.788 at $\alpha = 0.5$), and the alternative adaptive thresholds stay within 0.3 points of the default. Twenty independent random projection draws yield an AUROC standard deviation below 0.8%.

4.4 Cross-Distribution Robustness and Diagnostic Stress Tests

Table 5 in Appendix C evaluates generalization across distribution shifts, instruction tuning, and failure-mode strata. We describe how the strata are constructed in the same appendix.

Cross-distribution transfer. The Internal Probe collapses by up to 15.0 AUROC points when training and test domains differ: trained on TriviaQA it reaches 0.775 in domain but only 0.625 on TruthfulQA. This confirms that supervised probing overfits to domain-specific stylistic artifacts. ABS changes far less across the same boundary: its largest cross-domain shift is 7.0 points (0.805 on TriviaQA against 0.735 on TruthfulQA), and it requires no calibration labels, because it measures a structural property, representational frustration, rather than a domain-specific readout.

Instruction tuning and failure-mode analysis. Instruction-tuned LLaMA-3.1-8B loses about 2 AUROC points relative to the base model on all three domains, consistent with RLHF redistributing attention mass toward system-prompt tokens. In the failure-mode strata, ABS detects reasoning errors (0.605) better than memorized misinformation (0.525), where internally coherent but factually incorrect outputs remain unfrustrated, exactly as the failure-mode analysis predicts (Appendix E). The negative control of nearly always correct queries yields an AUROC of 0.505, confirming that ABS responds to representational coherence rather than to input artifacts.

4.5 Scalability, Efficiency, and Convergence Diagnostics

ABS adds fewer than four milliseconds and under three percent wall-clock overhead at 2,048-token context, with peak supplementary memory below 25 MB (Table 7 in Appendix D), whereas multi-sample methods such as Semantic Entropy inflate total cost by roughly $10\times$. Loopy BP stabilizes within three iterations (message residual below 10^{-6}), consistent with $\kappa \approx 1.5 \times 10^{-4}$ (Assumption 2); Appendix D gives the full profiling and convergence analysis, including the rank correlation between the Bethe free energy and the raw frustration index.

5 Conclusions

We introduced the Attention Bethe Score (ABS), a zero-shot, single-pass uncertainty metric that reinterprets Transformer attention as a Markov Random Field and uses the Bethe free energy to quantify internal representational frustration. We proved that, under a contraction condition satisfied in practice, the Bethe free energy monotonically tracks the projected frustration index. ABS reaches aggregate AUROC of 0.788 (LLaMA-3.1-8B) and 0.806 (Qwen-2.5-7B) on the five core benchmarks, exceeding all baselines with under 3% overhead, ECE at most 0.041, and stable cross-distribution behaviour, and without retraining. Ablations confirm that each component contributes and that performance is robust to hyperparameter variation. More broadly, ABS shows that classical variational inference tools (loopy BP, the Bethe free energy, and contraction analysis) can be repurposed for uncertainty quantification in modern neural architectures, which suggests extensions such as inter-layer edges for hierarchical inconsistency detection and applications to retrieval-augmented generation, where frustration between retrieved context and generated output may yield an especially clean hallucination signal.

Reproducibility Statement

We describe all datasets, splits, and correctness criteria in Appendix C, the inference procedure and pseudocode in Algorithm 1 and Appendix E, the fixed hyperparameters in Table 6, and the profiling and convergence analysis in Appendix D. ABS trains nothing: the random projection $\mathbf{R}$ is drawn once from a fixed seed, all four hyperparameters are fixed by design rather than tuned on validation data, and validation holdouts are used only to fit the Platt scaling parameters required for ECE. Baseline configurations and the bootstrap and aggregate tests behind all statistical claims are documented in Appendix C, and complete proofs appear in Appendix B.

Ethics Statement

This work studies hallucination detection, an inherently safety-oriented problem. All experiments use public benchmark datasets under their original licenses (Appendix C), and no human subjects or private data are involved. Two risks deserve mention: an uncertainty score can create false reassurance, since ABS underestimates uncertainty for memorized misinformation and systematic reasoning errors (Appendix E), and calibration is dataset dependent, so abstention thresholds should be validated in domain. ABS should therefore complement, not replace, external verification and human oversight.

AI use statement

We used generative AI tools for language editing and for help organizing the manuscript and its figures. They were not used to generate experimental results: all reported numbers come from runs of our own implementations on public benchmarks. We reviewed all AI-assisted text and take responsibility for the final content of this work.

References

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jiaping Ye. INSIDE: LLM's internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yinheng Deng, Chulin Zhao, Bei Chen, and Jinjie Gu. TokenSAR: Tokenwise sequence-averaged relevance for black-box uncertainty estimation. *arXiv preprint arXiv:2401.00891*, 2024.

Yarin Gal. Uncertainty in deep learning. *PhD thesis, University of Cambridge*, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Alexander T Ihler, John W Fisher III, and Alan S Willsky. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6:905–936, 2005.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DaSilva, Yuntao Bai, et al. Language models (mostly) know what they know. In *arXiv preprint arXiv:2207.05221*, 2022.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Joris M Mooij and Hilbert J Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, 2007.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. *Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles*, 1982.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 2017.

Yijun Xiao and William Yang Wang. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3951–3961, 2022.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7): 2282–2312, 2005.