Variational Distributional Neurons for Measurable Internal Uncertainty in Language Models

Yves Ruffenach
Conservatoire National des Arts et Métiers
Strasbourg, France
ORCID: 0009-0009-4737-0555
DOI: 10.5281/zenodo.21669216
Correspondence: yves@ruffenach.net

Abstract

Large language-model research increasingly relies on scale, which can make mechanism isolation difficult. We study the complementary small-scale regime as a controlled setting for architectural discovery. We examine EVE, a variational distributional neuron that replaces selected deterministic hidden activations with input-conditioned Gaussian latent states and exposes their local mean, variance, and KL activity during computation. We evaluate a 1% WritingPrompts-Filtered language-modeling benchmark with five matched seeds using a reproducible experimental pipeline. Relative to the deterministic baseline, EVE reduces mean test NLL by 4.63% (5.359 ± 0.039 vs. 5.620 ± 0.022), reduces perplexity by 22.88%, increases next-token accuracy by 11.93%, and improves CVaR-NLL by 3.69%; it wins all five paired seeds on NLL, accuracy, and CVaR-NLL. The paired NLL difference is -0.260 with an exploratory 95% interval of $[-0.330, -0.191]$, while mean calibration error remains comparable. Under covariate non-stationarity, reported aggregate NLL rises from 5.29 to 8.06 as context-token corruption increases from 0 to 0.30, while internal KL rises from 0.274 to 0.362. Validation-selected readouts further expose a risk-sensitive readout gap: on a representative run, an ECE-selected readout preserves NLL while reducing CVaR-NLL from 12.84 to 12.08, and a balanced readout increases accuracy to 0.170 while reducing CVaR-NLL to 11.94. These results establish EVE as a small-scale mechanism for measurable internal distributional activity and motivate parameter-matched and larger-scale evaluations.

Keywords: variational distributional neurons; internal uncertainty; probabilistic neural computation; uncertainty quantification; language models; distribution shift; risk-sensitive inference

1 Introduction

Current language-model research increasingly derives performance from scale. Large datasets, wider architectures, and longer training improve predictive capabilities, while controlled small-scale experiments provide a complementary setting for isolating the effect of individual computational mechanisms. They make repeated seeds, architectural ablations, stress-test sweeps, and direct inspection of hidden states practical.

An artificial neuron, or computational unit, receives an input representation, applies a learned transformation and usually a nonlinear function, and returns an activation. For fixed parameters and a fixed input, a conventional neuron returns one point-valued activation. Modern neural networks organize large numbers of these activations into layers. In Transformers, they form the coordinates of token representations and are repeatedly transformed by self-attention, feed-forward blocks, residual connections, and normalization operations. Although the final language-model output is a probability distribution over tokens, the intermediate hidden representations are generally point-valued vectors.

EVE, short for *Elemental Variational Expanse*, implements a variational distributional neuron. The term *distributional* means that the unit represents a probability distribution over possible activation states rather than a single activation value. The term *variational* means that this distribution is learned as an input-conditioned approximate posterior and regulated relative to a reference distribution, called the prior, through a variational objective.

For each input, an EVE unit predicts a mean and a variance that define a local Gaussian posterior over the activation. A latent activation is then drawn through reparameterization, a differentiable sampling procedure that enables the distribution parameters to be learned by gradient-based optimization. A Kullback–Leibler divergence, abbreviated KL, measures the displacement of the learned posterior from its prior and contributes a local regularization term to the global training objective.

The activation therefore becomes a measurable probabilistic state. Its posterior mean describes its central activation, its variance describes the spread of plausible activation states, and its KL activity measures the strength of its input-dependent departure from the prior. These quantities remain available during the forward computation and provide internal signals for inspection, stress testing, and inference-time decision making.

Within a Transformer, selected point-valued hidden coordinates can be replaced by EVE units while preserving the surrounding attention and feed-forward structure. The resulting architecture propagates sampled latent states together with measurable posterior statistics. This creates an internal representation of uncertainty at the level of the computational unit and complements the probability distribution produced over tokens at the model output.

This formulation has several potential uses. Internal posterior statistics can be tracked under distribution shift, meaning a change between the data encountered during training and the data processed at inference time. They can also support different posterior readouts, where a readout is the rule used to convert repeated samples or posterior statistics into a final prediction. Such readouts can emphasize average likelihood, calibration, predictive accuracy, or tail risk, defined here as the predictive loss concentrated among the most difficult examples.

The present study evaluates this mechanism in a controlled language-modeling setting. The experimental framework combines (i) a small-scale protocol for distributional neurons; (ii) internal diagnostics based on latent variance, local KL, mutual-information proxies, and top-1 sample instability; (iii) five-seed comparisons with deterministic and Monte Carlo dropout baselines; and (iv) a self-contained numerical record covering experimental configuration, seedwise outcomes, paired effects, posterior-readout searches, stress-test summaries, uncertainty intervals, and compute accounting.

2 Related Work and Conceptual Lineage

2.1 The point-valued neuron as the dominant computational primitive

The idea of a neural computational unit has taken several forms since the logical threshold elements of McCulloch and Pitts (McCulloch and Pitts, 1943), the perceptron (Rosenblatt, 1958), and the development of multilayer networks trained by back-propagation (Rumelhart et al., 1986). Modern convolutional networks (LeCun et al., 1998) and Transformers (Vaswani et al., 2017) are much larger and structurally richer. Their basic hidden computation is still usually point-valued: for fixed parameters and a fixed input, a unit or hidden coordinate returns one deterministic activation. Randomness may enter through data sampling, optimization, dropout, decoding, or an explicitly probabilistic output head, while the ordinary hidden unit carries a single activation state.

This observation characterizes the dominant primitive used in contemporary feed-forward and Transformer computation. It also separates uncertainty associated with parameters, outputs, masks, or model ensembles from uncertainty represented directly by the hidden computational state.

2.2 Stochastic units and local posterior structure

Stochastic neural computation has a long history. Boltzmann machines use stochastic binary units whose collective states define an energy-based probability distribution (Ackley et al., 1985). Later work developed deep generative stochastic networks and other architectures with stochastic hidden variables (Bengio et al., 2014). These models established randomness as a direct component of neural computation.

Dropout is the most widely used modern example of activation-level stochasticity. During training, Bernoulli masks randomly remove units or connections, primarily as a regularizer against co-adaptation (Srivastava et al., 2014). Monte Carlo dropout retains this stochastic masking at inference and averages repeated predictions; under specific assumptions it can be interpreted as approximate variational inference in a Bayesian neural network (Gal and Ghahramani, 2016). A dropout unit derives its random state from a mask distribution. A variational distributional unit instead constructs an input-conditioned posterior $q(z \mid x)$ with learned mean and variance and exposes a unit-specific prior-to-posterior divergence.

The same distinction applies to generic noise injection. A hidden value of the form $h + \epsilon$ is stochastic. Variational structure adds an explicit approximate posterior, a prior or reference distribution, differentiable sampling, and an inference objective that regulates the posterior. In this sense, "stochastic" describes how a value is produced, whereas "variational distributional" describes the probabilistic object that the unit constructs and optimizes.

From Boltzmann units to Gaussian Process Neurons. Two earlier neuron-level lineages locate EVE precisely. In a Boltzmann machine, each unit is a stochastic binary state whose conditional activation probability is determined by the states of the other units and the network energy; learning adjusts symmetric couplings so that equilibrium correlations under the model approach those observed with the data clamped (Ackley et al., 1985; Hinton and Sejnowski, 1986). The stochastic unit therefore acts as a sampling mechanism inside a globally coupled, undirected energy-based model. Its stochasticity belongs to global state dynamics, whereas EVE amortizes for each input a posterior $q_{\phi,i}(z_i \mid x_i)$ with separately learned mean and variance and contributes an input-dependent KL term attributable to the current state of an individual unit. Gaussian Process Neurons (GPNs) provide another probabilistic neuron-level primitive trained with variational Bayesian inference (Urban et al., 2017). Their principal random object is the activation function: a Gaussian-process prior is placed over the transfer function of each neuron, and the data induce a posterior over that function, represented in scalable variants through virtual observations or inducing variables (Rasmussen and Williams, 2006; Titsias, 2009; Urban et al., 2017). Consequently, uncertainty at a preactivation value arises by evaluating an uncertain function, and the GP kernel can correlate function values across inputs or examples. Urban et al. derive analytical propagation of means and covariances together with a deterministic variational training objective. EVE places the principal random object in the current hidden state: its mappings $x \mapsto \mu_\phi(x)$ and $x \mapsto \sigma_\phi(x)$ are ordinary parametric functions, while z_i is the latent variable. For every input, the unit amortizes a local Gaussian posterior, draws a reparameterized state that directly participates in computation, and contributes a locally measurable $\mathrm{KL}(q_{\phi,i}(z_i \mid x_i) \| p_i(z_i))$ term. Its formulation uses ordinary parametric mappings and local Gaussian posteriors in place of GP kernels and inducing observations. The resulting distinction separates *stochastic binary state dynamics* in Boltzmann machines, *posterior uncertainty over the activation function* in GPNs, and *input-conditioned posterior uncertainty over the current activation state* in EVE.

2.3 Bayesian neural networks: uncertainty over parameters and functions

Bayesian neural networks place distributions over weights or, more generally, over the functions induced by those weights. Practical variational methods approximate a posterior over network

parameters and optimize a variational free-energy or evidence-lower-bound objective (Graves, 2011; Blundell et al., 2015). Other scalable approaches propagate approximate distributions through the network or estimate posterior moments over weights (Hernández-Lobato and Adams, 2015). These methods are central to epistemic uncertainty: different plausible parameter settings induce different predictions.

A Bayesian network can therefore produce stochastic hidden activations when sampled weights are propagated forward. The primary inferential object remains the parameter or function posterior. The local reparameterization trick makes this distinction especially clear: global weight uncertainty can be transformed into data-point-specific activation noise to reduce gradient variance, while the learned posterior is still a posterior over weights (Kingma et al., 2015). In contrast, an EVE unit directly parameterizes a posterior over its local latent activation. Its weights may remain point estimates, while the activation state itself is the inferred random variable.

This architectural distinction supports a direct combination of Bayesian weight uncertainty and distributional activation uncertainty. They answer different questions: a Bayesian neural network asks which parameters or functions remain plausible after observing data; an EVE unit asks which local activation states remain plausible for the present input and how strongly that local posterior departs from its prior.

Deep ensembles provide another strong route to predictive uncertainty by aggregating independently trained deterministic networks (Lakshminarayanan et al., 2017), while post-hoc calibration changes the relationship between output confidence and empirical correctness (Guo et al., 2017). These methods provide valuable baselines while preserving the semantics of an individual hidden unit. Likewise, the usual distinction between epistemic and aleatoric uncertainty concerns sources of predictive uncertainty (Kendall and Gal, 2017). Internal EVE variance and KL are therefore treated as measurable internal distributional activity, with epistemic or aleatoric interpretation reserved for an explicit decomposition.

2.4 Variational autoencoders and stochastic latent-state sequence models

The closest methodological lineage comes from amortized variational inference. Variational autoencoders introduced a practical combination of an input-conditioned approximate posterior, a prior, reparameterized sampling, and an evidence lower bound (Kingma and Welling, 2014; Rezende et al., 2014). The encoder predicts the parameters of a distribution $q_\phi(z \mid x)$, samples a latent state, and regularizes that state through a KL term. This structure provides the conceptual template for the distributional neuron.

A large literature has inserted variational latent variables into sequential models. Variational recurrent neural networks add latent random variables to recurrent hidden dynamics (Chung et al., 2015); stochastic recurrent neural networks combine deterministic recurrence with state-space latent variables and structured variational inference (Fraccaro et al., 2016). In language generation, variational Transformers have introduced global, segment-wise, token-wise, or layer-wise latent variables to model diversity and higher-level structure (Hu et al., 2022). These approaches demonstrate that latent uncertainty can shape sequence computation through variables attached to an example, time step, segment, layer, or global representation. EVE brings the same variational structure to the semantics of an individual computational unit.

The variational information bottleneck provides another close relation. It learns a stochastic representation using a reparameterized conditional distribution and a KL-based compression objective (Alemi et al., 2017). This can improve robustness and provide uncertainty-related signals through a designated representation layer. EVE introduces a finer granularity by treating selected hidden units themselves as local variational bottlenecks whose posterior statistics remain available throughout computation.

Approach	Primary random object	Explicit $q(\cdot \mid x)$	Location of KL term	Typical role
Deterministic unit	None	No	None	Point-valued feature computation
Boltzmann unit	Binary state in global energy model	No amortized local posterior	Global energy objective	Sampling a joint undirected distribution
Dropout / MC dropout	Binary masks	Usually no	None	Regularization / approximate predictive uncertainty
Bayesian neural network	Weights or functions	Usually no	Parameter-level	Epistemic uncertainty
Gaussian Process Neuron	Activation function	Function posterior	Function / inducing-variable level	Uncertain learned transfer function
VAE / variational sequence model	Global, layer, step, or segment latent	Yes	Latent-level	Generative representation and sequence variability
EVE unit	Current local activation state	Yes, at the unit	Unit-level contribution	Internal distributional computation and diagnostics

Table 1: Conceptual positioning of the variational distributional neuron. "Local KL" denotes a KL contribution attributable to the corresponding hidden unit; the complete optimization objective remains global.

2.5 Positioning the variational distributional neuron

EVE builds on stochastic neurons, Bayesian inference, latent variables, and KL regularization by packaging these elements into a reusable computational primitive. For an input x_i, a selected unit constructs $q_{\phi,i}(z_i \mid x_i)$, samples z_i by reparameterization, uses that sample in the forward computation, and contributes a local KL term $\mathrm{KL}(q_{\phi,i} \| p_i)$ to the global objective. The global training loss retains a locally decomposable variational regularizer, and posterior variance, mean energy, KL, sample disagreement, and related diagnostics expose the unit's internal regime.

The term *distributional* emphasizes that the unit carries a fully parameterized distribution as its computational state. The term *variational* emphasizes that this distribution is an amortized approximation trained relative to a prior through an ELBO-style objective. The unit extends stochastic activation structure by relocating the principal uncertainty object from weights or outputs to the hidden activation state. This formulation makes the individual hidden unit an explicit, measurable variational latent state and yields internal uncertainty signals during language-model computation.

3 Distributional neurons

A standard hidden unit maps an input representation x to a deterministic activation

$$h = f(Wx + b). \tag{1}$$

This point-valued computation represents the hidden state as a single activation. Output-level uncertainty methods such as Monte Carlo dropout (Gal and Ghahramani, 2016), ensembles (Lakshminarayanan et al., 2017), or calibration methods (Guo et al., 2017) complement this deterministic hidden computation.

EVE unit. Operationally, an EVE unit is a drop-in replacement for a deterministic hidden activation: it receives the same input representation and returns a stochastic latent activation whose distribution is parameterized by learned mean and variance projections. An EVE neuron replaces a point hidden activation with a local variational latent state. Given input representation x, the unit computes

$$\mu_\phi(x) = W_\mu x + b_\mu, \tag{2}$$

$$\log \sigma^2_\phi(x) = W_\sigma x + b_\sigma, \tag{3}$$

and defines

$$q_\phi(z \mid x) = \mathcal{N}(\mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x))). \tag{4}$$

During training, a latent activation is sampled by reparameterization (Kingma and Welling, 2014):

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

The objective combines language-modeling loss and a local variational regularizer,

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \beta \sum_l \mathrm{KL}(q_{\phi,l}(z_l \mid x_l) \| p(z_l)). \tag{6}$$

Internal probabilistic activity. EVE exposes internal observables produced during computation: $\mu_\phi(x)$, $\sigma^2_\phi(x)$, local KL, stochastic sample disagreement, mutual-information proxies, and top-1 sample instability. A trained distributional model can also be read out in several ways: posterior mean, Monte Carlo mean, robust median, confidence-weighted aggregation, or validation-selected policies. We call the difference between information contained in the trained posterior and information exposed by a naive mean readout the *readout gap*.

4 Materials and Methods

4.1 Dataset and preprocessing

We evaluate EVE on `RLAIF/WritingPrompts-Filtered` using the GPT-2 tokenizer and frozen GPT-2 token embeddings. We use a fixed 1% prefix slice of the dataset: 1,394 train, 299 validation, and 299 test prompt–story examples. Teacher-forcing windows are generated with context length 24 and target stride 2, producing 134,086 train windows, 28,662 validation windows, and 28,898 test windows. All main runs use four epochs, five seeds, batch size 48, and the same split. The best checkpoint is selected on validation cross-entropy before final test evaluation.

4.2 Models, baselines, and evaluation

The deterministic baseline has 20.7M trainable parameters plus 38.6M frozen embedding parameters, for 59.3M total parameters. EVE has 134.9M trainable parameters plus the same frozen embeddings, for 173.5M total parameters. The comparison is designed as a small-scale mechanism study with explicit parameter and compute reporting.

The evaluation compares deterministic training, MC dropout applied directly to matching deterministic checkpoints, a five-member deterministic ensemble, and raw EVE readout. MC dropout uses the deterministic checkpoint for each seed, keeps dropout modules active at evaluation time, and averages four stochastic forward passes with dropout probability $p = 0.05$ in the deterministic decoder. Primary metrics are test NLL, perplexity, next-token accuracy, expected calibration error (ECE), and tail-risk CVaR-NLL. Internal metrics include mutual information, top-1 flip rate, internal KL, latent σ, latent μ^2, and active-unit fraction. Non-stationarity tests include context-token corruption, target-frequency slices, context-length shifts, temporal partitions, and post-shift recalibration variants. The covariate-shift monitor evaluates up to six batches per condition and preserves identical scenario definitions across seeds.

4.3 Statistical reporting

Main results are expressed as mean ± sample standard deviation over five matched seeds. The shared seed indices, data split, and evaluation protocol for DET, MC Dropout, and EVE also support seedwise win counts and paired differences. Exploratory 95% intervals use a Student-t interval for the mean paired difference, and paired t-test p-values provide descriptive evidence within the five-seed design. Validation-selected readout analyses form a complementary single-run diagnostic analysis alongside the five-seed comparison.

4.4 Compute reporting

For budget transparency, the accounting uses the heuristic estimate $6 \times N_{\text{trainable}} \times N_{\text{train tokens}}$, where $N_{\text{train tokens}}$ is the number of teacher-forcing windows times context length times epochs. For EVE, an additional factor of four accounts for latent training samples. Single-run budgets are 1.83×10^{15} FLOPs for DET and 4.32×10^{16} FLOPs for EVE. These accounting estimates derive from model size, token volume, epochs, and sampling multiplicity.

AI-assisted manuscript preparation. AI-assisted tools were used for language editing, consistency checking, figure generation from exported numerical results, and document formatting. The author reviewed the resulting text, tables, figures, and references and takes full responsibility for the scientific content and final manuscript.

5 Results

5.1 Predictive performance across five seeds

Across five matched seeds, EVE improves every primary predictive-risk mean relative to both deterministic baselines. Against DET, mean test NLL decreases from 5.620 to 5.359, a 4.63% relative reduction; perplexity decreases from 275.8 to 212.7, a 22.88% reduction; next-token accuracy increases from 0.1482 to 0.1659, an 11.93% relative gain; and CVaR-NLL decreases from 13.200 to 12.713, a 3.69% reduction. EVE wins all five matched seeds on NLL, accuracy, and CVaR-NLL against both DET and MC Dropout. Mean ECE remains comparable at 0.0418 for EVE, 0.0426 for DET, and 0.0456 for MC Dropout.

5.2 Paired effects and internal probabilistic activity

The mean paired EVE-minus-DET NLL difference is -0.2602, with an exploratory 95% Student-t interval of $[-0.3297, -0.1906]$. The paired accuracy difference is $+0.01768$, with an interval of $[+0.01571, +0.01965]$. CVaR-NLL improves in all five seeds, with an exploratory interval of $[-1.002, 0.029]$. Across the same seeds, the sampling-based mutual-information proxy is 0.190 ± 0.011 for EVE and 0.007 ± 0.003 for MC Dropout; the top-1 sample flip rate is 0.324 ± 0.012 and 0.061 ± 0.020, respectively. These larger values reflect greater measurable sample diversity under the selected perturbations.

5.3 Internal response under distribution shift

Under context-token corruption, aggregate raw-EVE NLL rises from 5.29 in the base condition to 5.54, 6.24, and 8.06 at corruption rates 0.05, 0.15, and 0.30. Internal KL rises from 0.274 to 0.291, 0.302, and 0.362, while latent μ^2 rises from approximately 0.051 to 0.075 at the strongest corruption. Across covariate-shift deltas, internal-KL change correlates with NLL change at $r = 0.67$, and latent-μ^2 change correlates with NLL and CVaR-NLL changes at $r = 0.97$ and $r = 0.87$. These descriptive diagnostic associations quantify the internal response accompanying changes in predictive risk.

5.4 Readout gap and risk-sensitive extraction

On a representative test run, an ECE-selected readout preserves EVE test NLL to three decimals while reducing CVaR-NLL from 12.836 to 12.082. A balanced confidence-weighted readout increases accuracy from 0.166 to 0.170 and reduces CVaR-NLL to 11.940 with test NLL 5.445. A CVaR-selected readout prioritizes tail risk, reaching CVaR-NLL 10.456 with test NLL 7.657. These validation-selected findings complement the five-seed raw-EVE comparison by illustrating objective-dependent posterior extraction.

6 Discussion and Research Perspectives

The five-seed results establish two complementary properties. Replacing selected point-valued activations with local variational states improves predictive performance in this controlled language-modeling regime: the NLL and accuracy effects are consistent across all matched seeds and have narrow exploratory paired intervals. The resulting unit also exposes local KL, posterior

scale, mean energy, mutual-information proxies, and sample instability as measurable internal quantities that evolve with the input distribution. The covariate dose response links changes in predictive risk to a measurable shift in the unit's internal variational regime.

The small-scale setting provides a controlled basis for mechanism isolation. The fixed 1% dataset slice, context length 24, four training epochs, and frozen embedding table define a precise experimental regime. These results provide a foundation for evaluation on larger corpora, longer contexts, additional tokenizers, and fully trained architectures. The validation trajectories also reveal distinct optimization and regularization dynamics between DET and EVE.

The present experiment prioritizes mechanism isolation, with EVE allocating approximately 6.5 times more trainable parameters and a larger heuristic FLOP budget to the construction and propagation of posterior states. This explicit capacity allocation motivates capacity-matched deterministic controls, latent-dimensionality ablations, and compute-matched scaling studies as the next experimental stage.

Five matched seeds reveal strong directional consistency across NLL, accuracy, and CVaR-NLL. CVaR-NLL improves in all five seeds, with an exploratory interval of $[-1.002, 0.029]$. The non-stationarity correlations provide descriptive measurements across the tested shift levels, and the validation-selected readouts demonstrate posterior flexibility on a representative run. Repeated-seed readout evaluations and denser shift sweeps form a natural extension of this evidence.

Internal variance, KL, mutual information, and flip rate provide diagnostic signals for monitoring and failure-mode analysis. Calibration against task-relevant outcomes can connect these internal observables to operational uncertainty categories, truthfulness criteria, and safety-oriented decision rules. Together, the results establish a measurable internal mechanism and a clear scaling trajectory.

7 Conclusion

We presented EVE as a small-scale testbed for variational distributional neurons in language models. Across five matched seeds, EVE reduces NLL and perplexity, increases next-token accuracy, and improves tail-risk CVaR-NLL relative to deterministic and MC-dropout baselines. More importantly for the architectural claim, the unit carries an explicit input-conditioned posterior whose KL and latent activity can be measured during computation and respond systematically to covariate degradation. Validation-selected readouts further show that different risk objectives can extract different information from the same frozen posterior. These results establish a controlled proof of mechanism for measurable internal uncertainty and motivate larger, parameter-matched, and multimodal studies of distributional neurons.

Ethics Statement

This study used a publicly available text dataset. Internal uncertainty measures serve as diagnostic signals for monitoring and failure-mode analysis, with operational safety interpretation grounded in task-specific calibration.

Author Contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization, and project administration: Y.R. The author has read and approved the manuscript.

Funding

This research received no external funding.

Data and Code Availability Statement

The experiments use the publicly available `RLAIF/WritingPrompts-Filtered` dataset. Appendices A.1–A.7 contain the aggregate numerical data supporting the empirical claims, tables, and figures, including complete primary seedwise metrics, training histories, readout-search values, non-stationarity summaries, bootstrap intervals, correlations, and compute estimates. This integrated numerical record supports direct inspection of the reported comparisons.

Conflicts of Interest

The author declares no conflicts of interest.

References

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943. doi:10.1007/BF02478259.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. doi:10.1037/h0042519.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi:10.1038/323533a0.

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.

Geoffrey E. Hinton and Terrence J. Sejnowski. Learning and relearning in Boltzmann machines. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 282–317. MIT Press, Cambridge, MA, 1986.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, 2009.

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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi:10.1109/5.726791.

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

Yoshua Bengio, Eric Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *Proceedings of the 31st International Conference on Machine Learning*, pages 226–234, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

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*, pages 1050–1059, 2016.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, 2011.

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

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1861–1869, 2015.

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

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

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*, pages 1278–1286, 2014.

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

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

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

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pages 159–168, 2018.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. Fuse it more deeply! A variational Transformer with layer-wise latent variable inference for text generation. In *Proceedings of NAACL-HLT*, pages 697–716, 2022. doi:10.18653/v1/2022.naacl-main.51.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. 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 *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

A Complete Numerical Appendix

The numerical record includes the experimental configuration, complete primary seedwise results, paired contrasts, training histories, validation-selected readout searches, non-stationarity summaries, bootstrap intervals, internal–external correlations, and compute accounting.

A.1 Experimental configuration

Table 2: Data construction and shared experimental settings.

Setting	Value
Dataset	RLAIF/WritingPrompts-Filtered
Tokenizer	gpt2
Dataset fraction	0.01
Validation fraction	0.15
Test fraction	0.15
Maximum prompt tokens	96
Maximum story tokens	192
Minimum story tokens	32
Context length	24
Target stride	2
Batch size	48
Training epochs	4
Weight decay	0.0001
Gradient clipping norm	1.0
Automatic mixed precision	True
AMP dtype	float16
Data-loader workers	0
Train / validation / test examples	1,394 / 299 / 299
Train / validation / test windows	134,086 / 28,662 / 28,898
Training seeds	0, 1, 2, 3, 4
Checkpoint selection	Minimum validation cross-entropy
Primary evaluation samples	4 Monte Carlo samples
Non-stationarity batches per condition	Up to 6
Non-stationarity bootstrap iterations	200

Table 3: Model architecture, optimization, and evaluation settings.

Setting	Value
DET trainable parameters	20,715,265
DET frozen embedding parameters	38,597,376
DET total parameters	59,312,641
EVE trainable parameters	134,852,868
EVE frozen embedding parameters	38,597,376
EVE total parameters	173,450,244
Deterministic learning rate	0.000200
EVE learning rate	0.000150
Decoder width / depth	1,024 / 1
Decoder dropout	0.05
Decoder gating	enabled
Transformer heads	12
Attention dropout	0.05
Residual dropout	0.05
Positional embeddings	enabled
MC Dropout probability	0.05
MC Dropout evaluation samples	4
Deterministic ensemble members	5

Table 4: Variational-unit and local-control settings used for the EVE runs.

EVE-specific setting	Value
Distributional units	1024
Latent dimension per unit	1
Posterior scale mode	per_unit_input
Initial posterior scale	0.8
Latent samples during training	4
Latent samples during extended evaluation	12
Latent aggregation	mean
Initial KL coefficient	0.0001
Initial local reconstruction coefficient	0.03
Local reconstruction ramp steps	400
Second-phase start epoch	3
Second-phase maximum posterior scale	5.8
Second-phase local reconstruction coefficient	0.022
Activity-band coefficient	0.004
Mean-energy target	0.1
Local control mode	homeo
Adaptive KL thermostat	True
Neuron activity regulator	True
Minimum adaptive KL coefficient	1e-06
Maximum adaptive KL coefficient	0.1
Minimum KL target	0.05
Maximum KL target	0.3
Posterior scale floor	0.15
Posterior scale ceiling	5.0
Control warm-up steps	200
Control update interval	10
Posterior mean clip	12.0
Minimum log-scale	-6.0
Maximum log-scale	2.0
Latent-state clip	12.0
Logit clip	50.0

A.2 Parameter and compute accounting

Operation	Trainable	Frozen	Total	Ctx.	Train windows	Epochs	Train mult.	Eval mult.	Train FLOPs	Eval FLOPs	Total FLOPs
eve	134,852,868	38,597,376	173,450,244	24	134,086	4	4	4	4.166e+16	1.490e+15	4.315e+16
det	20,715,265	38,597,376	59,312,641	24	134,086	4	1	4	1.600e+15	2.289e+14	1.829e+15
mc dropout	20,715,265	38,597,376	59,312,641	24	134,086	0	1	4	0.000e+00	2.289e+14	2.289e+14
ensemble det	20,715,265	38,597,376	59,312,641	24	134,086	0	1	4	0.000e+00	2.289e+14	2.289e+14

Table 5: Single-run parameter counts and heuristic FLOP accounting. The estimates cover trainable-parameter computation; tokenization, data loading, and frozen embedding lookup lie outside the accounting scope.

A.3 Complete five-seed primary results

Model	Seed	Best epoch	Val. CE	Test CE	MC NLL	PPL	Accuracy	ECE	CVaR-NLL
DET	0	1	5.6224	5.6352	5.5925	268.40	0.1469	0.0460	12.8893
DET	1	1	5.6227	5.6246	5.6140	274.25	0.1483	0.0312	13.1597
DET	2	2	5.6171	5.6204	5.6355	280.19	0.1503	0.0489	13.6673
DET	3	1	5.6505	5.6417	5.6090	272.86	0.1471	0.0422	13.0128
DET	4	1	5.6223	5.6257	5.6473	283.53	0.1486	0.0446	13.2705
MC Dropout	0	1	5.8466	5.5952	5.5952	269.14	0.1476	0.0479	12.9153
MC Dropout	1	1	5.8678	5.6153	5.6153	274.60	0.1441	0.0457	13.1835
MC Dropout	2	2	5.9276	5.6303	5.6303	278.74	0.1615	0.0601	13.6603
MC Dropout	3	1	5.8989	5.6116	5.6116	273.58	0.1476	0.0300	13.0119
MC Dropout	4	1	5.8808	5.6516	5.6516	284.75	0.1667	0.0441	13.2919
EVE raw	0	3	5.3757	5.3881	5.3918	219.60	0.1657	0.0552	12.8365
EVE raw	1	3	5.3661	5.3780	5.4005	221.52	0.1636	0.0413	12.3744
EVE raw	2	3	5.3688	5.3816	5.3632	213.40	0.1672	0.0331	12.6350
EVE raw	3	3	5.3734	5.3927	5.3354	207.56	0.1653	0.0355	12.8556
EVE raw	4	3	5.3611	5.3525	5.3065	201.63	0.1678	0.0439	12.8652

Table 6: Complete seedwise predictive metrics used in the primary comparison.

Model	Seed	MI	Cond. var.	Top-1 flip	Internal KL	KL/unit	σ	μ^2	Active frac.
DET	0	3.7770e-08	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DET	1	2.8250e-08	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DET	2	3.7046e-08	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DET	3	3.1872e-08	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DET	4	2.6698e-08	0.000000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MC Dropout	0	0.0057	8.856533e-05	0.0503	0.0000	0.0000	0.0000	0.0000	0.0000
MC Dropout	1	0.0060	5.366920e-05	0.0534	0.0000	0.0000	0.0000	0.0000	0.0000
MC Dropout	2	0.0130	0.000121	0.0959	0.0000	0.0000	0.0000	0.0000	0.0000
MC Dropout	3	0.0057	4.796186e-05	0.0495	0.0000	0.0000	0.0000	0.0000	0.0000
MC Dropout	4	0.0054	7.667884e-05	0.0560	0.0000	0.0000	0.0000	0.0000	0.0000
EVE raw	0	0.1834	0.001700	0.3082	0.2878	0.2878	0.8757	0.0551	1.0000
EVE raw	1	0.2075	0.001518	0.3338	0.2568	0.2568	0.8633	0.0422	1.0000
EVE raw	2	0.1893	0.001652	0.3377	0.2696	0.2696	0.8726	0.0499	1.0000
EVE raw	3	0.1941	0.002273	0.3194	0.2869	0.2869	0.8715	0.0513	1.0000
EVE raw	4	0.1779	0.001527	0.3194	0.2990	0.2990	0.8761	0.0570	1.0000

Table 7: Complete seedwise sampling-based and internal-activity metrics. Deterministic methods use fixed hidden states, so their variational-posterior fields are set to zero.

Model	n	NLL	Perplexity	Accuracy	ECE	CVaR-NLL	MI	Top-1 flip
DET	5	5.6196 ± 0.0218	275.84 ± 6.02	0.1482 ± 0.0014	0.0426 ± 0.0068	13.1999 ± 0.2985	0.0000 ± 0.0000	0.0000 ± 0.0000
EVE raw	5	5.3595 ± 0.0392	212.74 ± 8.29	0.1659 ± 0.0017	0.0418 ± 0.0087	12.7133 ± 0.2118	0.1904 ± 0.0113	0.3237 ± 0.0120
MC Dropout	5	5.6208 ± 0.0213	276.16 ± 5.89	0.1535 ± 0.0099	0.0456 ± 0.0107	13.2126 ± 0.2899	0.0072 ± 0.0033	0.0610 ± 0.0197

Table 8: Five-seed means and sample standard deviations.

Table 9: Exact paired differences, defined as EVE raw minus the stated baseline.

Seed	Comparison	Δ NLL	Δ Acc.	Δ ECE	Δ CVaR
0	EVE raw - DET	-0.2007	0.0188	0.0092	-0.0528
0	EVE raw - MC Dropout	-0.2034	0.0181	0.0073	-0.0788
1	EVE raw - DET	-0.2135	0.0153	0.0101	-0.7853
1	EVE raw - MC Dropout	-0.2148	0.0195	-0.0044	-0.8091
2	EVE raw - DET	-0.2723	0.0169	-0.0157	-1.0323
2	EVE raw - MC Dropout	-0.2671	0.0057	-0.0269	-1.0253
3	EVE raw - DET	-0.2735	0.0182	-0.0066	-0.1572
3	EVE raw - MC Dropout	-0.2762	0.0178	0.0056	-0.1563
4	EVE raw - DET	-0.3409	0.0192	-0.0007	-0.4053
4	EVE raw - MC Dropout	-0.3452	0.0012	-0.0003	-0.4267

Table 10: Paired EVE-minus-DET effect estimates. Intervals and p-values are exploratory because $n = 5$.

Metric	Mean difference	Exploratory 95% interval	Relative improvement	Wins	Paired p
NLL	-0.26017	[-0.32973, -0.19061]	4.63%	5/5	0.0005
Perplexity	-63.10175	[-79.33602, -46.86748]	22.88%	5/5	0.0004
Accuracy	0.01768	[0.01571, 0.01966]	11.93%	5/5	<0.0001
ECE	-0.00075	[-0.01430, 0.01279]	1.77%	3/5	0.8847
CVaR-NLL	-0.48657	[-1.00238, 0.02924]	3.69%	5/5	0.0589

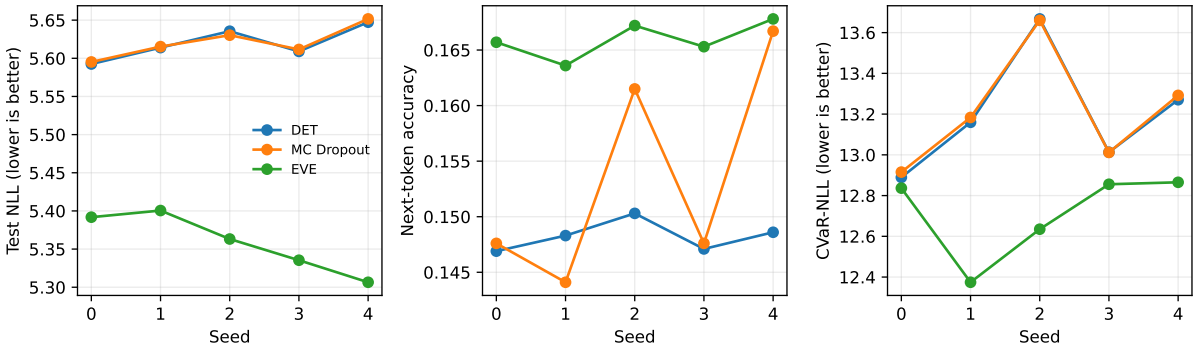

Figure 1: Matched-seed NLL, accuracy, and CVaR-NLL values.

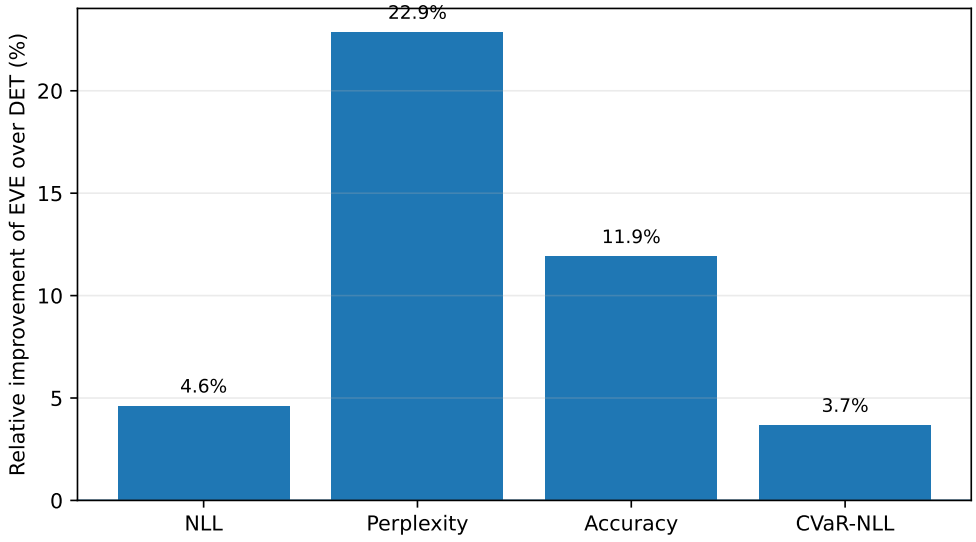

Figure 2: Relative improvements of EVE over DET computed from five-seed means.

A.4 Training and validation trajectories

Table 11: Per-epoch training and validation histories for the ten trained primary models.

Model	Seed	Epoch	Train CE	Train PPL	Train acc.	Val. CE	Val. PPL	Val. acc.	Val. KL	Val. μ^2
DET	0	1	5.9160	370.94	0.1346	5.6224	276.55	0.1534	0.0000	1.0273
DET	0	2	5.0155	150.73	0.1916	5.6307	278.85	0.1570	0.0000	0.5896
DET	0	3	3.7518	42.60	0.3060	6.2363	510.94	0.1334	0.0000	0.2747
DET	0	4	2.1842	8.88	0.5276	7.4492	1718.44	0.1265	0.0000	0.2332
DET	1	1	5.9144	370.34	0.1353	5.6227	276.64	0.1532	0.0000	1.0005
DET	1	2	5.0187	151.21	0.1918	5.6230	276.72	0.1542	0.0000	0.5987
DET	1	3	3.7703	43.39	0.3060	6.1667	476.59	0.1357	0.0000	0.2758
DET	1	4	2.2054	9.07	0.5265	7.4830	1777.51	0.1163	0.0000	0.2229
DET	2	1	5.9166	371.15	0.1339	5.6327	279.42	0.1509	0.0000	0.9704
DET	2	2	5.0192	151.28	0.1915	5.6171	275.09	0.1565	0.0000	0.6120
DET	2	3	3.7624	43.05	0.3039	6.3985	600.96	0.1392	0.0000	0.2504
DET	2	4	2.2014	9.04	0.5245	7.2689	1435.03	0.1213	0.0000	0.2252
DET	3	1	5.9138	370.09	0.1347	5.6505	284.44	0.1523	0.0000	1.0061
DET	3	2	5.0098	149.87	0.1923	5.6697	289.94	0.1577	0.0000	0.5992
DET	3	3	3.7350	41.89	0.3082	6.2664	526.60	0.1318	0.0000	0.2686
DET	3	4	2.1706	8.76	0.5309	7.2562	1416.82	0.1199	0.0000	0.2381
DET	4	1	5.9199	372.37	0.1354	5.6223	276.52	0.1549	0.0000	0.9931
DET	4	2	5.0255	152.25	0.1915	5.6452	282.93	0.1552	0.0000	0.5940
DET	4	3	3.7978	44.60	0.3020	6.2524	519.24	0.1357	0.0000	0.2595
DET	4	4	2.2512	9.50	0.5192	7.4256	1678.36	0.1206	0.0000	0.2366
EVE	0	1	5.8208	337.24	0.1381	5.5165	248.76	0.1609	317.2654	0.0241
EVE	0	2	5.4033	222.14	0.1661	5.4170	225.21	0.1689	313.3796	0.0469
EVE	0	3	5.1223	167.72	0.1858	5.3748	215.90	0.1693	273.1274	0.0523
EVE	0	4	4.7267	112.93	0.2179	5.4704	237.56	0.1668	220.0878	0.0394
EVE	1	1	5.8103	333.71	0.1382	5.5112	247.44	0.1595	340.1704	0.0336
EVE	1	2	5.4060	222.74	0.1659	5.3959	220.49	0.1675	276.9488	0.0434
EVE	1	3	5.1204	167.40	0.1869	5.3634	213.46	0.1737	222.2357	0.0455
EVE	1	4	4.7236	112.57	0.2172	5.4888	241.97	0.1709	226.3362	0.0513
EVE	2	1	5.8077	332.84	0.1394	5.5335	253.04	0.1588	342.1417	0.0345
EVE	2	2	5.4041	222.32	0.1670	5.4052	222.57	0.1717	282.4316	0.0316
EVE	2	3	5.1276	168.61	0.1863	5.3671	214.25	0.1758	254.9025	0.0574
EVE	2	4	4.7487	115.44	0.2157	5.4549	233.90	0.1709	225.3726	0.0541
EVE	3	1	5.8167	335.88	0.1385	5.5211	249.90	0.1603	329.3686	0.0233
EVE	3	2	5.4023	221.91	0.1657	5.4018	221.80	0.1680	277.1194	0.0341
EVE	3	3	5.1079	165.32	0.1866	5.3771	216.40	0.1711	267.3158	0.0804
EVE	3	4	4.6861	108.43	0.2200	5.5045	245.81	0.1690	238.0206	0.0837
EVE	4	1	5.8191	336.68	0.1398	5.5075	246.54	0.1634	332.1174	0.0299
EVE	4	2	5.3977	220.90	0.1662	5.4132	224.35	0.1681	312.4995	0.0345
EVE	4	3	5.1109	165.81	0.1864	5.3628	213.32	0.1716	285.7748	0.0582
EVE	4	4	4.7024	110.21	0.2197	5.4918	242.70	0.1687	224.4521	0.0557

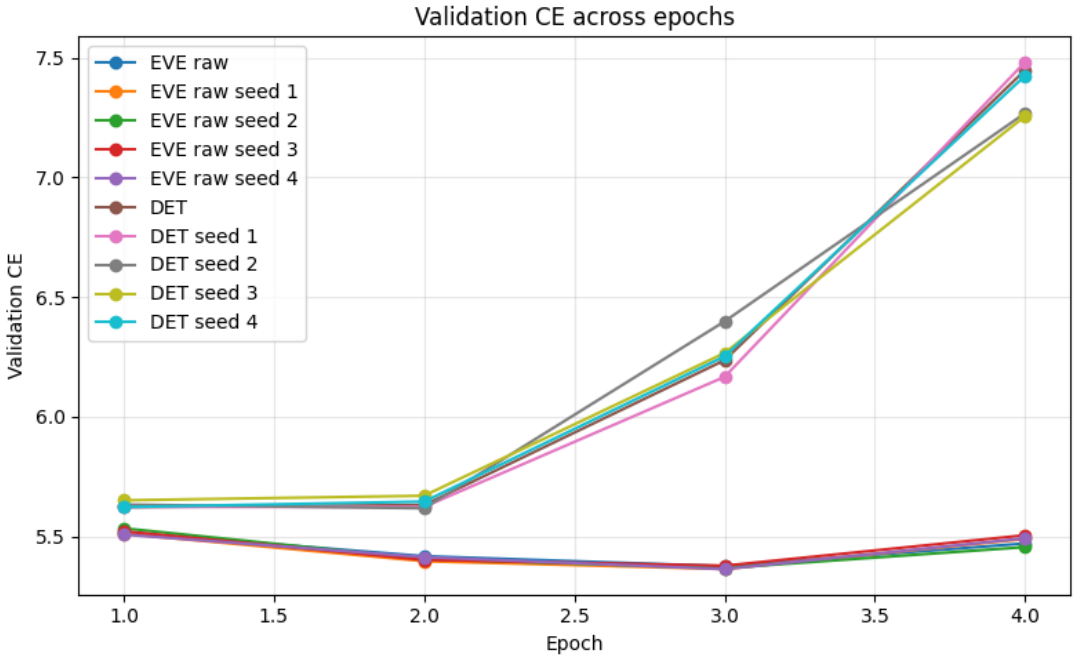

Figure 3: Validation cross-entropy trajectories for the five DET and five EVE runs.

A.5 Posterior readout analysis

Model / readout	Test NLL	PPL	Accuracy	ECE	MI	Top-1 flip	CVaR-NLL	KL	σ	μ^2	Readout
DET	5.5925	268.40	0.1469	0.0460	3.7770e-08	0.0000	12.8893	0.0000	0.0000	0.0000	–
MC Dropout	5.5952	269.14	0.1476	0.0479	0.0057	0.0503	12.9153	0.0000	0.0000	0.0000	–
EVE raw	5.3918	219.60	0.1657	0.0552	0.1834	0.3082	12.8365	0.2878	0.8757	0.0551	–
EVE-ECE-selected	5.3923	219.71	0.1615	0.0498	0.1699	0.3064	12.0820	0.2853	0.8759	0.0551	mean_prob
EVE-CVaR-selected	7.6571	2115.71	0.1615	0.1571	0.0329	0.3477	10.4563	0.2888	0.8759	0.0552	median_prob
EVE-balanced	5.4453	231.66	0.1701	0.0617	0.0908	0.3177	11.9398	0.2908	0.8759	0.0553	confidence_weighted
DET Ensemble	5.4545	233.81	0.1632	0.0466	0.1526	0.2920	12.7605	0.0000	0.0000	0.0000	–

Table 12: Representative-run test metrics for deterministic, MC-dropout, ensemble, raw EVE, and validation-selected EVE readouts.

Policy	Objective	Temp.	Readout	Selection score	Val. NLL	Val. acc.	Val. ECE	Val. CVaR
EVE-ECE-selected	ece	1.10	mean_prob	0.1010	5.6551	0.1684	0.0373	12.8643
EVE-CVaR-selected	cvar	3.00	median_prob	11.1050	7.7108	0.1701	0.1652	10.6782
EVE-balanced	balanced	1.20	confidence_weighted	6.0862	5.6525	0.1788	0.0488	12.3741

Table 13: Validation-selected EVE policies and their validation metrics.

Table 14: Complete 60-point validation search over temperatures and posterior readout functions. The three score columns are the objective values used for the ECE-first, CVaR-first, and balanced selections.

Temp.	Readout	NLL	Acc.	ECE	CVaR	MI	Flip	KL	σ	μ^2	ECE score	CVaR score	Balanced score
0.80	confidence_weighted	6.0375	0.1753	0.1797	15.9241	0.0843	0.3051	0.3121	0.8761	0.0609	0.2500	16.2709	6.8912
0.90	confidence_weighted	5.7699	0.1684	0.1184	14.4575	0.0861	0.3099	0.3020	0.8757	0.0606	0.1848	14.7756	6.4411
1.00	confidence_weighted	5.7136	0.1649	0.0726	13.6873	0.0863	0.3294	0.3180	0.8762	0.0608	0.1377	13.9912	6.2575
1.10	confidence_weighted	5.6667	0.1684	0.0465	13.0184	0.0889	0.3177	0.3030	0.8756	0.0607	0.1105	13.3134	6.1272
1.20	confidence_weighted	5.6525	0.1788	0.0488	12.3741	0.0897	0.3203	0.3109	0.8761	0.0609	0.1120	12.6690	6.0862
1.35	confidence_weighted	5.7910	0.1753	0.0685	11.9792	0.0892	0.3173	0.3211	0.8762	0.0609	0.1326	12.2859	6.2373
1.50	confidence_weighted	5.9498	0.1667	0.0895	11.6054	0.0866	0.3181	0.3045	0.8758	0.0607	0.1547	11.9253	6.4116
1.75	confidence_weighted	6.2626	0.1562	0.1066	11.1809	0.0784	0.3168	0.3352	0.8765	0.0614	0.1741	11.5206	6.7217
2.00	confidence_weighted	6.6053	0.1823	0.1517	10.9664	0.0688	0.3160	0.3039	0.8757	0.0607	0.2222	11.3346	7.1268
2.25	confidence_weighted	6.9194	0.1649	0.1447	10.8622	0.0599	0.3121	0.3187	0.8763	0.0609	0.2178	11.2443	7.4058
2.50	confidence_weighted	7.2090	0.1701	0.1587	10.7845	0.0496	0.3142	0.3148	0.8767	0.0611	0.2344	11.1846	7.7052
3.00	confidence_weighted	7.6795	0.1719	0.1666	10.7528	0.0366	0.3108	0.3185	0.8766	0.0611	0.2465	11.1784	8.1664
0.80	entropy_weighted	6.3899	0.1719	0.2430	17.1635	0.0043	0.3264	0.3061	0.8760	0.0608	0.3176	17.5438	7.4145
0.90	entropy_weighted	6.0923	0.1684	0.1755	16.0874	0.0045	0.3212	0.3152	0.8762	0.0608	0.2464	16.4359	6.9431
1.00	entropy_weighted	5.8685	0.1580	0.1216	14.8627	0.0029	0.3320	0.3080	0.8760	0.0607	0.1893	15.1865	6.5615
1.10	entropy_weighted	5.8263	0.1701	0.0685	14.5837	0.0021	0.3186	0.3161	0.8764	0.0609	0.1355	14.8921	6.4011
1.20	entropy_weighted	5.7815	0.1632	0.0412	13.7299	0.0027	0.3333	0.3141	0.8758	0.0608	0.1070	14.0293	6.2614
1.35	entropy_weighted	5.7598	0.1823	0.0558	12.5023	0.0026	0.3407	0.3085	0.8758	0.0607	0.1201	12.8043	6.2085
1.50	entropy_weighted	5.8687	0.1806	0.0813	11.9167	0.0021	0.3377	0.3051	0.8757	0.0606	0.1460	12.2305	6.3336
1.75	entropy_weighted	6.1676	0.1753	0.1091	11.5966	0.0043	0.3329	0.3086	0.8758	0.0608	0.1762	11.9322	6.6572
2.00	entropy_weighted	6.4822	0.1667	0.1282	11.0833	0.0049	0.3342	0.3039	0.8759	0.0606	0.1976	11.4395	6.9686
2.25	entropy_weighted	6.7908	0.1719	0.1479	11.0048	0.0061	0.3290	0.3142	0.8760	0.0608	0.2200	11.3813	7.2972
2.50	entropy_weighted	7.0941	0.1719	0.1576	10.8444	0.0070	0.3294	0.3021	0.8756	0.0606	0.2322	11.2385	7.5968
3.00	entropy_weighted	7.6057	0.1771	0.1708	10.7586	0.0077	0.3333	0.3114	0.8761	0.0608	0.2500	11.1816	8.1049
0.80	mean_prob	5.9590	0.1736	0.1334	15.7326	0.2120	0.3134	0.3088	0.8760	0.0607	0.2028	16.0639	6.7146
0.90	mean_prob	5.8062	0.1823	0.0803	14.8117	0.1964	0.3077	0.3047	0.8760	0.0607	0.1473	15.1221	6.4170
1.00	mean_prob	5.6910	0.1632	0.0645	13.7828	0.1834	0.3108	0.3072	0.8761	0.0609	0.1295	14.0835	6.2247
1.10	mean_prob	5.6551	0.1684	0.0373	12.8643	0.1678	0.3077	0.3145	0.8763	0.0608	0.1010	13.1563	6.0901
1.20	mean_prob	5.6834	0.1753	0.0488	12.3421	0.1593	0.3121	0.3045	0.8756	0.0605	0.1122	12.6385	6.1139
1.35	mean_prob	5.8076	0.1667	0.0680	11.7186	0.1385	0.3008	0.3226	0.8764	0.0610	0.1320	12.0260	6.2391
1.50	mean_prob	5.9720	0.1649	0.0922	11.5525	0.1208	0.3121	0.3058	0.8759	0.0607	0.1575	11.8742	6.4355
1.75	mean_prob	6.2806	0.1580	0.1134	11.1746	0.0967	0.3121	0.3028	0.8757	0.0606	0.1811	11.5170	6.7521
2.00	mean_prob	6.6257	0.1684	0.1395	10.9619	0.0795	0.3260	0.3203	0.8763	0.0608	0.2101	11.3281	7.1215
2.25	mean_prob	6.9195	0.1788	0.1603	10.8406	0.0641	0.3229	0.3137	0.8762	0.0609	0.2334	11.2267	7.4361
2.50	mean_prob	7.2109	0.1736	0.1627	10.7625	0.0539	0.3134	0.3123	0.8761	0.0607	0.2383	11.1637	7.7138
3.00	mean_prob	7.6930	0.1719	0.1669	10.7102	0.0381	0.3207	0.3093	0.8756	0.0607	0.2469	11.1366	8.1777
0.80	median_prob	6.0919	0.1788	0.1746	16.9215	0.0508	0.3464	0.3161	0.8762	0.0610	0.2464	17.2697	6.9826
0.90	median_prob	5.8665	0.1719	0.1221	15.3444	0.0471	0.3299	0.3099	0.8758	0.0607	0.1902	15.6682	6.5846
1.00	median_prob	5.7235	0.1736	0.0648	14.1818	0.0435	0.3112	0.3265	0.8761	0.0609	0.1305	14.4842	6.2760
1.10	median_prob	5.6904	0.1667	0.0447	13.6726	0.0607	0.3542	0.3338	0.8766	0.0611	0.1096	13.9683	6.1789
1.20	median_prob	5.6761	0.1753	0.0535	12.6694	0.0490	0.3316	0.3278	0.8767	0.0611	0.1173	12.9666	6.1329
1.35	median_prob	5.8053	0.1736	0.0706	12.1848	0.0524	0.3594	0.3221	0.8763	0.0608	0.1350	12.4927	6.2655
1.50	median_prob	5.9526	0.1788	0.1041	11.6556	0.0566	0.3442	0.3124	0.8761	0.0610	0.1693	11.9792	6.4459
1.75	median_prob	6.2898	0.1632	0.1154	11.2599	0.0611	0.3503	0.3149	0.8762	0.0607	0.1832	11.6033	6.7691
2.00	median_prob	6.6475	0.1823	0.1543	10.9751	0.0634	0.3468	0.3177	0.8765	0.0612	0.2251	11.3461	7.1724
2.25	median_prob	6.9702	0.1858	0.1676	10.8333	0.0525	0.3550	0.3215	0.8760	0.0607	0.2412	11.2237	7.4985
2.50	median_prob	7.2471	0.1736	0.1622	10.7845	0.0458	0.3351	0.3041	0.8758	0.0607	0.2382	11.1874	7.7483
3.00	median_prob	7.7108	0.1701	0.1652	10.6782	0.0328	0.3368	0.3170	0.8758	0.0606	0.2453	11.1050	8.1896
0.80	softmax_mean_logits	6.0425	0.1701	0.1649	16.4561	0.0673	0.3095	0.3088	0.8757	0.0607	0.2358	16.7995	6.8930
0.90	softmax_mean_logits	5.8084	0.1684	0.1134	15.1396	0.0603	0.3073	0.3278	0.8767	0.0610	0.1808	15.4584	6.5018
1.00	softmax_mean_logits	5.7008	0.1806	0.0566	14.2154	0.0579	0.3342	0.3172	0.8764	0.0609	0.1221	14.5146	6.2397
1.10	softmax_mean_logits	5.6627	0.1701	0.0406	13.2032	0.0602	0.3060	0.3082	0.8761	0.0609	0.1048	13.4965	6.1210

Continued on next page

Table 14 – continued

Temp.	Readout	NLL	Acc.	ECE	CVaR	MI	Flip	KL	σ	μ^2	ECE score	CVaR score	Balanced score
1.20	softmax_mean_logits	5.6864	0.1719	0.0379	12.6231	0.0647	0.3121	0.3115	0.8758	0.0606	0.1017	12.9169	6.1090
1.35	softmax_mean_logits	5.7715	0.1736	0.0695	11.8724	0.0650	0.3247	0.3052	0.8758	0.0607	0.1333	12.1783	6.2155
1.50	softmax_mean_logits	5.9405	0.1823	0.1056	11.5976	0.0668	0.3181	0.3077	0.8761	0.0607	0.1707	11.9210	6.4346
1.75	softmax_mean_logits	6.2964	0.1701	0.1226	11.1916	0.0694	0.2982	0.3132	0.8762	0.0609	0.1905	11.5371	6.7865
2.00	softmax_mean_logits	6.6448	0.1701	0.1415	11.0026	0.0652	0.3216	0.3184	0.8761	0.0608	0.2123	11.3703	7.1456
2.25	softmax_mean_logits	6.9347	0.1753	0.1570	10.8473	0.0572	0.3168	0.3069	0.8760	0.0608	0.2302	11.2333	7.4443
2.50	softmax_mean_logits	7.2405	0.1667	0.1559	10.7550	0.0506	0.3260	0.3085	0.8760	0.0608	0.2319	11.1560	7.7281
3.00	softmax_mean_logits	7.7071	0.1580	0.1535	10.7231	0.0366	0.3151	0.3038	0.8759	0.0608	0.2336	11.1468	8.1649

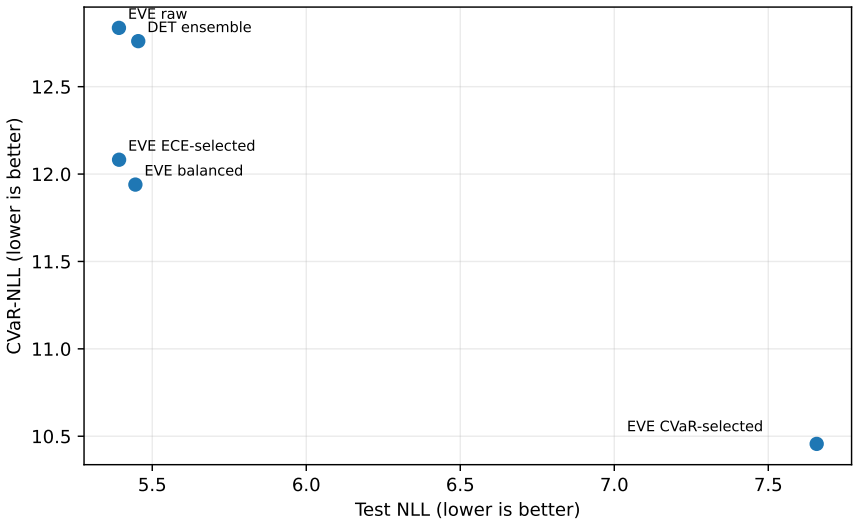

Figure 4: Representative-run NLL–CVaR trade-off among the selected readouts and the deterministic ensemble.

A.6 Multi-axis non-stationarity results

Table 15: Complete five-seed aggregate results for the base condition, context-length slices, covariate corruption, target-frequency slices, temporal partitions, and pre- or post-shift calibration/readout variants. Rows with $n < 5$ correspond to heterogeneous post-shift policy choices summarized by the selected readout.

Scenario	Model	n	NLL	NLL SD	Accuracy	ECE	CVaR	KL	σ	μ^2	Readout
Base test	DET	5	5.4782	0.0205	0.1590	0.0376	12.6848	0.0000	0.0000	0.0000	–
Base test	DET + temp. pre	5	5.4824	0.0211	0.1590	0.0412	12.5654	0.0000	0.0000	0.0000	temp.
Base test	EVE raw	5	5.2932	0.0406	0.1660	0.0599	12.3386	0.2737	0.8730	0.0512	raw
Base test	EVE balanced	5	5.4029	0.0641	0.1681	0.0750	11.4115	0.3043	0.8732	0.0518	conf.-weighted
Base test	MC Dropout	5	5.4779	0.0206	0.1618	0.0452	12.6798	0.0000	0.0000	0.0000	MC dropout
Base test	MC dropout + temp. pre	5	5.4879	0.0228	0.1556	0.0414	12.5784	0.0000	0.0000	0.0000	MC temp.
Long context	DET	5	5.3321	0.0121	0.1667	0.0611	12.4710	0.0000	0.0000	0.0000	–
Long context	DET + temp. pre	5	5.3318	0.0122	0.1667	0.0589	12.3464	0.0000	0.0000	0.0000	temp.
Long context	EVE raw	5	5.1107	0.0542	0.1840	0.0612	12.4387	0.2846	0.8740	0.0502	raw
Long context	EVE balanced	5	5.2016	0.0793	0.1840	0.0805	11.5279	0.2777	0.8742	0.0502	conf.-weighted
Long context	MC Dropout	5	5.3338	0.0143	0.1667	0.0668	12.4557	0.0000	0.0000	0.0000	MC dropout
Long context	MC dropout + temp. pre	5	5.3340	0.0141	0.1722	0.0659	12.3273	0.0000	0.0000	0.0000	MC temp.
Short context	DET	5	5.4764	0.0340	0.1375	0.0556	11.9887	0.0000	0.0000	0.0000	–
Short context	DET + temp. pre	5	5.4747	0.0348	0.1375	0.0582	11.8711	0.0000	0.0000	0.0000	temp.
Short context	EVE raw	5	5.1744	0.0105	0.1556	0.0552	12.0562	0.3373	0.8905	0.0648	raw
Short context	EVE balanced	5	5.2558	0.0197	0.1611	0.0618	11.3212	0.3370	0.8908	0.0650	conf.-weighted
Short context	MC Dropout	5	5.4807	0.0310	0.1382	0.0571	11.9768	0.0000	0.0000	0.0000	MC dropout
Short context	MC dropout + temp. pre	5	5.4772	0.0360	0.1375	0.0589	11.8600	0.0000	0.0000	0.0000	MC temp.
Covariate noise 0.05	DET	5	5.7035	0.0316	0.1507	0.0408	13.2150	0.0000	0.0000	0.0000	–
Covariate noise 0.05	DET + temp. pre	5	5.7044	0.0324	0.1507	0.0451	13.0740	0.0000	0.0000	0.0000	temp.
Covariate noise 0.05	EVE raw	5	5.5440	0.0378	0.1667	0.0585	13.0636	0.2910	0.8737	0.0522	raw
Covariate noise 0.05	EVE balanced	5	5.6210	0.0455	0.1556	0.0643	12.1927	0.3262	0.8740	0.0529	conf.-weighted
Covariate noise 0.05	MC Dropout	5	5.7023	0.0294	0.1542	0.0452	13.1959	0.0000	0.0000	0.0000	MC dropout
Covariate noise 0.05	MC dropout + temp. pre	5	5.7036	0.0343	0.1500	0.0487	13.0740	0.0000	0.0000	0.0000	MC temp.
Covariate noise 0.15	DET	5	6.3089	0.1140	0.1556	0.0597	16.0280	0.0000	0.0000	0.0000	–
Covariate noise 0.15	DET + temp. pre	5	6.2978	0.1061	0.1556	0.0629	15.8268	0.0000	0.0000	0.0000	temp.
Covariate noise 0.15	EVE raw	5	6.2373	0.0730	0.1431	0.0602	16.1177	0.3023	0.8784	0.0572	raw
Covariate noise 0.15	EVE balanced	5	6.1759	0.0528	0.1535	0.0792	14.4086	0.3094	0.8785	0.0571	conf.-weighted
Covariate noise 0.15	MC Dropout	5	6.3053	0.1147	0.1542	0.0573	15.9746	0.0000	0.0000	0.0000	MC dropout
Covariate noise 0.15	MC dropout + temp. pre	5	6.2899	0.1084	0.1569	0.0569	15.7604	0.0000	0.0000	0.0000	MC temp.
Covariate 0.15 post-cal.	DET + temp. post	5	6.2821	0.1108	0.1556	0.0675	14.9941	0.0000	0.0000	0.0000	temp.
Covariate 0.15 post-cal.	EVE shift CVaR	2	7.9284	0.0202	0.1476	0.1431	11.2057	0.3073	0.8791	0.0589	conf.-weighted
Covariate 0.15 post-cal.	EVE shift CVaR	2	8.0281	0.0703	0.1337	0.1312	11.1315	0.2972	0.8766	0.0540	mean-prob.
Covariate 0.15 post-cal.	EVE shift CVaR	1	8.0034	0.0000	0.1528	0.1503	11.1928	0.3149	0.8801	0.0602	mean-logits
Covariate 0.15 post-cal.	EVE shift ECE	3	6.1726	0.0406	0.1412	0.0556	15.1679	0.3132	0.8805	0.0597	mean-prob.
Covariate 0.15 post-cal.	EVE shift ECE	1	6.2420	0.0000	0.1354	0.0467	14.6432	0.2748	0.8682	0.0474	median-prob.
Covariate 0.15 post-cal.	EVE shift ECE	1	6.2758	0.0000	0.1493	0.0810	15.0677	0.3203	0.8823	0.0593	mean-logits
Covariate 0.15 post-cal.	EVE shift balanced	2	6.2466	0.0191	0.1424	0.0634	14.2441	0.2914	0.8755	0.0534	conf.-weighted
Covariate 0.15 post-cal.	EVE shift balanced	3	6.2120	0.0884	0.1609	0.0757	14.5571	0.3094	0.8806	0.0598	mean-prob.
Covariate 0.15 post-cal.	MC dropout + temp. post	5	6.2808	0.1093	0.1576	0.0629	15.0757	0.0000	0.0000	0.0000	MC temp.
Covariate noise 0.30	DET	5	8.1670	0.3643	0.1035	0.0925	18.5404	0.0000	0.0000	0.0000	–
Covariate noise 0.30	DET + temp. pre	5	8.1153	0.3289	0.1035	0.0868	18.2716	0.0000	0.0000	0.0000	temp.
Covariate noise 0.30	EVE raw	5	8.0573	0.1828	0.1028	0.0941	18.4830	0.3621	0.8994	0.0749	raw
Covariate noise 0.30	EVE balanced	5	7.7337	0.1471	0.1056	0.0728	16.4259	0.3660	0.8995	0.0747	conf.-weighted
Covariate noise 0.30	MC Dropout	5	8.1367	0.3642	0.1042	0.0916	18.4286	0.0000	0.0000	0.0000	MC dropout
Covariate noise 0.30	MC dropout + temp. pre	5	8.0936	0.3259	0.1056	0.0839	18.1625	0.0000	0.0000	0.0000	MC temp.
Covariate 0.30 post-cal.	DET + temp. post	5	7.8452	0.3829	0.1035	0.0534	16.0688	0.0000	0.0000	0.0000	temp.
Covariate 0.30 post-cal.	EVE shift CVaR	1	8.4921	0.0000	0.1111	0.1079	11.6614	0.3705	0.9120	0.0805	conf.-weighted
Covariate 0.30 post-cal.	EVE shift CVaR	3	8.5308	0.0476	0.1007	0.0973	11.5555	0.3601	0.8954	0.0710	mean-prob.
Covariate 0.30 post-cal.	EVE shift CVaR	1	8.5676	0.0000	0.1042	0.1010	11.7415	0.3987	0.9001	0.0814	median-prob.
Covariate 0.30 post-cal.	EVE shift ECE	1	7.5171	0.0000	0.1076	0.0684	15.0785	0.3793	0.8975	0.0762	conf.-weighted

Continued on next page

Table 15 – continued

Scenario	Model	n	NLL	NLL SD	Accuracy	ECE	CVaR	KL	σ	μ^2	Readout
Covariate 0.30 post-cal.	EVE shift ECE	1	7.8048	0.0000	0.0972	0.0465	15.9966	0.3432	0.8876	0.0630	entropy-weighted
Covariate 0.30 post-cal.	EVE shift ECE	2	7.7743	0.0576	0.1007	0.0584	16.2981	0.3900	0.9062	0.0809	mean-prob.
Covariate 0.30 post-cal.	EVE shift ECE	1	7.6948	0.0000	0.1181	0.0689	15.2002	0.4114	0.9011	0.0737	mean-logits
Covariate 0.30 post-cal.	EVE shift balanced	2	7.6950	0.0439	0.1059	0.0591	14.6524	0.3786	0.9057	0.0806	conf.-weighted
Covariate 0.30 post-cal.	EVE shift balanced	3	7.5682	0.0282	0.0984	0.0580	14.8546	0.3550	0.8952	0.0708	mean-prob.
Covariate 0.30 post-cal.	MC dropout + temp. post	5	7.8272	0.3731	0.1042	0.0541	15.9672	0.0000	0.0000	0.0000	MC temp.
Common targets	DET	5	2.0121	0.1005	0.6479	0.4448	4.3460	0.0000	0.0000	0.0000	–
Common targets	DET + temp. pre	5	2.0688	0.1606	0.6479	0.4549	4.3696	0.0000	0.0000	0.0000	temp.
Common targets	EVE raw	5	2.2891	0.1205	0.4938	0.3100	5.0297	0.2676	0.8712	0.0416	raw
Common targets	EVE balanced	5	2.7405	0.0992	0.4840	0.3668	5.2152	0.2646	0.8708	0.0415	conf.-weighted
Common targets	MC Dropout	5	2.0167	0.0993	0.6444	0.4419	4.3618	0.0000	0.0000	0.0000	MC dropout
Common targets	MC dropout + temp. pre	5	2.0726	0.1602	0.6431	0.4511	4.3755	0.0000	0.0000	0.0000	MC temp.
Rare targets	DET	5	10.3754	0.1663	0.0000	0.1226	15.6192	0.0000	0.0000	0.0000	–
Rare targets	DET + temp. pre	5	10.3025	0.1825	0.0000	0.1159	15.4394	0.0000	0.0000	0.0000	temp.
Rare targets	EVE raw	5	10.0247	0.0385	0.0007	0.1188	15.8360	0.3024	0.8696	0.0551	raw
Rare targets	EVE balanced	5	9.5510	0.0415	0.0021	0.0775	14.3337	0.3134	0.8698	0.0553	conf.-weighted
Rare targets	MC Dropout	5	10.3695	0.1671	0.0000	0.1218	15.5882	0.0000	0.0000	0.0000	MC dropout
Rare targets	MC dropout + temp. pre	5	10.3060	0.1781	0.0000	0.1154	15.4679	0.0000	0.0000	0.0000	MC temp.
Temporal bin 1	EVE raw	5	5.3085	0.0613	0.1688	0.0553	12.3061	0.2745	0.8731	0.0512	raw
Temporal bin 2	EVE raw	5	5.4281	0.0448	0.1576	0.0585	13.0687	0.3121	0.8711	0.0515	raw
Temporal bin 3	EVE raw	5	5.7873	0.0704	0.1340	0.0610	12.8727	0.2696	0.8746	0.0464	raw
Temporal bin 4	EVE raw	5	5.9056	0.0267	0.1694	0.0504	12.8982	0.2768	0.8761	0.0494	raw

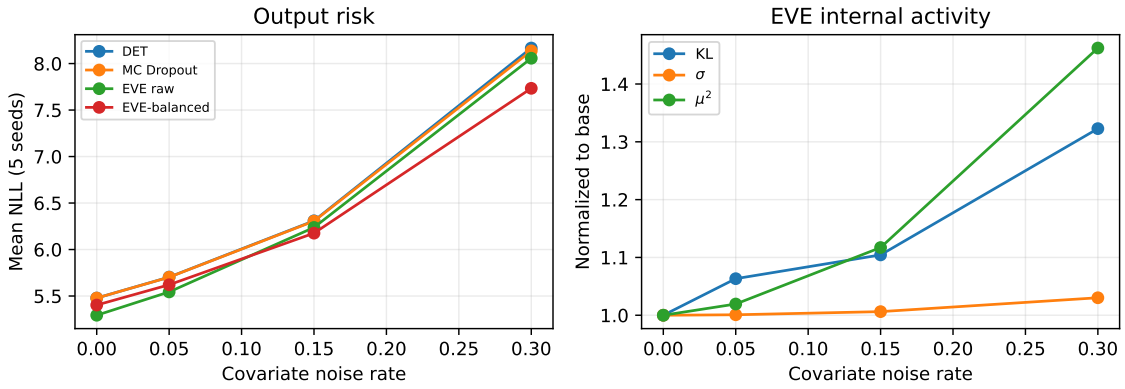

Figure 5: Covariate-corruption dose response for external risk and EVE internal activity.

Table 16: Bootstrap intervals for the raw EVE readout across all reported shift scenarios and key external and internal metrics. Intervals use 200 bootstrap resamples over five seeds.

Scenario	Metric	Mean	95% low	95% high	n
Base test	NLL	5.2932	5.2592	5.3247	5
Base test	Accuracy	0.1660	0.1528	0.1764	5
Base test	ECE	0.0599	0.0550	0.0644	5
Base test	CVaR-NLL	12.3386	12.2221	12.4553	5
Base test	Internal KL	0.2737	0.2647	0.2827	5
Base test	Latent σ	0.8730	0.8678	0.8764	5
Base test	Latent μ^2	0.0512	0.0468	0.0547	5
Long context	NLL	5.1107	5.0729	5.1574	5
Long context	Accuracy	0.1840	0.1715	0.1944	5
Long context	ECE	0.0612	0.0546	0.0699	5
Long context	CVaR-NLL	12.4387	12.2850	12.5361	5
Long context	Internal KL	0.2846	0.2699	0.2998	5
Long context	Latent σ	0.8740	0.8701	0.8780	5
Long context	Latent μ^2	0.0502	0.0444	0.0564	5
Short context	NLL	5.1744	5.1654	5.1825	5
Short context	Accuracy	0.1556	0.1438	0.1674	5
Short context	ECE	0.0552	0.0412	0.0693	5
Short context	CVaR-NLL	12.0562	11.9078	12.2110	5
Short context	Internal KL	0.3373	0.3071	0.3674	5
Short context	Latent σ	0.8905	0.8851	0.8965	5
Short context	Latent μ^2	0.0648	0.0568	0.0719	5
Covariate noise 0.05	NLL	5.5440	5.5071	5.5781	5
Covariate noise 0.05	Accuracy	0.1667	0.1562	0.1778	5
Covariate noise 0.05	ECE	0.0585	0.0445	0.0795	5
Covariate noise 0.05	CVaR-NLL	13.0636	12.5571	13.5808	5
Covariate noise 0.05	Internal KL	0.2910	0.2843	0.2966	5
Covariate noise 0.05	Latent σ	0.8737	0.8692	0.8774	5
Covariate noise 0.05	Latent μ^2	0.0522	0.0480	0.0555	5
Covariate noise 0.15	NLL	6.2373	6.1753	6.2995	5
Covariate noise 0.15	Accuracy	0.1431	0.1382	0.1479	5
Covariate noise 0.15	ECE	0.0602	0.0528	0.0706	5
Covariate noise 0.15	CVaR-NLL	16.1177	15.3664	16.7958	5
Covariate noise 0.15	Internal KL	0.3023	0.2840	0.3140	5
Covariate noise 0.15	Latent σ	0.8784	0.8735	0.8833	5
Covariate noise 0.15	Latent μ^2	0.0572	0.0524	0.0603	5
Covariate noise 0.30	NLL	8.0573	7.8904	8.2063	5
Covariate noise 0.30	Accuracy	0.1028	0.0882	0.1132	5
Covariate noise 0.30	ECE	0.0941	0.0766	0.1131	5
Covariate noise 0.30	CVaR-NLL	18.4830	17.7049	19.4103	5
Covariate noise 0.30	Internal KL	0.3621	0.3292	0.3829	5
Covariate noise 0.30	Latent σ	0.8994	0.8922	0.9067	5
Covariate noise 0.30	Latent μ^2	0.0749	0.0685	0.0795	5

Continued on next page

Table 16 – continued

Scenario	Metric	Mean	95% low	95% high	n
Common targets	NLL	2.2891	2.1737	2.3842	5
Common targets	Accuracy	0.4938	0.4028	0.5736	5
Common targets	ECE	0.3100	0.2171	0.3810	5
Common targets	CVaR-NLL	5.0297	4.7859	5.2648	5
Common targets	Internal KL	0.2676	0.2478	0.2816	5
Common targets	Latent σ	0.8712	0.8695	0.8727	5
Common targets	Latent μ^2	0.0416	0.0370	0.0462	5
Rare targets	NLL	10.0247	9.9927	10.0566	5
Rare targets	Accuracy	0.0007	0.0000	0.0021	5
Rare targets	ECE	0.1188	0.1109	0.1298	5
Rare targets	CVaR-NLL	15.8360	15.7188	15.9679	5
Rare targets	Internal KL	0.3024	0.2926	0.3132	5
Rare targets	Latent σ	0.8696	0.8664	0.8728	5
Rare targets	Latent μ^2	0.0551	0.0508	0.0599	5
Temporal bin 1	NLL	5.3085	5.2625	5.3590	5
Temporal bin 1	Accuracy	0.1688	0.1611	0.1771	5
Temporal bin 1	ECE	0.0553	0.0528	0.0585	5
Temporal bin 1	CVaR-NLL	12.3061	12.1607	12.4744	5
Temporal bin 1	Internal KL	0.2745	0.2596	0.2884	5
Temporal bin 1	Latent σ	0.8731	0.8678	0.8769	5
Temporal bin 1	Latent μ^2	0.0512	0.0467	0.0549	5
Temporal bin 2	NLL	5.4281	5.3883	5.4652	5
Temporal bin 2	Accuracy	0.1576	0.1528	0.1625	5
Temporal bin 2	ECE	0.0585	0.0444	0.0726	5
Temporal bin 2	CVaR-NLL	13.0687	12.8022	13.2966	5
Temporal bin 2	Internal KL	0.3121	0.2880	0.3486	5
Temporal bin 2	Latent σ	0.8711	0.8680	0.8743	5
Temporal bin 2	Latent μ^2	0.0515	0.0472	0.0554	5
Temporal bin 3	NLL	5.7873	5.7404	5.8571	5
Temporal bin 3	Accuracy	0.1340	0.1250	0.1486	5
Temporal bin 3	ECE	0.0610	0.0499	0.0728	5
Temporal bin 3	CVaR-NLL	12.8727	12.6458	13.2120	5
Temporal bin 3	Internal KL	0.2696	0.2469	0.2894	5
Temporal bin 3	Latent σ	0.8746	0.8699	0.8783	5
Temporal bin 3	Latent μ^2	0.0464	0.0421	0.0504	5
Temporal bin 4	NLL	5.9056	5.8796	5.9273	5
Temporal bin 4	Accuracy	0.1694	0.1618	0.1792	5
Temporal bin 4	ECE	0.0504	0.0412	0.0605	5
Temporal bin 4	CVaR-NLL	12.8982	12.7149	13.0888	5
Temporal bin 4	Internal KL	0.2768	0.2589	0.2946	5
Temporal bin 4	Latent σ	0.8761	0.8720	0.8806	5
Temporal bin 4	Latent μ^2	0.0494	0.0445	0.0553	5

Table 17: All reported correlations between changes in EVE internal activity and changes in external metrics. These descriptive correlations pool the available seed-by-scenario deltas within each shift family.

Shift family	Internal change	External change	Pearson r	n
all	Δ internal KL	Δ CVaR-NLL	0.4507	90
all	Δ internal KL	Δ NLL	0.3431	90
all	Δ latent μ^2	Δ CVaR-NLL	0.6978	90
all	Δ latent μ^2	Δ NLL	0.5175	90
all	Δ latent σ	Δ ECE	-0.2323	90
context length	Δ internal KL	Δ CVaR-NLL	-0.4236	20
context length	Δ internal KL	Δ NLL	0.5931	20
context length	Δ latent μ^2	Δ CVaR-NLL	-0.3775	20
context length	Δ latent μ^2	Δ NLL	0.6955	20
context length	Δ latent σ	Δ ECE	-0.4034	20
covariate	Δ internal KL	Δ CVaR-NLL	0.5791	30
covariate	Δ internal KL	Δ NLL	0.6696	30
covariate	Δ latent μ^2	Δ CVaR-NLL	0.8715	30
covariate	Δ latent μ^2	Δ NLL	0.9685	30

Continued on next page

Table 17 – continued

Shift family	Internal change	External change	Pearson r	n
covariate	Δ latent σ	Δ ECE	0.3578	30
target frequency	Δ internal KL	Δ CVaR-NLL	0.4537	20
target frequency	Δ internal KL	Δ NLL	0.4519	20
target frequency	Δ latent μ^2	Δ CVaR-NLL	0.9134	20
target frequency	Δ latent μ^2	Δ NLL	0.9117	20
target frequency	Δ latent σ	Δ ECE	0.1968	20
temporal order	Δ internal KL	Δ CVaR-NLL	0.0143	20
temporal order	Δ internal KL	Δ NLL	-0.3052	20
temporal order	Δ latent μ^2	Δ CVaR-NLL	-0.0678	20
temporal order	Δ latent μ^2	Δ NLL	-0.4744	20
temporal order	Δ latent σ	Δ ECE	-0.5442	20

A.7 Metric-specific primary figures

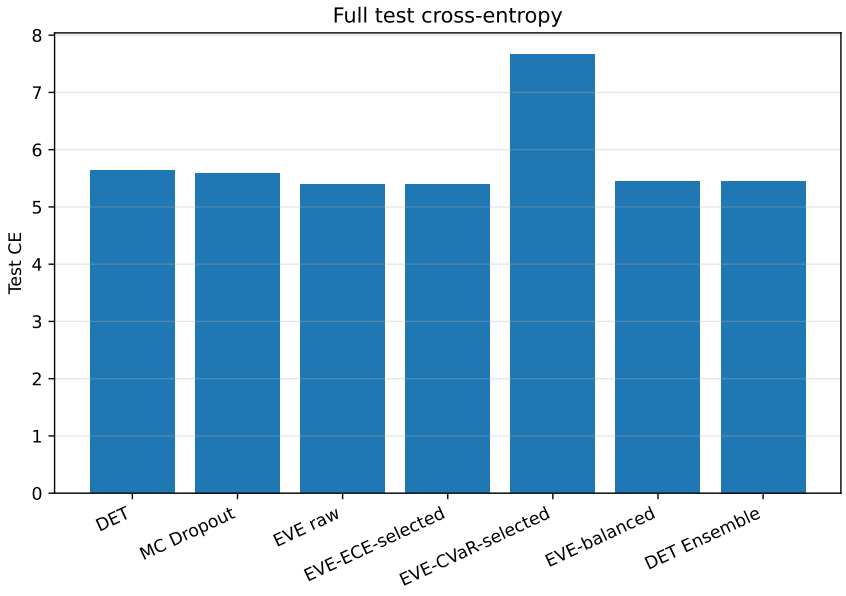

Figure 6: Full-test cross-entropy by evaluated model and readout.

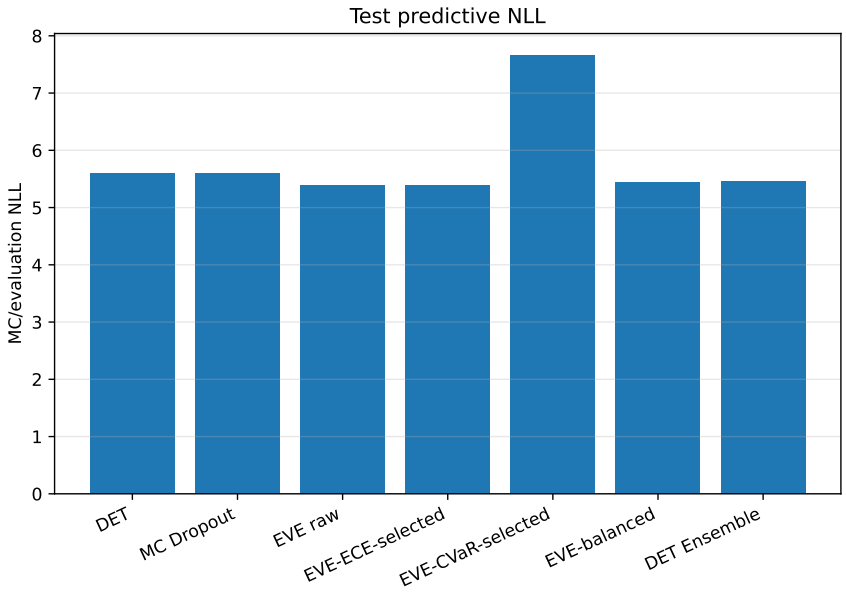

Figure 7: Monte Carlo test NLL by evaluated model and readout.

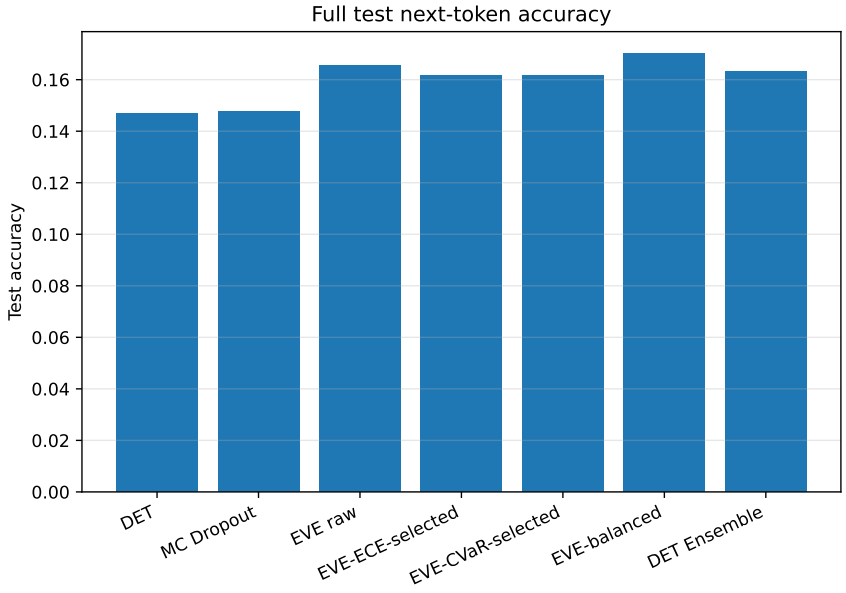

Figure 8: Next-token test accuracy by evaluated model and readout.

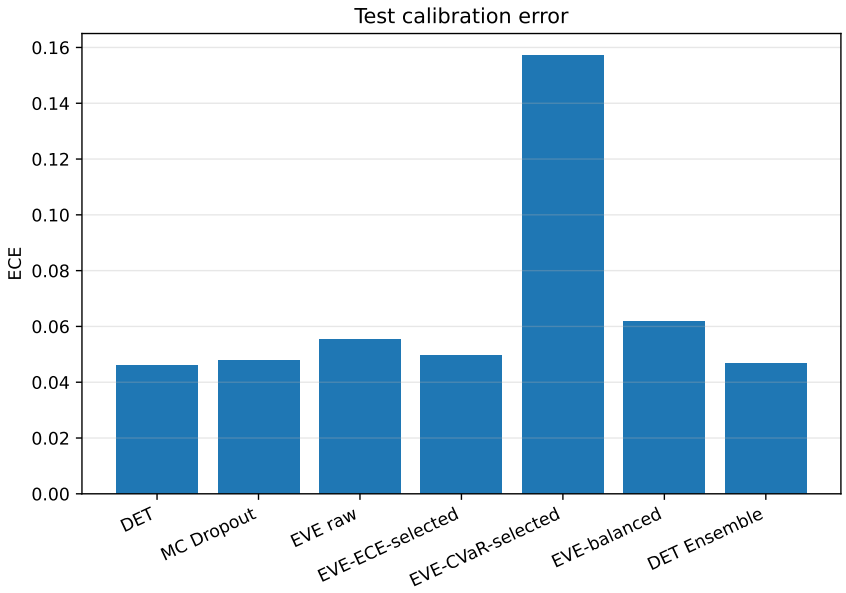

Figure 9: Expected calibration error by evaluated model and readout.

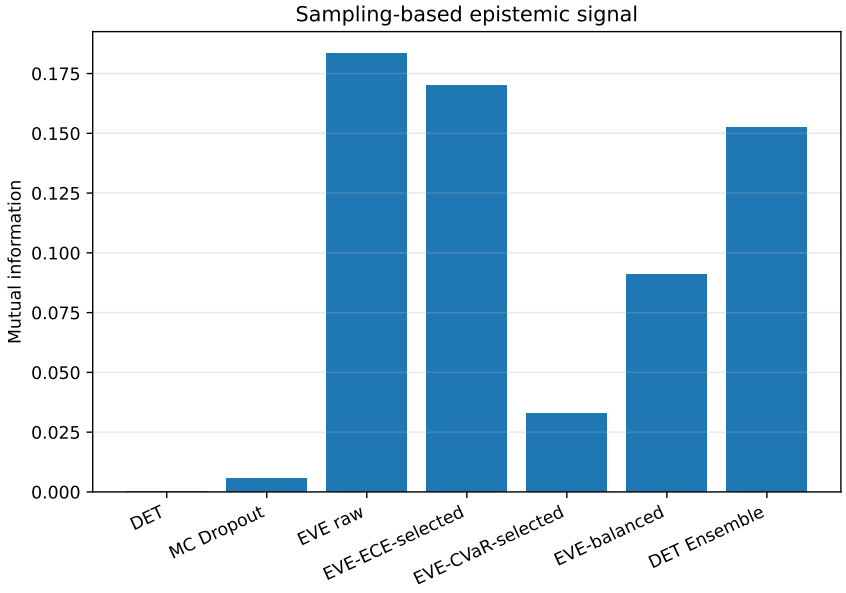

Figure 10: Sampling-based mutual-information proxy by evaluated model and readout.

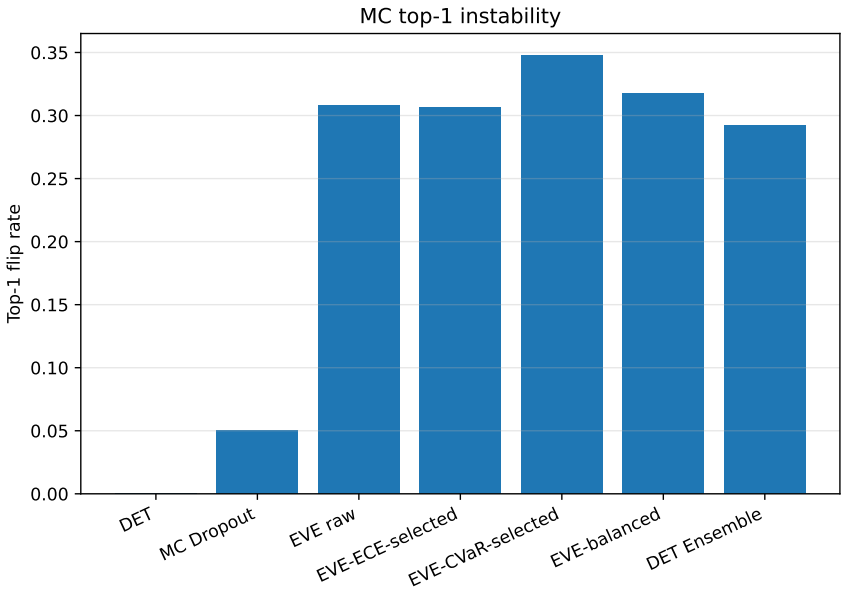

Figure 11: Top-1 sample flip rate by evaluated model and readout.

A.8 Scope of the reported numerical record

The tables above contain the aggregate and seedwise numerical values used for every empirical statement and plotted result in the article.