# Measuring Local Posterior Activity in Variational Language Model Units

Yves Ruffenach
Conservatoire National des Arts et Métiers
Strasbourg, France
ORCID: 0009-0009-4737-0555
yves@ruffenach.net
DOI: 10.5281/zenodo.21641827

**Abstract**

Language-model evaluation commonly focuses on final predictive distributions through likelihood, calibration, accuracy, entropy, and variability across predictions. Internal probabilistic activity offers a complementary measurement level by describing how distributions are represented within the computation that produces those predictions.

EVE, the *Elemental Variational Expanse*, is a variational language-model unit that maps a selected hidden activation to an input-conditioned Gaussian posterior. During the forward computation, the unit exposes divergence from a reference prior, posterior scale, posterior mean energy, and active-unit fraction. These quantities form an internal measurement panel evaluated alongside deterministic, dropout-based, and ensemble-based predictive references.

Controlled evaluations show competitive predictive behavior together with directly observable posterior activity. Validation-selected posterior readouts provide distinct calibration and tail-risk operating points, while controlled changes in context quality and target frequency produce systematic variation in both output metrics and internal posterior statistics. The resulting framework combines predictive evaluation with explicit probabilistic measurements inside language-model computation.

## 1 Introduction

A language model (LM) receives a sequence of tokens and estimates the probability of each possible next token. The input passes through successive computational layers, each producing *hidden activations*: numerical representations that carry contextual information from one stage to the next. In a conventional neural network, a fixed input and fixed model parameters produce a single activation value or vector at each unit. The final hidden representation is then converted into a probability distribution over the vocabulary.

Output-level evaluation characterizes this final distribution through likelihood, calibration, prediction accuracy, entropy, disagreement across stochastic predictions, and the concentration of loss in difficult examples. Internal probabilistic measurement addresses a complementary question: how a selected computation represents a distribution while the prediction is being formed. It describes the location, dispersion, and prior-relative specialization of an internal activation distribution for each input.

Predictive uncertainty and internal probabilistic activity therefore operate at distinct and complementary levels. Predictive uncertainty summarizes the model's final distribution over possible outputs. Internal probabilistic activity describes an explicitly modeled distribution inside the computation. Their joint analysis connects final predictive behavior with the probabilistic state carried by selected hidden units.

This connection is especially informative under *non-stationarity*, where evaluation or deployment data exhibit changes in vocabulary, writing style, context quality, topic frequency, target rarity, or observation order. Such changes can alter aggregate metrics and data-slice profiles. Internal variables add a second measurement surface that can be tracked across inputs, slices, and operating conditions.

EVE stands for *Elemental Variational Expanse*. An EVE unit converts a deterministic hidden activation into the parameters of a local probability distribution. For each input, the unit produces a posterior mean and posterior scale, draws or summarizes a latent activation from that distribution,

and transmits the resulting activation to the next computation. The *posterior* is the input-conditioned distribution learned by the unit. The *prior* is the reference distribution used to regularize that posterior.

The construction follows the logic of *variational inference* (VI), which learns a tractable approximation to a target posterior. Variational models commonly optimize an evidence lower bound (ELBO), balancing data fit with divergence between an approximate posterior and a prior. The Kullback–Leibler divergence (KL) measures this prior-relative difference. A smaller local KL indicates closer alignment with the prior, while a larger value indicates stronger input-conditioned specialization. Predictive calibration remains assessed through output-level metrics, while local KL characterizes the variational activation distribution.

The local posterior exposes four principal measurements. Posterior scale describes conditional dispersion. Posterior mean energy, defined from the squared posterior mean, describes displacement from zero. Internal KL describes divergence from the prior. Active-unit fraction summarizes the proportion of instrumented units exceeding a stated activity threshold. Together, these variables form an internal probabilistic measurement panel.

The principal evaluation terms are defined as follows. DET denotes the deterministic baseline. Monte Carlo (MC) sampling repeats a stochastic computation to approximate an expectation or predictive distribution; MC Dropout keeps dropout active during evaluation and aggregates repeated forward passes. Negative log-likelihood (NLL) measures the penalty assigned to the observed target, and perplexity (PPL) is the exponential of average NLL. Cross-entropy (CE) is the token-prediction training loss. Expected calibration error (ECE) summarizes the alignment between confidence and empirical accuracy. Conditional value at risk (CVaR) summarizes the mean loss within a specified high-loss tail; CVaR–NLL therefore describes negative log-likelihood among the most difficult evaluated cases. Mutual information (MI) summarizes disagreement attributable to variability across stochastic predictions. The notation $\mathrm{KL}_{\mathrm{int}}$ denotes unit-level internal Kullback–Leibler activity.

Several neighboring probabilistic traditions operate at different inferential scales. A variational autoencoder (VAE) commonly introduces a latent variable at the example, sequence, or time-step level. A Bayesian neural network places a probability distribution over model parameters. Dropout and deterministic ensembles generate predictive variability through masks or independently trained models. A Gaussian process (GP) defines a probability distribution over functions, and a Gaussian Process Neuron (GPN) applies this principle to a neuron's activation function. Uncertainty quantification (UQ) encompasses methods that estimate or propagate predictive and internal uncertainty. EVE places its explicitly modeled variational object at the activation transmitted by a selected unit for the current input.

The controlled evaluation compares raw EVE with deterministic, MC Dropout, and deterministic ensemble references. Validation-selected posterior readouts target calibration, high-loss tail behavior, and a balanced combination of likelihood, calibration, and tail risk. Controlled perturbations and data slices then characterize the responsiveness of internal KL, posterior scale, posterior mean energy, and active-unit fraction across changing evaluation conditions.

The central research question is: *can a language-model unit expose an explicit, input-conditioned activation posterior whose statistics provide useful information alongside standard predictive metrics?* The analysis addresses four objectives:

1. define local probabilistic activity as an explicit measurement target inside language-model computation;
2. compare the predictive behavior of EVE with deterministic, MC Dropout, and ensemble references under shared data splits, seeds, target windows, and evaluation panels;
3. characterize posterior readouts for likelihood, calibration, and high-loss tail objectives; and
4. measure the response of local KL activity, posterior scale, posterior mean energy, and active-unit fraction under controlled changes in the evaluation distribution.

## 2 Related probabilistic architectures

**Taxonomy by probabilistic object.** Probabilistic neural methods differ primarily in the location of their probabilistic object. Randomness may be attached to binary hidden states in graphical models, global or time-indexed latent variables, network parameters, sampled model identities, activation functions, deterministic representation densities, or individual activation variables. Each choice defines a specific inferential objective, set of observables, and computational profile. Table 1 summarizes these complementary constructions.

**Stochastic neurons and belief networks.** Stochastic hidden units have a long history in Boltzmann machines, sigmoid belief networks, and stochastic feed-forward networks [Ackley et al., 1985, Neal, 1992]. Gradient estimators and variational objectives later extended this lineage to networks with discrete stochastic neurons [Bengio et al., 2013, Mnih and Gregor, 2014]. These models establish the neuron as a random variable, commonly through Bernoulli or related conditional distributions in latent-variable generative systems. EVE belongs to this broad lineage while using a continuous, input-conditioned Gaussian posterior, an explicit unit-level prior, and posterior statistics designed for direct internal measurement.

**Variational latent-variable and sequential models.** Variational autoencoders and stochastic backpropagation introduced scalable amortized inference for continuous latent-variable models [Kingma and Welling, 2014, Rezende et al., 2014]. Extensions to text and sequences include sentence-level latent variables, recurrent latent states, and multiple stochastic layers [Bowman et al., 2016, Chung et al., 2015, Fraccaro et al., 2016]. EVE shares reparameterized sampling and KL regularization with this family. Its inferential granularity is unit-indexed: the posterior governs the activation transmitted by a local computational unit. Work on posterior collapse further motivates explicit activity diagnostics for latent channels [He et al., 2019]; the active-unit fraction provides an operational threshold-based view of utilization.

**Bayesian neural networks and local reparameterization.** Bayesian neural networks place distributions over weights and use approximate inference to represent parameter uncertainty [Graves, 2011, Blundell et al., 2015]. Variational dropout and the local reparameterization trick express weight uncertainty as activation-space noise for efficient gradient estimation [Kingma et al., 2015]. In that construction, the posterior remains parameter-centered. In EVE, $q_\phi(z_i \mid h_i)$ is the modeled variational object, $p(z_i)$ is its local prior, and $\mathrm{KL}(q_\phi(z_i \mid h_i) \| p(z_i))$ is available for each instrumented unit and input. The distinction concerns the locus and semantics of inference.

MC Dropout estimates predictive uncertainty through repeated stochastic forward passes [Gal and Ghahramani, 2016]. Deep ensembles use disagreement among independently trained predictors [Lakshminarayanan et al., 2017]. Both provide strong predictive references under standard and shifted data conditions [Ovadia et al., 2019]. Dropout supplies mask-driven activation variability, ensembles supply model-level variability, and post-hoc instrumentation can summarize their internal activations. EVE adds an amortized per-input activation posterior, a corresponding unit prior, and a unitwise KL readout. The interpretation of every variational approximation remains linked to its posterior family [Hron et al., 2018], which makes explicit reporting of the modeled object especially important.

**Probabilistic activation functions and activation-level propagation.** The Gaussian Process Neuron (GPN) places a Gaussian-process prior and posterior over an individual neuron's activation *function*, producing stochastic neuron-specific nonlinearities [Urban et al., 2017]. GPNs and EVE both locate probabilistic structure at neuron granularity. A GPN represents uncertainty over the activation function, while EVE maps $h_i$ parametrically to $(\mu_i, \sigma_i)$ and represents a posterior over the transmitted activation variable $z_i$. The resulting EVE measurements describe posterior scale, mean energy, and KL for the current activation distribution.

Stochastic monotonic activations and activation-moment propagation provide additional activation-level approaches [Ravanbakhsh et al., 2016, Postels et al., 2019, Haußmann et al., 2020]. Sampling-free and deterministic uncertainty methods derive efficient scores from propagated moments, learned evidence, distances, or related deterministic quantities [Postels et al., 2022]. Together, these methods establish a broad comparison space spanning posterior-based units, moment propagation, and deterministic uncertainty surrogates.

**Variational information bottlenecks and adaptive activation noise.** The variational information bottleneck (VIB) introduces a stochastic, input-conditioned representation and regularizes retained information through a variational bound [Alemi et al., 2017]. Information Dropout introduces data-dependent multiplicative activation noise and connects its regularizer to representation minimality and variational inference [Achille and Soatto, 2018]. VIB objectives also operate at neuron granularity for compression and pruning [Dai et al., 2018]. This lineage is closely related to EVE through input-conditioned stochastic representations inside the network. EVE uses additive Gaussian activation posteriors with standard-normal priors across 1,024 local units, propagates sampled activations through the language model, and records unitwise posterior statistics as an internal measurement panel.

Table 1: Probabilistic neural families organized by the locus of the modeled object. Local readout denotes the quantity directly available from the canonical formulation.

| Family | Primary object | Conditioning | Regularization locus | Local readout | Predictive passes | Relation to EVE |
|---|---|---|---|---|---|---|
| Deterministic network | Point activation | Deterministic input map | Fixed parameters | Activation values | Single | Point-activation reference. |
| Stochastic/belief-network units | Sampled hidden state | Often input-conditioned | Model-dependent latent objective | Sampled state statistics | One or several | Discrete and latent-state stochastic-neuron lineage. |
| VAE / variational sequence model | Global, sequence, or time-step latent state | Input-conditioned | Latent-state KL | Latent posterior statistics | One or several | Shared amortized VI at broader latent granularity. |
| Bayesian neural network | Weights and induced functions | Data-conditioned parameter posterior | Parameter-level divergence | Parameter posterior summaries | Several | Parameter-centered uncertainty. |
| MC Dropout | Dropout mask and model sample | Mask-driven | Dropout regularization | Predictive and activation variability | Several | Mask-based stochastic prediction. |
| Deep ensemble | Model identity | Training-data conditioned members | Independent objectives | Cross-model disagreement | Several | Model-level variability. |
| Gaussian Process Neuron | Neuron activation function | GP prediction | Function-level prior and posterior | Function-posterior moments | One or several | Function-posterior granularity. |
| VIB / Information Dropout | Stochastic representation or activation noise | Input-conditioned | Representation-level variational bound | Representation or noise statistics | One or several | Closely related activation-level variational lineage. |
| Density uncertainty layer | Density-conditioned predictive layer | Input-density conditioned | Density-based criterion | Layer-specific uncertainty | Single or several | Density-grounded predictive variance. |
| Moment propagation / deterministic UQ | Moments, evidence, or distance | Method-dependent | Method-specific | Propagated or deterministic scores | Often single | Efficient internal uncertainty surrogates. |
| EVE | Activation variable $z_i$ | $q_\phi(z_i \mid h_i)$ | Unit-level prior and KL | KL, $\sigma$, $\mu^2$, active fraction | Optional | Input-conditioned activation posterior as measurement layer. |

Density Uncertainty Layers provide a density-aware stochastic layer whose predictive variance is coupled to empirical input density [Park and Blei, 2024]. Their density-conditioned criterion complements EVE's activation-posterior diagnostics and broadens the family of single-model uncertainty layers relevant to comparative evaluation.

**Output-level uncertainty and calibration in language models.** Language-model uncertainty is commonly evaluated through token probabilities, entropy, calibration error, selective prediction, sample disagreement, and semantic consistency. Calibration research examines the relationship between Transformer confidence and correctness across domains [Guo et al., 2017, Desai and Durrett, 2020]. More recent approaches rank uncertainty measures against generation quality and aggregate generations by semantic equivalence [Huang et al., 2024, Farquhar et al., 2024]. Output-level methods characterize uncertainty in predictions or generated meanings; EVE characterizes probabilistic activity within selected computational units. Their joint use links internal posterior behavior with predictive risk and task outcomes.

**Hidden representations and distribution-shift measurement.** Mahalanobis-distance methods, activation-based probes, and related representation tests derive uncertainty or out-of-distribution scores from hidden representations [Lee et al., 2018, Slobodkin et al., 2023]. Dataset-shift measurement also compares input or representation distributions before labels become available [Rabanser et al., 2019]. These methods treat internal states as measurement surfaces through distances, densities, classifiers, or statistical tests. EVE complements them by emitting the parameters of a learned posterior directly during the forward computation. The controlled stress panel characterizes the responsiveness of these posterior variables across input perturbations and data slices.

**EVE formulation.** EVE combines an amortized posterior over a local activation variable, an explicit unit-level prior and KL regularizer, direct posterior statistics during the forward computation, and joint evaluation with predictive quality, calibration, tail risk, and controlled non-stationarity. The term "internal probabilistic activity" denotes these model-defined posterior quantities and their behavior across inputs and operating conditions.

# 3 Method

## 3.1 Computational path of the variational units

Let a token context contain $L$ positions and let $X^{(\ell-1)} \in \mathbb{R}^{L\times d}$ denote the sequence entering variational block $\ell$. The implementation uses $L = 24$, embedding width $d = 768$, three blocks, and $N = 1{,}024$ local latent units per block. Token vectors are taken from the frozen GPT-2 token-embedding matrix, while contextual computation is learned by the three variational blocks. A learned positional embedding is added before the first block.

Each block first applies layer normalization, multi-head self-attention, and a residual connection:

$$A^{(\ell)} = X^{(\ell-1)} + \mathrm{MHA}_\ell\Big(\mathrm{LN}_\ell(X^{(\ell-1)})\Big). \tag{1}$$

The attended sequence is flattened. For blocks after the first, the implementation concatenates the flattened sequence with the latent activations produced by preceding blocks. A linear map, Gaussian error linear unit (GELU), and dropout then produce a block summary $h^{(\ell)}$ of width 1,024.

For each local unit $i$, two input-conditioned affine maps parameterize a Gaussian posterior:

$$q_\phi\left(z_i^{(\ell)} \mid h^{(\ell)}\right) = \mathcal{N}\left(\mu_i^{(\ell)}(h^{(\ell)}), \left[\sigma_i^{(\ell)}(h^{(\ell)})\right]^2\right), \qquad p(z_i^{(\ell)}) = \mathcal{N}(0,1). \tag{2}$$

The logarithm of the posterior scale is clipped to $[-6, 2]$ before exponentiation; the resulting scale is constrained to a finite positive interval. During most of training the upper scale bound is 7.0 and is reduced to 5.8 from epoch 3. With one latent dimension per unit, reparameterized samples are

$$z_i^{(\ell,s)} = \mu_i^{(\ell)} + \sigma_i^{(\ell)} \epsilon_i^{(\ell,s)}, \qquad \epsilon_i^{(\ell,s)} \sim \mathcal{N}(0,1). \tag{3}$$

Four latent samples are used in each training forward pass. Their unit activations are projected back to an $L \times d$ sequence and added to the attended sequence:

$$X^{(\ell)} = A^{(\ell)} + \mathrm{reshape}\left(W_z^{(\ell)} \tilde{z}^{(\ell)} + b_z^{(\ell)}\right). \tag{4}$$

The latent activation $z^{(\ell)}$ is projected explicitly into a sequence-shaped residual update. The three block-level latent representations are combined with learned softmax weights. A gated prediction head of width 1,024 maps the resulting representation back to the vocabulary using the same frozen token-embedding matrix as a tied output matrix.

## 3.2 Training objective and control terms

The training objective combines next-token prediction, adaptive KL regulation, posterior-energy control, local reconstruction, and temporary auxiliary supervision. For a minibatch, it is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \sum_{i=1}^{N} \beta_i \, \overline{\mathrm{KL}}_i + \lambda_{\mathrm{band}} \mathcal{B}(\mu^{(1)}) + \lambda_{\mathrm{loc}} \mathcal{R}_{\mathrm{loc}} + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{CE}}$ is next-token cross-entropy, $\overline{\mathrm{KL}}_i$ sums the local KL contribution of unit $i$ across the three variational blocks and averages it over the minibatch, and $\beta_i$ is an adaptive per-unit coefficient. The controller starts at $2 \times 10^{-4}$, is constrained to $[10^{-6}, 10^{-1}]$, and adjusts the coefficients to keep local KL activity within configured operating targets. A fixed reference coefficient of $10^{-4}$ is also configured.

The mean-energy band term $\mathcal{B}$ is a squared hinge penalty applied to the first block's batch-mean $\mu_i^2$. Its nominal target is 0.1, with lower and upper bounds equal to 0.8 and 1.45 times that target; its base coefficient is $\lambda_{\mathrm{band}} = 0.004$. This explicit homeostatic control defines the operating range of posterior mean energy.

The local term $\mathcal{R}_{\mathrm{loc}}$ is a per-neuron reconstruction loss. Each sampled latent unit is decoded into a one-dimensional target obtained from a learned projection of the block summary; the target is detached before reconstruction. Its coefficient ramps to 0.03 over 400 optimization steps and is reduced to 0.022 from epoch 3. A final-block auxiliary cross-entropy term with coefficient 0.05 is active during epoch 3 only. KL-floor and layer-weight-floor coefficients are set to zero in the reported runs. Together, these terms define the complete evaluated system and provide a clear basis for component-wise ablation studies.

## 3.3 Internal measurements

For evaluated inputs, EVE exposes four model-defined measurements:

- **Internal KL**: the mean unitwise divergence from the standard-normal prior;
- **Posterior scale** $\sigma$: the mean conditional posterior standard deviation;
- **Posterior mean energy** $\mu^2$: the mean squared posterior location;
- **Active-unit fraction**: the proportion of units whose mean evaluation KL exceeds $10^{-4}$.

The first three statistics are averaged over evaluated examples, stochastic predictive passes, and the implementation's block aggregation. They characterize the learned posterior construction and complement output-level calibration and error metrics. All units exceed the operational activity threshold in the reported runs, providing a threshold-specific utilization measure.

### 3.4 Validation-selected posterior readouts

The trained seed-0 checkpoint is held fixed while a grid of temperatures and posterior-bank aggregation rules is evaluated on validation data. Candidate readouts include the mean of probabilities, the softmax of mean logits, the median probability, entropy weighting, and confidence weighting. The selected temperature/readout pair is then frozen and applied to the test panel.

The calibration-first policy minimizes

$$S_{\mathrm{ECE}} = \mathrm{ECE} + 0.01\,\mathrm{NLL} + 0.001\max(0, \mathrm{CVaR} - \mathrm{NLL}), \tag{6}$$

the tail-risk-first policy minimizes

$$S_{\mathrm{CVaR}} = \mathrm{CVaR} + 0.05\,\mathrm{NLL} + 0.25\,\mathrm{ECE}, \tag{7}$$

and the balanced policy minimizes

$$S_{\mathrm{bal}} = \mathrm{NLL} + 2\,\mathrm{ECE} + 0.05\max(0, \mathrm{CVaR} - \mathrm{NLL}). \tag{8}$$

Policy selection uses validation labels, after which the selected rule is frozen and applied to the test panel.

## 4 Experimental protocol

**Dataset and windowing.** The evaluation uses the `RLAIF/WritingPrompts-Filtered` dataset and the GPT-2 tokenizer [Fan et al., 2018]. A one-percent dataset sample contains 1,992 prompt–story pairs, split reproducibly into 1,394 training, 299 validation, and 299 test pairs. Prompts are truncated to 96 tokens, stories to 192 tokens, and selected stories contain at least 32 tokens. Fixed-context next-token windows use 24 context tokens and a target stride of two, producing 134,086 training, 28,662 validation, and 28,898 test windows.

**Architectures and parameter counts.** All systems use the same frozen GPT-2 token-embedding matrix, containing 38,597,376 frozen parameters, and tied vocabulary projection. The deterministic reference flattens the frozen context embeddings, maps them to 1,024 deterministic units, and applies a gated prediction head. It has 20,715,265 trainable parameters. The MC Dropout reference uses the deterministic architecture and activates its dropout modules during stochastic evaluation. The deterministic ensemble averages five independently trained deterministic checkpoints.

The EVE system uses the three variational Transformer blocks described above and has 134,852,868 trainable parameters. The comparison is system-level: data splits, seeds, target windows, and evaluation code are shared, while architecture, capacity, and training cost follow each system's specified design. Component-level attribution is addressed through the ablation program described in Section 7.

**Optimization.** Five training seeds, 0 through 4, are used for DET, MC Dropout, and raw EVE. Models are trained for four epochs with AdamW, weight decay $10^{-4}$, gradient clipping at 1.0, minibatches of 48, and mixed-precision `float16` execution on a single CUDA device. The learning rate is $2 \times 10^{-4}$ for deterministic models and $1.5 \times 10^{-4}$ for EVE. The checkpoint with the lowest validation cross-entropy is retained.

**Evaluation panels.** Two fixed evaluation panels support complementary analyses. The five-seed comparison and non-stationarity analysis evaluate six minibatches, hence 288 windows per seed and scenario. The representative seed-0 policy search and five-member ensemble use twelve minibatches, hence 576 windows. The full test split supplies separately logged cross-entropy, perplexity, and accuracy fields. The five-seed table uses the dedicated 288-window stochastic panel, preserving a consistent evaluation basis across seeds.

**Predictive metrics.** Four repeated predictive passes are used for stochastic metrics. If $p_s(y \mid x)$ is the predictive distribution from pass $s$, the averaged distribution is $\bar{p} = S^{-1}\sum_s p_s$. NLL is $-\log \bar{p}(y^\star \mid x)$ averaged over windows, and perplexity is exp(NLL). Accuracy uses the top-1 class under $\bar{p}$. ECE uses 15 equal-width confidence bins. Mutual information is

$$\mathrm{MI} = H(\bar{p}) - \frac{1}{S}\sum_{s=1}^{S} H(p_s), \tag{9}$$

and top-1 flip rate is the fraction of stochastic passes whose argmax differs from the argmax of $\bar{p}$. CVaR–NLL is the mean NLL among the worst five percent of evaluated windows [Rockafellar and Uryasev, 2000].

**Statistical summaries.** Tables report means and sample standard deviations across five seeds. The paired base comparison includes two-sided 95% Student-$t$ intervals for seedwise differences. Stress scenarios include percentile intervals obtained from 200 resamples of the five seed-level values. These summaries quantify seed-level variation, while hierarchical prompt–story resampling provides the natural extension for data-level uncertainty decomposition.

# 5 Results

Table 2: Fixed-panel predictive metrics. Rows marked $5s$ report mean ± sample standard deviation across seeds 0–4 on the six-minibatch (288-window) panel. The deterministic ensemble is a five-member result on the complementary twelve-minibatch (576-window) panel and provides an output-level reference. Lower is better for NLL, PPL, ECE, and CVaR-NLL; higher is better for accuracy. Dashes identify EVE-specific posterior quantities.

| Model / policy | Setting | NLL | PPL | Acc. | ECE | CVaR | $\mathrm{KL}_{\mathrm{int}}$ | $\sigma$ | $\mu^2$ |
|---|---|---|---|---|---|---|---|---|---|
| DET | $5s$ | 5.478±0.020 | 239.5±4.9 | 0.159±0.013 | 0.038±0.009 | 12.685±0.388 | – | – | – |
| MC Dropout | $5s$ | 5.482±0.023 | 240.3±5.4 | 0.159±0.014 | 0.043±0.013 | 12.678±0.428 | – | – | – |
| EVE raw | $5s$ | **5.289±0.035** | **198.3±6.9** | **0.165±0.012** | 0.048±0.007 | **12.205±0.132** | 0.271±0.019 | 0.873±0.006 | 0.051±0.005 |
| DET Ensemble | 5 members | 5.455 | 233.8 | 0.163 | 0.047 | 12.761 | – | – | – |

Table 2 reports the fixed-panel seed comparison. Raw EVE records lower mean NLL, perplexity, and CVaR–NLL than both single-model references, together with slightly higher mean accuracy and higher mean ECE. The rows characterize system-level predictive behavior under a shared data and evaluation protocol, with architecture and parameter counts reported explicitly.

Table 3: Paired seedwise differences, raw EVE minus comparator, on the 288-window panel. Values are the mean difference and a two-sided 95% Student-$t$ interval over five paired seeds. For NLL, PPL, ECE, and CVaR, negative values indicate lower values for EVE; for accuracy, positive values indicate higher values for EVE. Intervals summarize variation across the five paired seeds.

| Comparator | Metric | Mean difference | 95% interval |
|---|---|---|---|
| DET | NLL | −0.189 | [−0.240, −0.138] |
| DET | PPL | −41.19 | [−51.61, −30.78] |
| DET | Accuracy | +0.0063 | [−0.0020, +0.0145] |
| DET | ECE | +0.0104 | [−0.0112, +0.0319] |
| DET | CVaR–NLL | −0.480 | [−0.953, −0.008] |
| MC Dropout | NLL | −0.192 | [−0.247, −0.138] |
| MC Dropout | PPL | −42.00 | [−53.37, −30.64] |
| MC Dropout | Accuracy | +0.0063 | [−0.0040, +0.0165] |
| MC Dropout | ECE | +0.0051 | [−0.0191, +0.0294] |
| MC Dropout | CVaR–NLL | −0.474 | [−0.992, +0.045] |

The paired differences are most concentrated for NLL and perplexity: all five seedwise comparisons place raw EVE below both references. Accuracy and ECE intervals span both directions, while CVaR–NLL is lower in every seed and its interval against MC Dropout includes both directions. The intervals summarize seed-level variation within the defined 288-window panel.

The validation-selected EVE readouts in Table 4 illustrate how the same internal posterior can be operationalized for different measurement objectives. EVE-ECE-selected preserves similar likelihood while attaining lower ECE than the raw representative checkpoint. EVE-CVaR-selected obtains the strongest tail-risk score while shifting the readout toward a specialized tail-risk operating point with higher NLL and ECE. EVE-balanced provides a tail-risk-oriented compromise while retaining competitive likelihood and calibration.

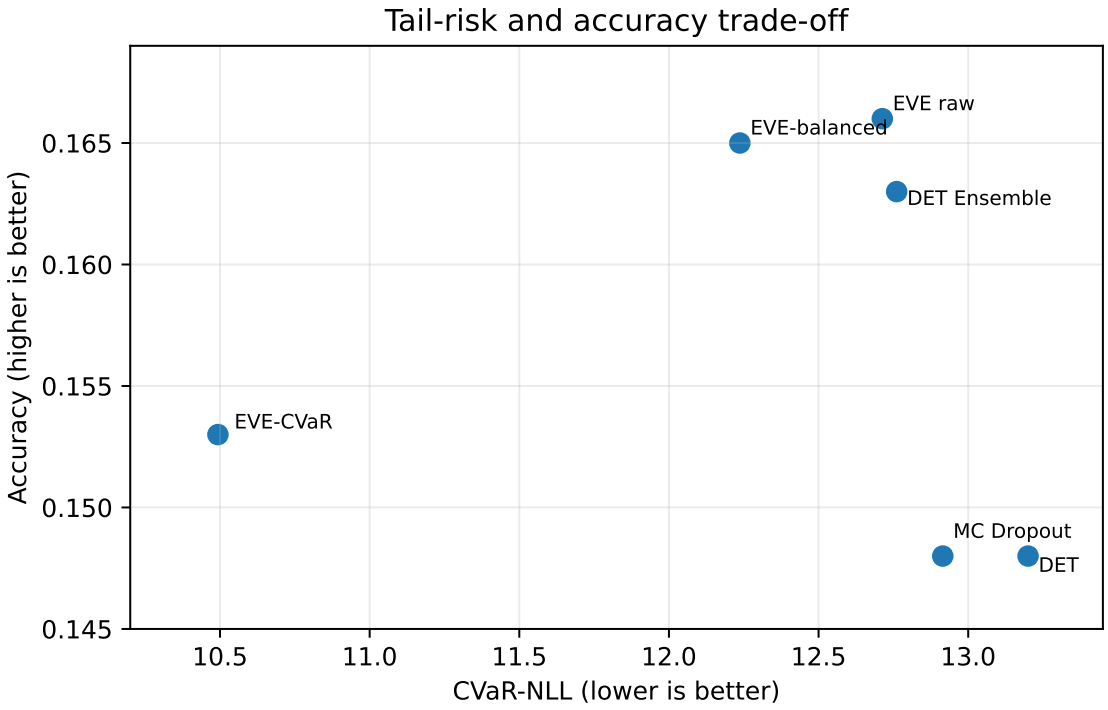


Figure 1: Accuracy and tail-risk trade-off. The plot combines seed-matched base-test rows for DET, MC Dropout, and raw EVE with representative validation-selected EVE readouts. In the plotted values, EVE-balanced occupies a higher-accuracy/lower-CVaR point than the deterministic ensemble reference; the figure provides a visual system-level comparison across the two reported evaluation panels. EVE-CVaR-selected identifies a specialized high-loss-tail operating point with higher NLL and ECE.

## 5.1 Policy-dependent measurement

Table 4: Validation-selected readouts for the fixed seed-0 EVE checkpoint on the twelve-minibatch (576-window) test panel. The exact validation scores are defined in the Method section; only temperature and posterior-bank aggregation are selected.

| Policy | Val. target | $T$ | Readout | NLL | ECE | CVaR |
|---|---|---|---|---|---|---|
| EVE raw | raw CE | – | raw | 5.392 | 0.055 | 12.836 |
| EVE-ECE-selected | ECE-first score | 1.1 | mean probability | 5.360 | 0.044 | 12.234 |
| EVE-CVaR-selected | CVaR-first score | 3.0 | confidence weighted | 7.599 | 0.148 | 10.493 |
| EVE-balanced | balanced score | 1.1 | mean probability | 5.410 | 0.046 | 12.237 |

Table 4 shows that post-hoc EVE readouts are policy-dependent. EVE-CVaR-selected achieves the best CVaR-NLL, while shifting the readout toward a specialized tail-risk operating point with higher NLL and ECE. This row defines a specialized tail-risk operating point. EVE-ECE-selected and EVE-balanced provide complementary readouts that attain lower ECE or CVaR values while keeping NLL close to raw EVE in this representative checkpoint.

This result is important for uncertainty-aware model evaluation. It shows that uncertainty measurement is multi-objective. The same internal posterior can be operationalized differently depending on the measurement target. Calibration-oriented, tail-risk-oriented, and balanced predictive measurements therefore provide complementary views of the same probabilistic activity.

## 5.2 Internal probabilistic activity

The internal posterior panel remains numerically stable across the four readouts, whereas the output metrics in Table 4 can change substantially. This is expected because the checkpoint and its local posterior parameterization are fixed; the policies primarily alter how posterior samples are aggregated and temperature-scaled. The result illustrates that a readout operating point and the underlying posterior

Table 5: Posterior diagnostics for the fixed seed-0 checkpoint under the representative readouts. Differences arise from stochastic evaluation and aggregation; the checkpoint parameters are identical.

| Policy | $KL_{int}$ | $\sigma$ | $\mu^2$ | Active fraction |
|---|---|---|---|---|
| EVE raw | 0.288 | 0.876 | 0.055 | 1.000 |
| EVE-ECE-selected | 0.296 | 0.876 | 0.055 | 1.000 |
| EVE-CVaR-selected | 0.291 | 0.876 | 0.055 | 1.000 |
| EVE-balanced | 0.287 | 0.876 | 0.055 | 1.000 |

diagnostics are related but distinct objects.

### 5.3 Lightweight non-stationarity stress panel

To test responsiveness under controlled non-stationarity, the stress pipeline reuses the fixed six-minibatch panel for each seed. In the covariate perturbation, every non-padding token in both context tensors is independently replaced by the end-of-sequence token with probability 0.05, 0.15, or 0.30; targets remain unchanged and the corruption is deterministically seeded by batch and intensity. Target-frequency slices rank test windows by the add-one-smoothed frequency of their target token in the training stories and define the rare and common quarters. Context-length slices use the shortest and longest quarters by non-padding length. Four ordered monitoring bins contain consecutive, non-overlapping groups of 288 test windows. The main table reports the base panel, the covariate dose response, and the rare-target slice; Appendix B reports the remaining summaries.

Table 6: Five-seed non-stationarity stress panel. The goal is to test whether EVE internal measurements move under controlled shift-like conditions. Lower is better for NLL, ECE, and CVaR-NLL.

| Scenario | Model | Seeds | NLL | ECE | CVaR | $\mathrm{KL}_{\mathrm{int}}$ | $\sigma$ | $\mu^2$ |
|---|---|---|---|---|---|---|---|---|
| Base | EVE raw | 5 | 5.289 | 0.048 | 12.205 | 0.271 | 0.873 | 0.051 |
| Covariate shift 0.05 | EVE raw | 5 | 5.545 | 0.057 | 13.073 | 0.299 | 0.874 | 0.052 |
| Covariate shift 0.15 | EVE raw | 5 | 6.230 | 0.056 | 16.043 | 0.305 | 0.878 | 0.057 |
| Covariate shift 0.30 | EVE raw | 5 | 8.054 | 0.103 | 18.507 | 0.373 | 0.900 | 0.075 |
| Covariate shift 0.30 | EVE-balanced | 5 | 7.838 | 0.073 | 17.227 | 0.363 | 0.899 | 0.075 |
| Rare targets | EVE raw | 5 | 10.033 | 0.119 | 15.839 | 0.308 | 0.870 | 0.055 |
| Rare targets | EVE-balanced | 5 | 9.732 | 0.093 | 14.858 | 0.303 | 0.870 | 0.055 |

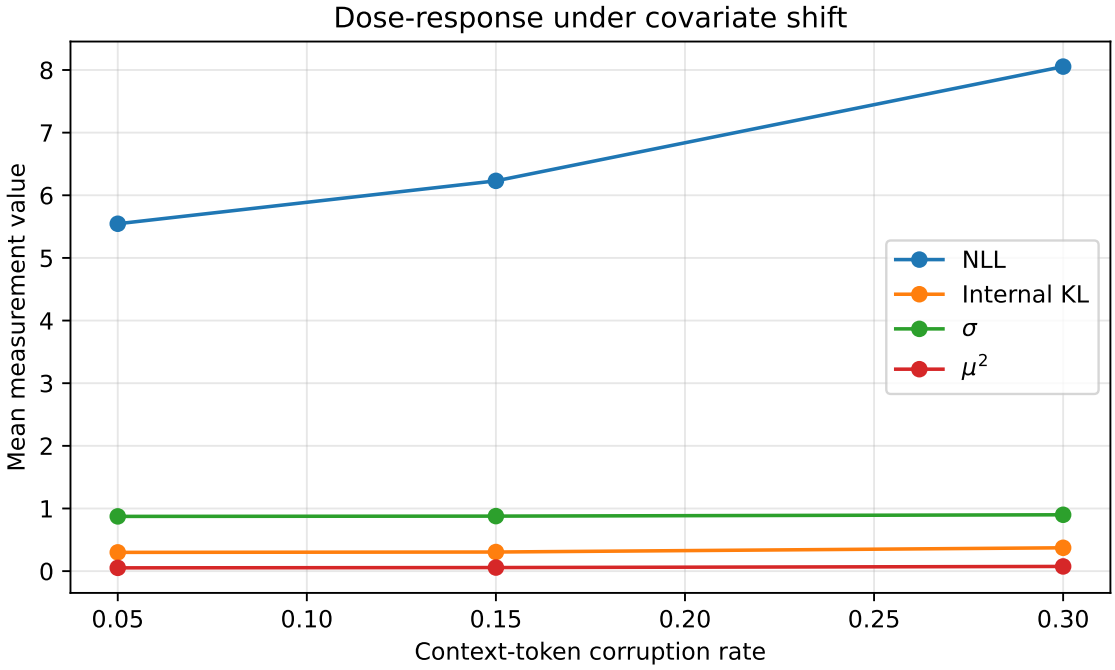


Figure 2: Five-seed dose response for raw EVE under deterministic context-token corruption. The figure presents the measured dose response under the controlled context-token corruption protocol.

The covariate dose response changes both output metrics and internal measurements. Relative to the base mean, raw EVE NLL rises progressively to 8.054 at corruption 0.30; internal KL rises to 0.373, posterior scale to 0.900, and posterior mean energy to 0.075. The seed bootstrap gives a 95% percentile interval of $[7.910, 8.184]$ for NLL and $[0.349, 0.393]$ for internal KL at corruption 0.30. Across the 15 raw-EVE seed–intensity observations, change in posterior mean energy correlates with change in NLL at $r = 0.971$ and with change in CVaR–NLL at $r = 0.888$; change in internal KL correlates with change in NLL at $r = 0.823$. These pooled correlations characterize dose-linked co-movement across ordered corruption levels. Incremental predictive contribution, temporal forecasting, and calibrated shift detection

form complementary validation targets.

Readout policies also support distinct post-shift measurement objectives. Under severe covariate shift and rare-target slices, EVE-balanced attains lower NLL, ECE, and CVaR than raw EVE in the reported averages. Validation-selected readouts further illustrate that calibration-oriented and tail-risk-oriented policies are distinct: the same internal posterior can support different operational choices depending on the measurement objective.

# 6 Discussion

**EVE as a measurement layer.** EVE provides model-defined local observability through an input-conditioned activation posterior and unitwise prior-relative KL. Deterministic models, dropout, ensembles, and post-hoc representation methods provide complementary uncertainty and internal-state signals. EVE adds posterior mean, posterior scale, local KL, and active-unit fraction directly within the forward computation. The reported comparison is system-level: EVE has approximately 6.5 times as many trainable parameters as the deterministic reference and incorporates attention, local reconstruction, homeostatic control, and auxiliary supervision.

**Connection to non-stationarity.** Controlled context perturbations, rare-target slices, context-length slices, and ordered monitoring bins reveal systematic variation in local KL, posterior scale, posterior mean energy, and output-level risk. This dose responsiveness establishes an empirical basis for extending EVE instrumentation to covariate shift, label shift, construct drift, and deployment-time recalibration.

**Trade-offs across measurement targets.** Calibration-first, tail-risk-first, and balanced readouts produce distinct operating points from the same posterior bank. The CVaR-first readout prioritizes the high-loss tail and yields higher average NLL and ECE, while the calibration-first and balanced policies preserve likelihood closer to the raw readout. An incremental-utility evaluation can compare output metrics with joint output-and-internal feature sets on held-out shifts and pre-specified thresholds.

# 7 Scope and research extensions

**Scale and sampling.** The reference configuration uses a one-percent dataset sample, a 24-token context, four training epochs, five seeds, and fixed stochastic panels of 288 or 576 windows. This configuration supports matched seed-level comparisons and efficient stress testing. Full-test evaluation and prompt–story-level resampling extend the analysis toward separate estimates of data, training-seed, checkpoint-selection, and Monte Carlo variability.

**Architecture and component analysis.** EVE contains 134.9 million trainable parameters, while the deterministic and MC Dropout references contain 20.7 million. Its architecture also includes three attention blocks, dense inter-block connections, local decoders, adaptive KL coefficients, an energy-band controller, and temporary auxiliary supervision. The resulting comparison characterizes complete systems. A component-wise analysis naturally includes a deterministic backbone matched in architecture and parameter count, $\beta = 0$, fixed and input-conditioned variance, mean-only inference, local-reconstruction and energy-control variants, prior sensitivity, latent width, and unit placement.

**Comparative activation-level methods.** A broader matched-compute benchmark can include VIB, Information Dropout, Gaussian Process Neurons, Density Uncertainty Layers, deterministic uncertainty methods, activation-moment propagation, and hidden-state shift detectors. Such a benchmark would quantify the relationship between activation-posterior instrumentation, representation-based monitoring, and output-level uncertainty.

**Interpretation of internal variables.** Internal KL quantifies divergence from the selected standard-normal prior. Posterior scale quantifies learned conditional dispersion under the configured clipping and control ranges. Posterior mean energy reflects posterior location together with the explicit energy-band objective. Active-unit fraction provides a threshold-indexed utilization measure. Threshold sweeps, calibration analyses, and epistemic–aleatoric decomposition experiments can refine the interpretation of each variable.

**Non-stationarity validation.** The stress panel establishes a controlled dose response under end-of-sequence corruption, frequency slices, length slices, and ordered subsets. Held-out natural shifts, area under the receiver-operating-characteristic curve, precision–recall analysis, temporal lead–lag measurement, and comparisons with output-only and representation-based detectors provide the next validation layer. Joint models using output metrics together with internal KL, $\sigma$, and $\mu^2$ can quantify incremental predictive and detection value.

# 8 Conclusion

An input-conditioned local activation posterior provides a direct internal measurement surface for language-model computation. The EVE architecture exposes unitwise KL activity, posterior scale, posterior mean energy, and active-unit fraction during prediction. Across five seeds on fixed evaluation panels, the complete EVE system records lower mean NLL and perplexity than the deterministic and MC Dropout references, while its internal statistics vary systematically with controlled context corruption. Validation-selected aggregation rules further provide distinct calibration and tail-risk operating points from the same checkpoint.

These results establish internal probabilistic activity as a measurable complement to output-level evaluation. The system-level comparison, posterior-readout analysis, and controlled stress response together define a foundation for parameter-matched component studies, broader activation-level baselines, natural distribution shifts, and direct measurement of the incremental value contributed by internal posterior variables.

# A Detailed configuration

Table 7: Experimental configuration.

| Component | Setting |
|---|---|
| Data | `RLAIF/WritingPrompts-Filtered`, fraction 0.01; GPT-2 tokenizer |
| Splits | 1,394 / 299 / 299 prompt–story pairs; 134,086 / 28,662 / 28,898 windows |
| Windowing | prompt cap 96, story cap 192, minimum story 32, context 24, stride 2 |
| Frozen representation | GPT-2 token-embedding matrix, 38,597,376 parameters; contextual computation learned by the EVE blocks |
| EVE backbone | three variational Transformer blocks; 12 attention heads; width 768; residual and attention dropout 0.05 |
| Local posterior | 1,024 units per block; latent width 1; input-conditioned log scale; initial $\sigma = 0.8$ |
| Sampling | four latent samples in training; configured evaluation aggregation 12; four repeated passes for stochastic metrics |
| Prediction head | gated width-1,024 head; tied frozen token matrix; score-weighted sample aggregation with 0.35 uniform mixing |
| Optimization | AdamW; batch 48; four epochs; LR $1.5 \times 10^{-4}$ (EVE), $2 \times 10^{-4}$ (DET); weight decay $10^{-4}$; gradient clip 1.0 |
| Posterior controls | adaptive per-unit $\beta$: initial $2 \times 10^{-4}$, range $[10^{-6}, 10^{-1}]$; standard-normal prior |
| Additional losses | mean-energy band weight 0.004; local reconstruction 0.03, reduced to 0.022 from epoch 3; final-block auxiliary CE 0.05 in epoch 3 |
| Metrics | four predictive passes; ECE 15 equal-width bins; CVaR worst 5%; active KL threshold $10^{-4}$ |
| Parameters | EVE 134,852,868 trainable; DET 20,715,265 trainable; both 38,597,376 frozen |

# B Additional stress summaries

The common-target slice combines low NLL with high ECE because its accuracy and confidence distributions differ from the base panel, illustrating the complementary roles of likelihood and calibration. The ordered bins display non-monotonic internal behavior across the recorded test order; chronological experiments can extend this view toward temporal drift analysis.

Table 8: Additional five-seed raw-EVE stress summaries on 288-window panels. The rows summarize data slices and ordered monitoring bins.

| Scenario | NLL | ECE | CVaR | $KL_{int}$ | $\sigma$ | $\mu^2$ |
|---|---|---|---|---|---|---|
| Common targets | 2.290 | 0.318 | 4.999 | 0.262 | 0.871 | 0.042 |
| Short contexts | 5.191 | 0.058 | 12.263 | 0.340 | 0.891 | 0.065 |
| Long contexts | 5.140 | 0.059 | 12.346 | 0.292 | 0.874 | 0.050 |
| Ordered bin 1 | 5.296 | 0.050 | 12.220 | 0.272 | 0.873 | 0.051 |
| Ordered bin 2 | 5.431 | 0.058 | 13.078 | 0.305 | 0.871 | 0.051 |
| Ordered bin 3 | 5.761 | 0.063 | 12.718 | 0.382 | 0.875 | 0.049 |
| Ordered bin 4 | 5.901 | 0.040 | 12.961 | 0.266 | 0.876 | 0.049 |

# References


Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018. doi: 10.1109/TPAMI.2017.2784440.

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. doi: 10.1207/s15516709cog0901_7.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *International Conference on Learning Representations*, 2017. arXiv:1612.00410.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622. PMLR, 2015.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21. Association for Computational Linguistics, 2016. doi: 10.18653/v1/K16-1002.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, volume 28, pages 2980–2988, 2015.

Bin Dai, Chen Zhu, Baining Guo, and David Wipf. Compressing neural networks using the variational information bottleneck. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1135–1144. PMLR, 2018.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.21.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, volume 29, pages 2199–2207, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR, 2016.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, volume 24, pages 2348–2356, 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Manuel Haußmann, Fred A. Hamprecht, and Melih Kandemir. Sampling-free variational inference of bayesian neural networks by variance backpropagation. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 563–573. PMLR, 2020.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019.

Jiri Hron, Alexander G. de G. Matthews, and Zoubin Ghahramani. Variational bayesian dropout: Pitfalls and fixes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2019–2028. PMLR, 2018.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.18.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28, pages 2575–2583, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pages 7167–7177, 2018.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1791–1799. PMLR, 2014.

Radford M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56(1):71–113, 1992. doi: 10.1016/0004-3702(92)90065-6.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yookoon Park and David Blei. Density uncertainty layers for reliable uncertainty estimation. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 163–171. PMLR, 2024.

Janis Postels, Francesco Ferroni, Sinem Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2931–2940, 2019. doi: 10.1109/ICCV.2019.00302.

Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17870–17909. PMLR, 2022.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Siamak Ravanbakhsh, Barnabás Póczos, Jeff Schneider, Dale Schuurmans, and Russell Greiner. Stochastic neural networks with monotonic activation functions. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 809–818. PMLR, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, 2014.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–42, 2000. doi: 10.21314/JOR.2000.038.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.220.

Sebastian Urban, Marcus Basalla, and Patrick van der Smagt. Gaussian process neurons learn stochastic activation functions. *arXiv preprint arXiv:1711.11059*, 2017. doi: 10.48550/arXiv.1711.11059.