An Isotropic Approach to Efficient Uncertainty Quantification with Gradient Norms

Anonymous Authors

Abstract

Existing methods for quantifying predictive uncertainty in neural networks are either computationally intractable for large language models or require access to training data that is typically unavailable. We derive a lightweight alternative through two approximations: a first-order Taylor expansion that expresses uncertainty in terms of the gradient of the prediction and the parameter covariance, and an isotropy assumption on the parameter covariance. Together, these yield epistemic uncertainty as the squared gradient norm and aleatoric uncertainty as the Bernoulli variance of the point prediction, from a single forward-backward pass through an unmodified pretrained model. We justify the isotropy assumption by showing that covariance estimates built from non-training data introduce structured distortions that isotropic covariance avoids, and that theoretical results on the spectral properties of large networks support the approximation at scale. Validation against reference Markov Chain Monte Carlo estimates on synthetic problems shows strong correspondence that improves with model size. We then use the estimates to investigate when each uncertainty type carries useful signal for predicting answer correctness in question answering with large language models, revealing a benchmark-dependent divergence: the combined estimate achieves the highest mean AUROC on TruthfulQA, where questions involve genuine conflict between plausible answers, but falls to near chance on TriviaQA's factual recall, suggesting that parameter-level uncertainty captures a fundamentally different signal than self-assessment methods.

1. Introduction

As large language models (LLMs) are increasingly deployed in consequential applications—from medical diagnosis assistance to legal document analysis and financial advisory services (Chen et al., 2024)—ensuring their trustworthiness becomes paramount. A fundamental requirement is the ability to assess when a model's predictions may be unreliable, yet contemporary LLMs provide no built-in mechanism for distinguishing what they know from what they do not, invariably delivering outputs with the same authoritative tone regardless of whether their training provides adequate foundation for the claims they make (Ji et al., 2023; Zhou et al., 2024).

Unreliable predictions can arise for two fundamentally different reasons. Some predictions are uncertain because the question itself admits multiple valid answers, an inherent ambiguity that no amount of additional training data would resolve. Others are uncertain because the model has seen too little relevant training data to have settled on a reliable answer, a gap in knowledge that more data could, in principle, fill. These two sources are known as *aleatoric* and *epistemic* uncertainty, respectively, and distinguishing between them is essential: the former signals an irreducibly hard problem, while the latter flags a prediction that should not be trusted.

The Bayesian framework provides a natural language for quantifying both sources of uncertainty (Neal, 2012; Kendall and Gal, 2017), but applying it is intractable for modern neural networks

(Blundell et al., 2015), and existing approximations—deep ensembles (Lakshminarayanan et al., 2017), Monte Carlo dropout (Gal and Ghahramani, 2016), Laplace approximations (MacKay, 1992; Daxberger et al., 2021a)—each impose severe practical limitations: training multiple models, requiring architectural support, or computing Hessian matrices from training data that is typically unavailable for contemporary LLMs.

We address this gap through two approximations. First, a first-order Taylor expansion, also known as the delta method (Doob, 1935), expresses uncertainty as a product of two factors: the gradient of the prediction with respect to the parameters, and the covariance of those parameters. Second, we assume isotropic parameter covariance, leaving only the gradient factor as our epistemic estimate. The same expansion yields the aleatoric estimate directly from the model's output, giving a complete uncertainty decomposition from a single forward-backward pass through an unmodified pretrained model, with no ensembles, no sampling, and no Hessian estimation.

The isotropy assumption is the key simplification: we argue that proxy covariance estimates built from non-training data can introduce structured distortions worse than isotropy's uniform error, supported by theoretical results on Hessian spectra at scale and empirical precedent across adjacent fields. We validate against reference MCMC estimates on synthetic problems (Spearman ρ of 0.44–0.99, improving with model scale as the theory predicts) and then investigate when aleatoric and epistemic uncertainty carry useful signal for predicting answer correctness in LLM question answering, revealing a benchmark-dependent divergence: the combined estimate achieves the highest area under the receiver operating characteristic curve (AUROC) (0.63) on TruthfulQA, while epistemic uncertainty falls to near chance on TriviaQA, suggesting that parameter-level uncertainty is most informative when the task involves genuine conflict between plausible answers rather than factual memorization.

Our contributions are as follows:

1. We derive epistemic and aleatoric uncertainty estimators from a first-order Taylor expansion under an isotropy assumption, providing the first systematic justification for the isotropy assumption through proxy bias analysis, spectral theory, and empirical precedent.

2. We validate both estimators against reference Markov Chain Monte Carlo (MCMC) estimates on synthetic problems, demonstrating strong correspondence in classification and an improving trend with model scale.

3. We investigate the utility of aleatoric and epistemic uncertainty for predicting answer correctness in LLM question answering, revealing a benchmark-dependent divergence that illuminates when parameter-level uncertainty provides useful signal.

2. Background

2.1. Uncertainty in Machine Learning

The Bayesian framework reasons about uncertainty by maintaining a posterior distribution $p(\theta \mid \mathcal{D})$ over model parameters θ given data $\mathcal{D}$ rather than a point estimate; predictions for a new input x are made by marginalizing over this posterior, $p(y \mid x) = \mathbb{E}_\theta[p(y \mid x, \theta)]$, where y denotes the output. The uncertainty in this predictive distribution decomposes into two components (Hüllermeier and Waegeman, 2021): *aleatoric uncertainty*, capturing irreducible randomness in the data-generating process, and *epistemic uncertainty*, reflecting incomplete knowledge about the parameters due to finite training data. Regularized loss minimization is mathematically equivalent to maximum a posteriori (MAP) estimation (Bishop and Nasrabadi, 2006; Goodfellow et al., 2016), so the loss surface of any regularized neural network is a posterior parameter distribution; standard training collapses this distribution to its mode, discarding the information necessary for uncertainty quantification.

2.2. Uncertainty Decomposition

The most widely used decomposition operates on the entropy of the predictive distribution (Smith and Gal, 2018):

$$\underbrace{\mathbb{H}[y \mid x]}_{\text{total}} = \underbrace{\mathbb{E}_\theta[\mathbb{H}[y \mid x, \theta]]}_{\text{aleatoric}} + \underbrace{\mathbb{I}[y; \theta \mid x]}_{\text{epistemic}}, \tag{1}$$

where the aleatoric component is the expected conditional entropy and the epistemic component is the mutual information between the prediction and the parameters (Gal and Ghahramani, 2016; Kuhn et al., 2022). However, Wimmer et al. (2023) show that mutual information violates several desirable axiomatic properties—it is not maximal under complete ignorance, not monotone under mean-preserving spreads, and not invariant under location shifts—and existing estimators degrade to near-random performance under non-trivial aleatoric uncertainty (Tomov et al., 2025).

An alternative decomposition considers variance rather than entropy and operates label-wise (Sale et al., 2023, 2024). For a given class c, the model's prediction $p(y_c \mid x, \theta)$ can be viewed as the success probability of a Bernoulli variable. Applying the law of total variance to the posterior gives:

$$\underbrace{\mathrm{Var}[y_c \mid x]}_{\text{total}} = \underbrace{\mathbb{E}_\theta[p(y_c \mid x, \theta)(1 - p(y_c \mid x, \theta))]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_\theta[p(y_c \mid x, \theta)]}_{\text{epistemic}}. \tag{2}$$

This label-wise decomposition provides a more fine-grained view than the entropy-based one and does not suffer from the axiomatic violations above (Sale et al., 2024); we adopt it throughout this work.

Both decompositions assume that the prediction at the MAP estimate equals the posterior predictive expectation, $p(y \mid x, \theta^*) = \mathbb{E}_\theta[p(y \mid x, \theta)]$, a necessary condition for both the law of total variance and the entropy identity. This assumption is incorrect in practice, but these decompositions remain a standard analytical tool (Smith and Gal, 2018; Depeweg et al., 2018; Jayasekera et al., 2025).

3. Gradient-Based Epistemic Uncertainty Quantification

We approximate the predictive distribution via a first-order Taylor expansion around the parameter point estimate θ^*: $p(y_c \mid x, \theta) \approx p(y_c \mid x, \theta^*) + g^\top(\theta - \theta^*)$, where $g = \nabla_\theta p(y_c \mid x, \theta)\big|_{\theta^*}$. Substituting into the variance-based definition of epistemic uncertainty, $\mathrm{Var}_\theta[p(y_c \mid x, \theta)]$:

$$\begin{aligned} \mathrm{Var}_\theta[p(y_c \mid x, \theta)] &\approx \mathrm{Var}_\theta\left[p(y_c \mid x, \theta^*) + g^\top(\theta - \theta^*)\right] \\ &= \mathrm{Var}_\theta\left[g^\top(\theta - \theta^*)\right] = \mathrm{Var}_\theta\left[g^\top \theta\right] \\ &= g^\top \mathrm{Cov}[\theta]\, g \end{aligned} \tag{3}$$

This is known as the delta method. However, at this point we are still left with a notoriously difficult object: the parameter covariance matrix. Most work on the delta method accordingly focuses on different estimations of this matrix (Nilsen et al., 2022; Schmitt et al., 2025). In contrast, we take our approximation a step further by assuming isotropic parameter covariance. Since any isotropic covariance $\sigma^2 I$ differs from the identity only by a constant factor that scales all estimates equally, we can set $\mathrm{Cov}[\theta] = I$ without loss of generality:

$$\mathrm{Var}_\theta[p(y_c \mid x, \theta)] \approx g^\top \mathrm{Cov}[\theta]\, g \approx g^\top g \tag{4}$$

We are left with $\|g\|^2$, the squared gradient norm, as our estimate of epistemic uncertainty. This is a strong assumption that requires justification.

We motivate the isotropy assumption from three complementary angles: (i) the available alternatives to estimate the true covariance may introduce structured distortions worse than assuming isotropy (Section 3.1), (ii) theoretical results on the Hessian spectrum suggest that the identity is a reasonable proxy for the true covariance as model size grows, and (iii) empirical precedent confirms that it performs well across tasks adjacent to uncertainty quantification (Section 3.2).

3.1. Proxy Covariance Estimates Introduce Structured Bias

For LLMs, the true training data is typically unavailable even for ostensibly open models (Touvron et al., 2023; Grattafiori et al., 2024; Abdin et al., 2024; Gemma Team et al., 2025), so any covariance estimate must be built from proxy data: corpora that plausibly overlap with the true training distribution but inevitably do not match it exactly.

Consider a diagonal Laplace approximation (Daxberger et al., 2021a), the closest widely used alternative to isotropic covariance (Denker and LeCun, 1990; Gui et al., 2021; Ortega et al., 2024). Let Σ^*_{diag} denote the true diagonal of Σ and $\hat{\Sigma}_{\text{diag}}$ the diagonal estimated from proxy data. The total error decomposes as

$$g^\top \Sigma\, g - g^\top \hat{\Sigma}_{\text{diag}}\, g = \underbrace{g^\top (\Sigma - \Sigma^*_{\text{diag}})\, g}_{\text{(i) structural error}} + \underbrace{g^\top (\Sigma^*_{\text{diag}} - \hat{\Sigma}_{\text{diag}})\, g}_{\text{(ii) estimation error}}, \tag{5}$$

and identically for our isotropic approximation:

$$g^\top \Sigma\, g - \|g\|^2 = \underbrace{g^\top (\Sigma - \Sigma^*_{\text{diag}})\, g}_{\text{(i) structural error}} + \underbrace{g^\top (\Sigma^*_{\text{diag}} - I)\, g}_{\text{(iii) anisotropy error}}. \tag{6}$$

Since the structural error is identical, the difference reduces to comparing terms (ii) and (iii):

$$\text{(ii)} = \sum_{i=1}^{P} g_i^2 (\Sigma_{ii} - \hat{\Sigma}_{ii}), \qquad \text{(iii)} = \sum_{i=1}^{P} g_i^2 (\Sigma_{ii} - 1). \tag{7}$$

Both errors are weighted by g_i^2, but the proxy biases $\Sigma_{ii} - \hat{\Sigma}_{ii}$ reflect the coverage of the proxy corpus: parameters that the proxy data activates receive large curvature and small variance, while parameters it does not activate retain poorly constrained variance. Since g_i^2 upweights the parameters the model relies on for a given prediction, the proxy's errors concentrate precisely where accuracy matters most. The identity's biases $\Sigma_{ii} - 1$ encode no data-dependent structure and therefore cannot introduce spatially structured distortions of this kind.

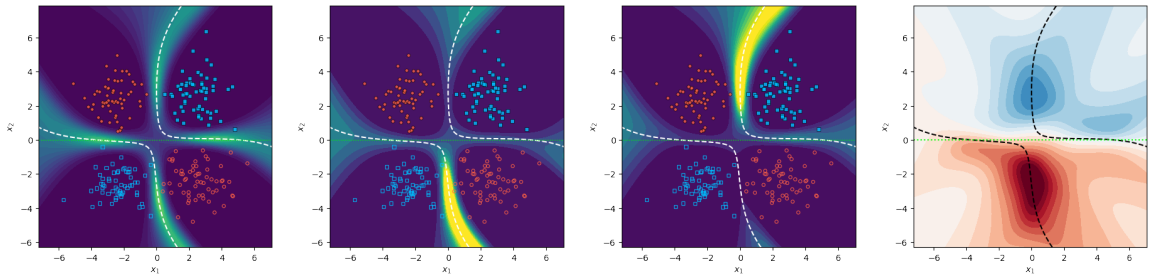
We demonstrate this empirically on three 2D classification problems with spatially symmetric decision boundaries (details in Appendix A). Splitting the training data by location and computing empirical proxies of the Fisher information matrix (FIM) from each half yields two Hessian estimates H_A and H_B, with corresponding uncertainty estimates $U_A(x) = g^\top H_A^{-1} g$ and $U_B(x) = g^\top H_B^{-1} g$. As Fig. 1 shows for the XOR problem, each proxy inflates uncertainty in the half of input space absent from its data and suppresses it where its data is concentrated (Cohen's $d = -2.5$), while the identity produces spatially symmetric estimates that peak at the decision boundary, as the problem's symmetry demands.

To confirm that this extends beyond synthetic settings to a practical natural language processing task, we repeat the experiment on a DistilBERT (Sanh et al., 2020) domain classification task (science vs. sports text from 20 Newsgroups; details in Appendix A.1). The results mirror the synthetic case: the log-ratio $\log(U_A/U_B)$ flips sign between domains (median -0.61 on science, $+1.25$ on sports), with a Cohen's d of -5.3, confirming that each proxy inflates uncertainty on the domain absent from its data. The identity produces no such domain-dependent distortion (Fig. 4).

In short, any covariance estimate for an LLM must be built from proxy data, and any such estimate imposes structured, data-dependent distortions. The identity avoids this: ignorance may be preferable to bias.

3.2. The Covariance of Large Models Approaches the Identity

Further, the isotropy assumption is a theoretically well-grounded approximation of the true covariance for large models. The Hessian of deep networks exhibits a characteristic spectral pattern—a small



(a) Identity ($C = I$) (b) Proxy H_A^{-1} (top half) (c) Proxy H_B^{-1} (bottom half) (d) $\log(U_A/U_B)$

Figure 1: Normalized epistemic uncertainty on the XOR problem under three covariance assumptions, and the log ratio $\log(U_A/U_B)$. The identity produces spatially symmetric estimates that peak along the two decision boundaries. Each proxy inflates uncertainty in the half of input space *absent* from its data and suppresses it where its data is concentrated, despite the problem's underlying symmetry; the log ratio makes this asymmetry explicit. See Appendix A for experimental details and additional problems.

number of large eigenvalues with a vast bulk near zero (Sagun et al., 2017; Pennington and Worah, 2018; Karakida et al., 2019)—and inverting it amplifies the near-zero eigenvalues, so the damping term λI universally added for stabilization dominates in most parameter directions, causing the damped inverse $(F + \lambda I)^{-1}$ to converge to approximately $(1/\lambda)I$. Li et al. (2025) verify empirically that for LLMs, the damping overwhelms the Hessian so the damped inverse is effectively proportional to the identity, and Kwon et al. (2023) independently observe that iterative Hessian inversion collapses to the identity baseline on a 13B-parameter model. Weight decay imposes an isotropic Gaussian prior (Bishop and Nasrabadi, 2006), and pretrained language models have very low intrinsic dimensionality relative to their parameter count (Aghajanyan et al., 2021; Hu et al., 2021), so the posterior is driven by this isotropic prior in all but a small subspace. The isotropy assumption has also been employed, often implicitly, across data attribution (Pruthi et al., 2020; Charpiat et al., 2019; Yang et al., 2024; Jaburi et al., 2025; Kowal et al., 2026), out-of-distribution detection (Bergamin et al., 2022; Zhdanov et al., 2025), and dataset pruning (Paul et al., 2021), consistently matching or outperforming more elaborate curvature corrections. A detailed review of these theoretical results and empirical precedents is provided in Appendix B.

3.3. Estimating Aleatoric Uncertainty

The same Taylor expansion as in Eq. (3) can be used to derive an estimate of aleatoric uncertainty. Applying it to the Bernoulli variance $h(\theta) = p(y_c \mid x, \theta)(1 - p(y_c \mid x, \theta))$ and taking expectations gives:

$$\mathbb{E}_\theta[h(\theta)] \approx h(\theta^*) + \nabla_\theta h(\theta)\big|_{\theta^*}^\top (\mathbb{E}_\theta[\theta] - \theta^*) \tag{8}$$

The remaining term vanishes: combining the Taylor approximation with the assumption $p(y_c \mid x, \theta^*) = \mathbb{E}_\theta[p(y_c \mid x, \theta)]$ required for the variance decomposition (Section 2) yields $\mathbb{E}_\theta[\theta] = \theta^*$ (proof in Appendix C). Thus, the estimate of aleatoric uncertainty reduces to $h(\theta^*) = p(y_c \mid x, \theta^*)(1 - p(y_c \mid x, \theta^*))$. Beyond the Taylor expansion, the only additional requirement is $p(y_c \mid x, \theta^*) = \mathbb{E}_\theta[p(y_c \mid x, \theta)]$, which is itself necessary for the uncertainty decomposition.

3.4. Extension to Sequences

The derivations above treat y_c as a single discrete output. For generative language models, the object of interest is a sequence $y = (y_{c_1}, \ldots, y_{c_T})$. A direct extension would consider $\mathrm{Var}_\theta[p(y \mid x, \theta)]$, but

the joint probability scales exponentially with sequence length, making it unsuitable for comparing sequences of different lengths. Instead, we apply the Taylor expansion to the mean predicted probability,

$$\bar{p}(y \mid x, \theta) = \frac{1}{T} \sum_{t=1}^{T} p(y_{c_t} \mid y_{<t}, x, \theta), \tag{9}$$

preserving consistency with the single-token derivation in Section 3. This yields

$$\mathrm{Var}_\theta[\bar{p}(y \mid x, \theta)] \approx \bar{g}^\top \mathrm{Cov}[\theta]\, \bar{g}, \tag{10}$$

where $\bar{g} = \frac{1}{T} \sum_{t=1}^{T} \nabla_\theta p(y_{c_t} \mid y_{<t}, x, \theta^*)$. Under isotropic covariance, epistemic uncertainty reduces to $\|\bar{g}\|^2$, computable from a single backward pass.

Expanding $\|\bar{g}\|^2 = \left\| \frac{1}{T} \sum_t g_t \right\|^2$ yields $\frac{1}{T^2} \left(\sum_t \|g_t\|^2 + \sum_{t \neq s} g_t^\top g_s \right)$. The cross-terms $g_t^\top g_s$ capture correlations in parameter space between token predictions, present in the sequence-level formulation but absent from an average of per-token norms. These cross-terms allow the sequence-level estimate to reflect uncertainty about the sequence as a whole, rather than treating each token independently. If these correlations are negligible, the two approaches coincide; if significant, the sequence-level estimate captures them at no additional cost, while per-token backward passes scale linearly with T.

Aleatoric uncertainty extends symmetrically as the mean per-token Bernoulli variance, $\frac{1}{T} \sum_{t=1}^{T} p(y_{c_t} \mid y_{<t}, x, \theta^*)(1 - p(y_{c_t} \mid y_{<t}, x, \theta^*))$, requiring only the forward pass.

4. Experiments

We first validate $\|g\|^2$ against MCMC estimates on synthetic problems (Section 4.1), then investigate the utility of aleatoric and epistemic uncertainty for predicting answer correctness in LLM question answering (Section 4.2).

4.1. Validation

	Linear	XOR	Rings
Epistemic (GN)			
r	0.95	0.65	0.86
ρ	0.99	0.68	0.44
Epistemic (LA)			
r	0.95	0.68	0.86
ρ	0.99	0.70	0.46
Aleatoric			
r	0.99	0.76	0.95
ρ	1.00	0.74	0.58

(a) Binary classification

	Clusters	Spirals	Rings
Epistemic (GN)			
r	0.86	0.76	0.88
ρ	0.97	0.91	0.97
Aleatoric			
r	0.95	0.96	0.96
ρ	0.99	0.97	0.98

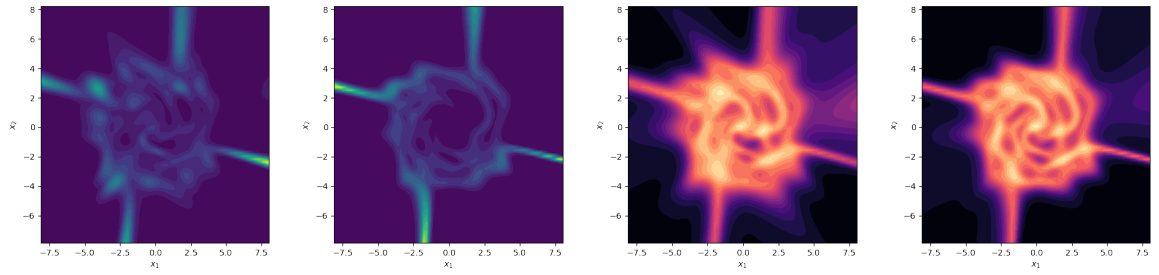
(b) Multiclass classification

	Linear	Nonlin.
Epistemic (GN)		
r	0.98	0.73
ρ	0.99	0.81
Epistemic (LA)		
r	1.00	0.93
ρ	1.00	0.97

(c) Regression

Table 1: Pearson (r) and Spearman (ρ) correlations between our estimates and MCMC estimates. GN: gradient norm $\|g\|^2$; LA: Laplace $g^\top H^{-1} g$. Aleatoric: $p(y_c \mid x, \theta^*)(1 - p(y_c \mid x, \theta^*))$.

We compare our estimates directly against the quantity they are designed to approximate, rather than using out-of-distribution (OOD) detection as a proxy. OOD detection assumes that inputs far from the training data produce high epistemic uncertainty, but Bayesian epistemic uncertainty depends on the space of plausible parameterizations, not on distance from training data alone—a linear classifier, for instance, cannot exhibit high epistemic uncertainty far from its boundary regardless of how distant the input is from any training point. This disconnect has been observed in practice (Ulmer et al., 2020), so failures on OOD benchmarks may reflect a mismatch between the validation assumption



(a) Epistemic, MCMC (b) Epistemic, GN (c) Aleatoric, MCMC (d) Aleatoric, point est.

Figure 2: Multiclass spirals uncertainty maps. Left two panels: epistemic uncertainty (MCMC vs. gradient norm $\|g\|^2$). Right two panels: aleatoric uncertainty (MCMC vs. point estimate). All maps are individually normalized to $[0, 1]$. Additional problems in Appendix D.

and the quantity being measured rather than a deficiency of the method. On synthetic problems where the parameter count permits full posterior inference, we use Hamiltonian Monte Carlo (HMC) (Betancourt, 2018) with dual-averaging step-size adaptation (Nesterov, 2009; Hoffman and Gelman, 2014) to compute $\mathrm{Var}_\theta[p(y_c \mid x, \theta)]$ and measure how well $\|g\|^2$ tracks it, using Pearson and Spearman correlation. We also evaluate the Laplace approximation $g^\top H^{-1} g$ as a point of comparison.

We conduct experiments across classification (logistic/softmax regression and multilayer perceptrons (MLPs) on 2D problems of varying complexity), regression (single-hidden-layer MLP on 1D problems), and a scaling study (12 to $\sim 10^6$ parameters). Full setup details are given in Appendix F.

4.1.1. Classification and Regression

Across six classification problems, $\|g\|^2$ consistently tracks the MCMC estimates (Table 1). Correlations are highest when the model is correctly specified and lower for MLPs with anisotropic posteriors, though the correct spatial pattern is always recovered (Fig. 2). The Laplace approximation provides almost no improvement over $\|g\|^2$ in classification, indicating that posterior curvature is negligible in this setting. Aleatoric point estimates are uniformly strong across all multiclass problems.

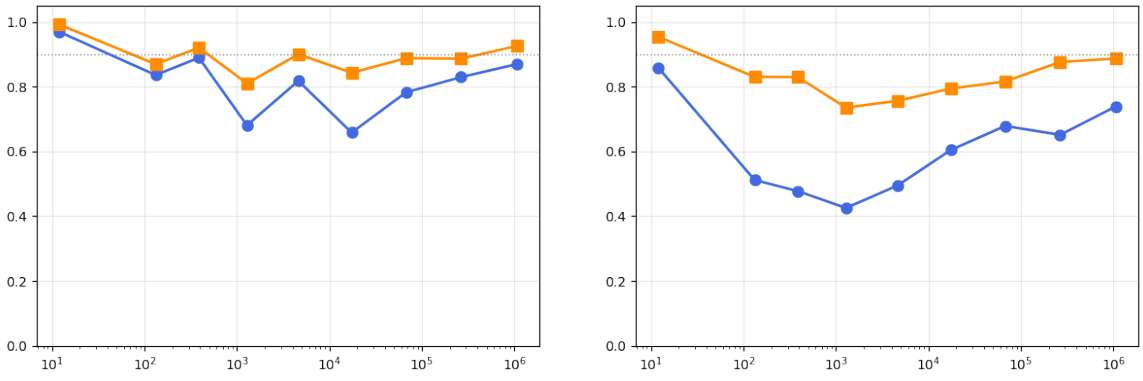
In regression, the isotropy assumption breaks down: on a nonlinear problem the Laplace approximation substantially outperforms $\|g\|^2$ (Table 1 and Appendix E), confirming that the isotropy assumption rather than the shared Taylor expansion is the primary source of error.

4.1.2. Scaling with Model Size

The theoretical arguments in Section 3.2 predict that the identity becomes a better proxy as network scale increases. We test this by training models from 12 to approximately 10^6 parameters on a concentric rings problem. As Fig. 3 shows, the epistemic correlation follows a U-shaped trajectory: high for the smallest models (where the posterior is trivially isotropic), dipping at intermediate scales (where the posterior is anisotropic but spectral concentration has not yet taken effect), and recovering to $\rho = 0.87$ at approximately 10^6 parameters. The recovery begins roughly when the number of parameters exceeds the training samples, placing the model in the overparameterized regime where the FIM becomes increasingly low-rank, suggesting that the approximation is weakest in precisely the regime where these synthetic experiments operate and should improve further at the scales of LLMs.

4.2. Utility of Uncertainty for Question Answering

We now investigate when aleatoric and epistemic uncertainty carry useful signal for predicting answer correctness in LLM question answering. We evaluate four LLMs on TriviaQA (Joshi et al., 2017) and



(a) Spearman correlation

(b) Pearson correlation

Figure 3: Correlation between our estimates and MCMC estimates as a function of model size (number of parameters) on a concentric rings problem. Both epistemic (blue) and aleatoric (orange) correlations follow a U-shaped trajectory, dipping at intermediate scales and recovering at larger model sizes. Full per-model results in Table 10.

TruthfulQA (Lin et al., 2022), measuring each uncertainty type's ability to predict correctness via AUROC. Following Farquhar et al. (2024), we use their semantic equivalence criterion and compare against naïve and semantic entropy (Kuhn et al., 2022) as well as P(True), which prompts the model to assess its own answer correctness (Kadavath et al., 2022); we exclude deep ensembles, which require training multiple models, and the Laplace approximation, which faces the proxy-data concerns of Section 3.1. Results are averaged over four models and 300 bootstrap runs; full details in Appendix G.

4.2.1. Results

Table 2 reports the AUROC scores averaged across models. On TriviaQA, P(True) (0.69) dominates. On TruthfulQA, the pattern reverses: the combined estimate achieves 0.63, the highest score on this benchmark, significantly outperforming P(True) (0.55) and the entropy baselines ($p < 0.01$ after Benjamini–Hochberg (BH) correction; Appendix G.7).

Method	TriviaQA	TruthfulQA
P(True)	**0.69** $\pm$ 0.06	0.55 $\pm$ 0.06
Sem. Entropy	0.55 $\pm$ 0.03	0.54 $\pm$ 0.08
Naïve Entropy	0.52 $\pm$ 0.04	0.51 $\pm$ 0.08
Aleatoric	0.60 $\pm$ 0.04	0.60 $\pm$ 0.09
Epistemic	0.52 $\pm$ 0.07	0.55 $\pm$ 0.07
Epi. & Alea.	0.61 $\pm$ 0.05	**0.63** $\pm$ 0.08

Table 2: AUROC (mean $\pm$ std over 300 bootstrap runs, 4 LLMs) for predicting answer correctness. Higher is better; 0.50 is chance. Best per column in bold.

This divergence between the two benchmarks is the most instructive finding. TriviaQA tests factual recall, where a model may be equally confident in correct and incorrect answers, so uncertainty and correctness are largely independent. TruthfulQA targets common misconceptions with genuinely ambiguous answer spaces, creating both inherent output ambiguity and epistemic conflict between popular and truthful answers. In this setting, the aleatoric estimate (0.60) reflects output-level hedging, the epistemic estimate (0.55) captures parameter-level sensitivity, and their combination (0.63) outperforms all baselines. P(True) loses its advantage because the model's self-assessment is precisely what TruthfulQA is designed to defeat, while aleatoric and epistemic uncertainty carry genuinely useful signal, suggesting that these uncertainty types are most informative when the task involves conflict between plausible parameterizations rather than factual memorization. The epistemic estimate and P(True) are only

weakly correlated (Spearman $\rho \approx -0.2$ on both benchmarks; Appendix G.8), confirming they capture largely distinct signal.

The per-model breakdown (Table 11 in Appendix G.5) reveals substantial model-level variation, but several trends are consistent. The benchmark divergence is universal: on TruthfulQA, the combined estimate is at least on par with the best baseline for every model, while on TriviaQA the reverse holds for three of four models. The relative utility of aleatoric and epistemic uncertainty is model-dependent: both Llama models favor aleatoric uncertainty on TruthfulQA, while OLMo and Phi-4 favor epistemic. Models from the same family behaving alike suggests training data as a driver—models that have seen more relevant data would have less epistemic uncertainty to exploit, making output ambiguity the dominant signal, while models with genuine knowledge gaps benefit more from the epistemic estimate. Notably, the aleatoric estimate alone (0.60) is competitive; the epistemic component provides a significant lift when combined ($0.60 \rightarrow 0.63$, paired bootstrap $p = 0.018$ on TruthfulQA, not significant on TriviaQA; Appendix G.7) and on individual models carries substantial independent signal (e.g., Phi-4 epistemic alone: 0.63).

Beyond accuracy, the gradient-based estimates offer a substantial computational advantage, as our method requires only a single backward pass after generation and is 46–107× faster per sample than the baselines, which require multiple forward passes for sampling or self-evaluation (Appendix G.6).

5. Conclusion

By approximating uncertainty via a first-order Taylor expansion under isotropic covariance, we reduce epistemic uncertainty to the squared norm of the prediction gradient and aleatoric uncertainty to the Bernoulli variance of the point prediction, giving a complete uncertainty decomposition from a single forward-backward pass through an unmodified pretrained model.

Validation against reference MCMC estimates on synthetic problems shows strong correspondence in classification (Spearman ρ of 0.44–0.99 across settings), with an improving trend at larger model sizes that supports the isotropy assumption. The downstream question answering experiments reveal that uncertainty estimates are most informative when the model faces genuine conflict between plausible parameterizations (as on TruthfulQA, where at least one uncertainty estimate exceeds all baselines for every model), rather than when correctness depends on factual memorization, though the relative utility of aleatoric and epistemic uncertainty varies substantially between models. More broadly, the near-chance epistemic AUROC on TriviaQA suggests that epistemic uncertainty may not be as useful for hallucination detection as previously assumed (Xiao and Wang, 2021; Han et al., 2025; Park et al., 2026; Liu et al., 2026), since factual errors need not coincide with parameter-level disagreement; gradient-based uncertainty captures a complementary signal to self-assessment methods like P(True), with the two excelling on fundamentally different question types. More generally, even when the Bayesian calibration of the squared gradient norm is approximate, it retains a meaningful ranking of inputs as a measure of local sensitivity to parameter perturbations.

Limitations. The estimates are on the scale of squared gradient norms, which lack intuitive interpretation and do not generalize across model architectures: training an answer correctness classifier on the uncertainty estimates from three models and evaluating on the fourth yields chance-or-below performance, with the relationship between gradient norm and correctness occasionally inverting on the held-out model, even after normalizing by the squared parameter norm (Appendix G.9). The isotropy assumption, while well-motivated at scale, introduces measurable error at intermediate model sizes and in regression settings where the posterior is highly anisotropic. Further, it distorts the relative scaling of the aleatoric and epistemic uncertainty estimates, such that their ratio is no longer meaningful. The scaling study validates up to $\sim 10^6$ parameters while the LLM experiments operate at 10^9–10^{10}; although the trend is monotonically improving and multiple lines of evidence support continued improvement, there is no formal guarantee. On the downstream task, the substantial variance across models and bootstrap runs currently precludes reliable deployment for assessing individual predictions.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report, December 2024. arXiv preprint arXiv:2412.08905.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.568.

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Fisher Information and Natural Gradient Learning in Random Deep Networks. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 694–702. PMLR, April 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Federico Bergamin, Pierre-Alexandre Mattei, Jakob Drachmann Havtorn, Hugo Sénétaire, Hugo Schmutz, Lars Maaløe, Soren Hauberg, and Jes Frellsen. Model-agnostic out-of-distribution detection using combined statistical tests. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 10753–10776. PMLR, May 2022.

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo, July 2018. arXiv preprint arXiv:1701.02434.

Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1613–1622. PMLR, June 2015.

Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. Input Similarity from the Neural Network Perspective. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zhiyu Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Ruth Petzold, and William Yang Wang. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. *Transactions on Machine Learning Research*, June 2024. ISSN 2835-8856.

Sam Dauncey, Christopher C. Holmes, Christopher Williams, and Fabian Falck. Approximations to the Fisher Information Metric of Deep Generative Models for Out-Of-Distribution Detection. *Transactions on Machine Learning Research*, February 2024. ISSN 2835-8856.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20089–20103. Curran Associates, Inc., 2021a.

Erik Daxberger, Eric Nalisnick, James U. Allingham, Javier Antoran, and Jose Miguel Hernandez-Lobato. Bayesian Deep Learning via Subnetwork Inference. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2510–2521. PMLR, July 2021b.

John Denker and Yann LeCun. Transforming Neural-Net Output Levels to Probability Distributions. In *Advances in Neural Information Processing Systems*, volume 3. Morgan-Kaufmann, 1990.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1184–1193. PMLR, July 2018.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115, December 2023.

J. L. Doob. The Limiting Distributions of Certain Statistics. *The Annals of Mathematical Statistics*, 6(3):160–169, 1935. ISSN 0003-4851.

Runa Eschenhagen, Alexander Immer, Richard Turner, Frank Schneider, and Philipp Hennig. Kronecker-Factored Approximate Curvature for Modern Neural Network Architectures. *Advances in Neural Information Processing Systems*, 36:33624–33655, December 2023.

Sebastian Farquhar, Lewis Smith, and Yarin Gal. Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations. In *Advances in Neural Information Processing Systems*, volume 33, pages 4346–4357. Curran Associates, Inc., 2020.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0.

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. PMLR, June 2016.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar,

Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report, March 2025. arXiv preprint arXiv:2503.19786.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei

Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. arXiv preprint arXiv:2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur,

Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841.

Ming Gui, Ziqing Zhao, Tianming Qiu, and Hao Shen. Laplace Approximation with Diagonalized Hessian for Over-parameterized Neural Networks. In *NeurIPS Workshop on Bayesian Deep Learning*, 2021.

David Han, Adrienne Raglin, and Doug Summers-Stay. Negative Nudging to Quantify the LLM Hallucination. In Helmut Degen and Stavroula Ntoa, editors, *Artificial Intelligence in HCI*, pages 152–167, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-93418-6. doi: 10.1007/978-3-031-93418-6_11.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. ISSN 1533-7928.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, October 2021.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3.

Louis Jaburi, Gonçalo Paulo, Stepan Shabalin, Lucia Quirke, and Nora Belrose. Mitigating Emergent Misalignment with Data Attribution. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, September 2025.

I. Shavindra Jayasekera, Jacob Si, Filippo Valdettaro, Wenlong Chen, Aldo A. Faisal, and Yingzhen Li. Variational Uncertainty Decomposition for In-Context Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, March 2023. ISSN 0360-0300. doi: 10.1145/3571730.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, May 2017. arXiv preprint arXiv:1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, November 2022. arXiv preprint arXiv:2207.05221.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, April 2019.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks. *Neural Computation*, 33(8):2274–2307, July 2021. ISSN 0899-7667. doi: 10.1162/neco_a_01411.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, October 2017. arXiv preprint arXiv:1703.04977.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. doi: 10.1073/pnas.1611835114.

Matthew Kowal, Goncalo Paulo, Louis Jaburi, Tom Tseng, Lev E. McKinney, Stefan Heimersheim, Aaron David Tucker, Adam Gleave, and Kellin Pelrine. Concept Influence: Leveraging Interpretability to Improve Performance and Efficiency in Training Data Attribution, February 2026. arXiv preprint arXiv:2602.14869.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5436–5446. PMLR, November 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*, September 2022.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. DataInf: Efficiently Estimating Data Influence in LoRA-tuned LLMs and Diffusion Models. In *The Twelfth International Conference on Learning Representations*, October 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, November 2017. arXiv preprint arXiv:1612.01474.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. Do Influence Functions Work on Large Language Models? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14367–14382, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.775.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100, May 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods, May 2022. arXiv preprint arXiv:2109.07958.

Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, Yubing Jian, Yao Qin, and Roland Memisevic. Enhancing Hallucination Detection through Noise Injection, March 2026. arXiv preprint arXiv:2502.03799.

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, May 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, August 2009. ISSN 1436-4646. doi: 10.1007/s10107-007-0149-x.

Geir K. Nilsen, Antonella Z. Munthe-Kaas, Hans J. Skaug, and Morten Brun. Epistemic uncertainty quantification in deep learning classification by the Delta method. *Neural Networks*, 145:164–176, January 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.10.014.

Luis A. Ortega, Simon Rodriguez Santana, and Daniel Hernández-Lobato. Variational Linearized Laplace Approximation for Bayesian Deep Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 38815–38836. PMLR, July 2024.

Vardan Papyan. Traces of Class/Cross-Class Structure Pervade Deep Learning Spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020. ISSN 1533-7928.

Seonghyeon Park, Jewon Yeom, Jaewon Sok, Jeongjae Park, Heejun Kim, and Taesup Kim. Efficient Epistemic Uncertainty Estimation for Large Language Models via Knowledge Distillation, February 2026. arXiv preprint arXiv:2602.01956.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems*, volume 34, pages 20596–20607. Curran Associates, Inc., 2021.

Jeffrey Pennington and Pratik Worah. The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural Networks. In *International Conference on Learning Representations*, February 2018.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond, October 2017. arXiv preprint arXiv:1611.07476.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks, May 2018. arXiv preprint arXiv:1706.04454.

Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-Order Uncertainty Quantification: Variance-Based Measures, December 2023. arXiv preprint arXiv:2401.00276.

Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise Aleatoric and Epistemic Uncertainty Quantification. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 3159–3179. PMLR, September 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, March 2020. arXiv preprint arXiv:1910.01108.

Simon Schmitt, John Shawe-Taylor, and Hado van Hasselt. General Uncertainty Estimation with Delta Variances. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):20318–20328, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i19.34238.

Lewis Smith and Yarin Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569. AUAI Press, 2018.

Samuel L. Smith, Benoit Dherin, David Barrett, and Soham De. On the Origin of Implicit Regularization in Stochastic Gradient Descent. In *International Conference on Learning Representations*, October 2020.

TheBloke. Llama-2-7B-Chat-AWQ. https://huggingface.co/TheBloke/Llama-2-7B-Chat-AWQ, July 2023.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. Entropy Is Not Enough: Uncertainty Quantification for LLMs fails under Aleatoric Uncertainty. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference {\backslash&} Generative Modeling*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. arXiv preprint arXiv:2307.09288.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust Issues: Uncertainty Estimation Does Not Enable Reliable OOD Detection On Medical Tabular Data. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, pages 341–354. PMLR, November 2020.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bern Bischl, and Eyke Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures?, June 2023. arXiv preprint arXiv:2209.03302.

Yijun Xiao and William Yang Wang. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.236.

Ziao Yang, Han Yue, Jian Chen, and Hongfu Liu. Revisit, Extend, and Enhance Hessian-Free Influence Functions, October 2024. arXiv preprint arXiv:2405.17490.

Maksim Zhdanov, Stanislav Dereka, and Sergey Kolesnikov. Identity Curvature Laplace Approximation for Improved Out-of-Distribution Detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7019–7028, February 2025. doi: 10.1109/WACV61041.2025.00682.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.198.

A. Proxy Covariance Bias Experiment

This appendix provides the full experimental details for the synthetic proxy bias experiment summarized in Section 3.1.

Setup. We train a binary classifier—a two-hidden-layer MLP with tanh activations and 1 185 parameters—on three 2D problems: a linearly separable boundary, an XOR pattern, and concentric rings. Each problem is designed so that the decision boundary is spatially symmetric. For each problem, we split the training data into two halves by spatial location (top vs. bottom) and compute the empirical Fisher information matrix (FIM) from each half separately, obtaining two proxy Hessians H_A (from top-half data) and H_B (from bottom-half data). We then evaluate epistemic uncertainty under three covariance assumptions: the identity ($C = I$, yielding $\|g\|^2$), the Laplace approximation using H_A^{-1}, and the Laplace approximation using H_B^{-1}, writing $U_A(x) = g^\top H_A^{-1} g$ and $U_B(x) = g^\top H_B^{-1} g$ for the two proxy estimates. All uncertainty maps are evaluated on a dense grid covering the input space and individually normalized to $[0, 1]$ for comparison.

Results. On all three problems, the identity produces uncertainty estimates that respect the spatial symmetry of the decision boundary, while each proxy Hessian introduces a severe asymmetry: it suppresses uncertainty in the half of input space where its data is concentrated and inflates it in the complementary half. Table 3 reports the full results.

	Top/bottom ratio			Welch t-test on $\log(U_A/U_B)$		
Problem	r_{id}	r_A	r_B	t	p	Cohen's d
Linear	0.93	0.21	4.56	-8.1	3.0×10^{-14}	-1.0
XOR	1.45	0.62	4.39	-19.6	3.8×10^{-50}	-2.5
Rings	1.06	0.05	19.10	-22.2	2.0×10^{-61}	-2.8

Table 3: Proxy covariance bias across three synthetic problems. r_{id}, r_A, r_B: ratio of mean top-half to mean bottom-half normalized uncertainty under the identity, proxy H_A^{-1} (top-half data), and proxy H_B^{-1} (bottom-half data), respectively. A ratio of 1.0 indicates perfect spatial symmetry. The Welch t-test is performed on the log-ratio map $\log(U_A(x)/U_B(x))$ between the two halves of the input space.

The symmetry ratio under isotropic covariance, r_{id} (ratio of mean uncertainty in each half of the input space), remains close to 1.0 on all three problems, confirming that it preserves the true spatial symmetry. The corresponding ratios under each proxy covariance, r_A and r_B (same metric, using H_A^{-1} and H_B^{-1} respectively), deviate strongly in opposite directions, with the effect increasing with decision boundary complexity: from the linear problem ($d = -1.0$) through XOR ($d = -2.5$) to rings ($d = -2.8$). All effects are highly significant ($p < 10^{-13}$). The weaker effect on the linear problem is expected: a linear decision boundary constrains fewer parameter directions, so there is less room for the proxy to distort the covariance structure.

A.1. Language Model Extension

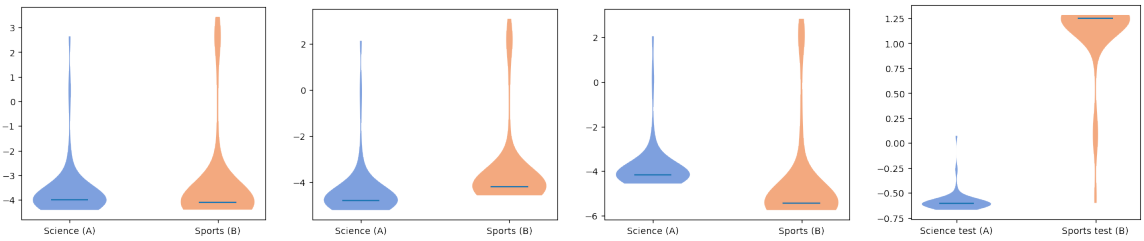
To verify that the proxy bias phenomenon extends beyond synthetic settings, we repeat the experiment on a text classification task using a pretrained language model.

Setup. We fine-tune DistilBERT on a binary domain classification task using the 20 Newsgroups corpus. Domain A consists of scientific text (`sci.med`, `sci.space`) and domain B of sports text (`rec.sport.hockey`, `rec.sport.baseball`), with 400 training and 100 test samples per domain. The model is trained for 3 epochs with AdamW ($\eta = 2 \times 10^{-5}$, weight decay 10^{-2}), achieving 96.5% overall test accuracy (98% on science, 95% on sports), confirming that both domains are well-learned.

We apply a last-layer Laplace approximation (Kristiadi et al., 2020), restricting the FIM to the classifier head parameters ($768 \times 2 + 2 = 1{,}538$ parameters). For each domain, we compute a separate empirical FIM from 200 proxy samples using the model's own predictive distribution, with prior precision $\lambda = 1$ matching the weight decay. We then evaluate epistemic uncertainty under three covariance assumptions on each test set: the identity ($\|g\|^2$), the Laplace approximation using H_A^{-1} (science proxy), and H_B^{-1} (sports proxy).

Results. The results are shown in Fig. 4 and Table 4. The log-ratio $\log(U_A/U_B)$ flips sign between domains: the median is -0.61 on science test data and $+1.25$ on sports test data, meaning the science proxy assigns relatively lower uncertainty to science inputs and higher to sports inputs, and vice versa for the sports proxy. A Welch t-test on the log-ratio between domains yields $t = -37.5$, $p < 10^{-4}$, Cohen's $d = -5.3$, confirming that the proxy-induced distortion is highly significant and large in effect size, substantially larger than in any of the synthetic problems (Table 3).

Unlike the spatially symmetric synthetic problems, the two text domains need not have identical baseline uncertainty, so the mean ratio r_{id} under the identity is not expected to equal 1.0. The relevant comparison is the relative shift introduced by each proxy: $r_A < r_{\text{id}} < r_B$, confirming that the science proxy suppresses science uncertainty and the sports proxy suppresses sports uncertainty, consistent with the synthetic findings.



(a) Identity ($C = I$) (b) Proxy H_A^{-1} (science) (c) Proxy H_B^{-1} (sports) (d) $\log(U_A/U_B)$

Figure 4: Epistemic uncertainty distributions on the DistilBERT domain classification task under three covariance assumptions, and the log ratio between proxy estimates. Under the identity, both domains have similar uncertainty distributions. Each proxy shifts the relative uncertainty between domains: H_A^{-1} (estimated from science data) suppresses uncertainty on science inputs relative to sports, while H_B^{-1} does the reverse. The log ratio makes this asymmetry explicit, flipping sign between domains.

	Domain ratio (science / sports)			Welch t-test on $\log(U_A/U_B)$		
	r_{id}	r_A	r_B	t	p	Cohen's d
DistilBERT	0.28	0.25	0.30	-37.5	$< 10^{-4}$	-5.3

Table 4: Proxy covariance bias on the DistilBERT domain classification task. r_{id}, r_A, r_B: ratio of mean science to mean sports uncertainty under the identity, proxy H_A^{-1} (science data), and proxy H_B^{-1} (sports data), respectively. The Welch t-test is performed on $\log(U_A(x)/U_B(x))$ between science and sports test sets. Bootstrap 95% CIs for the domain ratio: identity $[0.05, 1.04]$, H_A^{-1} $[0.03, 0.93]$, H_B^{-1} $[0.06, 1.13]$.

B. Theoretical and Empirical Support for Isotropic Covariance

This appendix provides the detailed theoretical arguments and empirical precedents supporting the isotropy assumption summarized in Section 3.2.

B.1. Hessian Structure Simplifies with Scale

A convergent body of theoretical and empirical work shows that the Hessian of deep networks exhibits a characteristic spectral pattern: a small number of large eigenvalues with a vast bulk near zero. Much of this theory has been developed for the FIM, a standard positive semi-definite approximation to the Hessian: Amari et al. (2019) prove that the FIM is approximately unit-wise

block-diagonal, with off-block elements of order $O(1/\sqrt{n})$ for layer width n; within each block, Dauncey et al. (2024) demonstrate diagonal dominance, with diagonal entries roughly five times larger than off-diagonal ones; and the eigenvalue distribution has been characterized analytically via mean-field theory (Karakida et al., 2019) and nonlinear random matrix theory (Pennington and Worah, 2018). Karakida et al. (2021) show that this bulk spectral concentration holds equally for the empirical FIM and the generalized Gauss–Newton matrix. Empirical studies of the full Hessian confirm the same two-component structure in trained networks (Sagun et al., 2017, 2018; Papyan, 2020).

This spectral structure has a direct consequence for the covariance. Inverting the Hessian amplifies the near-zero eigenvalues, so the damping term λI universally added for stabilization dominates in most parameter directions, causing the damped inverse $(F + \lambda I)^{-1}$ to converge to approximately $(1/\lambda)I$. This effect becomes more pronounced at scale: Li et al. (2025) show empirically that for LLMs, the damping term overwhelms the Hessian, so that the damped inverse is effectively proportional to the identity. Corroborating this, Kwon et al. (2023) report that the LiSSA algorithm for iterative inverse-Hessian approximation collapses to the Hessian-free (identity) baseline across all tasks on a 13B-parameter model, which they attribute to the high dimensionality of large-scale models. These results also have implications for the structural error in Eq. (5) and Eq. (6): the off-block-diagonal decay at $O(1/\sqrt{n})$ and the within-block diagonal dominance together suggest that $\|\Sigma_{\text{off}}\|_{\text{op}}$ shrinks relative to $\|\Sigma^*_{\text{diag}}\|_{\text{op}}$ as scale increases, so that term (i) becomes a diminishing fraction of the total error for both the identity and any diagonal approximation.

B.2. Precedent from Laplace Approximations

The Laplace approximation for neural networks provides perhaps the most direct precedent for simplifying the covariance structure without sacrificing downstream performance: practitioners routinely employ drastic simplifications of the Hessian with little loss. The full Hessian or generalized Gauss–Newton (GGN) matrix is typically replaced by a diagonal (Kirkpatrick et al., 2017; Ritter et al., 2018), Kronecker-factored (Ritter et al., 2018; Eschenhagen et al., 2023), or block-diagonal approximation. More aggressive still, last-layer Laplace restricts the posterior to only the final layer's parameters (Kristiadi et al., 2020; Daxberger et al., 2021a), and subnetwork Laplace (Daxberger et al., 2021b) selects an arbitrary subset of parameters for Bayesian treatment. Daxberger et al. (2021a) systematically compare these approximations and find that the cheapest variants—diagonal and last-layer—often match or exceed the predictive performance of more faithful Hessian estimates, suggesting that the precise structure of the covariance matters far less than one might expect. Our isotropy assumption goes further by replacing per-parameter curvature magnitudes with a uniform scalar, but the spectral results above imply that at scale the damped inverse is already approximately proportional to the identity, and obtaining accurate magnitudes without the true training data introduces structured distortions (Section 3.1) that may outweigh the benefit.

B.3. Bayesian and Optimization Arguments

Most modern LLMs are trained with weight decay, which is equivalent to imposing a Gaussian prior with covariance proportional to the identity (Bishop and Nasrabadi, 2006; Goodfellow et al., 2016). Pretrained language models have very low intrinsic dimensionality relative to their full parameter count (Aghajanyan et al., 2021; Hu et al., 2021), so the training data determines the posterior in only a small subspace; in the remaining directions, the posterior is driven by the isotropic prior. Moreover, Farquhar et al. (2020) prove that diagonal weight-space covariance in deep networks can induce function-space distributions comparable to structured covariance approximations in shallower networks, suggesting that the off-diagonal structure our approximation discards is largely redundant. A complementary argument comes from optimization: Smith et al. (2020) show that stochastic gradient descent (SGD) implicitly regularizes the squared norms of per-sample loss gradients, suppressing $\|g\|^2$ in well-learned regions while leaving it unconstrained for unfamiliar

inputs, directionally consistent with the contrast needed for epistemic uncertainty estimation without an explicit covariance correction.

B.4. Empirical Success of the Isotropy Assumption

The isotropy assumption, whether explicit or implicit, has been employed across multiple tasks with strong empirical performance.

In out-of-distribution detection, Bergamin et al. (2022) propose a model-agnostic method that scores anomalies using gradient information weighted by different approximations to the Hessian; the identity performs competitively with more elaborate curvature approximations. Zhdanov et al. (2025) introduce the Identity Curvature Laplace Approximation (ICLA), which replaces the Hessian entirely with the identity and outperforms standard last-layer Laplace using the empirical FIM, GGN, and K-FAC on OOD detection benchmarks.

In data attribution, several methods define training sample influence via the gradient dot product $\nabla_\theta \ell(z_{\text{test}})^\top \nabla_\theta \ell(z_{\text{train}})$, corresponding to the influence function with the inverse Hessian replaced by the identity. Charpiat et al. (2019) define input similarity as the cosine similarity of parameter gradients; Pruthi et al. (2020) formalize this as TracIn; Yang et al. (2024) show that this identity approximation is order-consistent with true influence in many practical regimes and that the inverse Hessian can introduce errors making it worse than the identity; Jaburi et al. (2025) find essentially no performance loss from using the identity for mitigating emergent behaviors in LLMs; and Kowal et al. (2026) show that an even more aggressive double-identity approximation to concept-based influence functions matches or exceeds full EK-FAC performance at 7B scale while being 20× faster.

In dataset pruning, Paul et al. (2021) introduce the Gradient Norm (GraNd) score, the expected L^2 norm of the per-sample loss gradient, and use it to prune significant fractions of training data without sacrificing test accuracy.

C. Proof that $\mathbb{E}_\theta[\theta] = \theta^*$

From the first-order Taylor expansion used in Section 3:

$$p(y_c \mid x, \theta) \approx p(y_c \mid x, \theta^*) + \nabla_\theta p(y_c \mid x, \theta)\big|_{\theta^*}^\top (\theta - \theta^*) \tag{11}$$

Taking expectations on both sides:

$$\mathbb{E}_\theta[p(y_c \mid x, \theta)] \approx p(y_c \mid x, \theta^*) + \nabla_\theta p(y_c \mid x, \theta)\big|_{\theta^*}^\top (\mathbb{E}_\theta[\theta] - \theta^*) \tag{12}$$

By the assumption that $p(y_c \mid x, \theta^*) = \mathbb{E}_\theta[p(y_c \mid x, \theta)]$, which is necessary for the variance-based uncertainty decomposition (Section 2), the left-hand side equals the first term on the right, so:

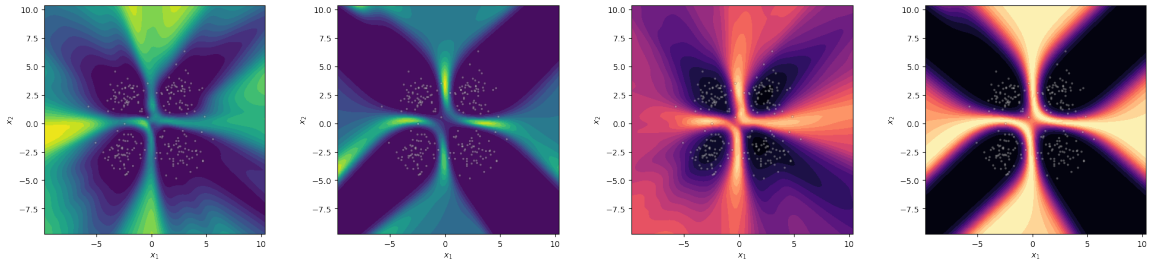
$$\nabla_\theta p(y_c \mid x, \theta)\big|_{\theta^*}^\top (\mathbb{E}_\theta[\theta] - \theta^*) = 0, \tag{13}$$

$$\mathbb{E}_\theta[\theta] = \theta^*. \tag{14}$$

Since $\mathbb{E}_\theta[\theta]$ and θ^* are global quantities independent of x, the difference $\mathbb{E}_\theta[\theta] - \theta^*$ is a single fixed vector; a single input x with nonzero gradient therefore suffices to constrain it via $g^\top(\mathbb{E}_\theta[\theta] - \theta^*) = 0$. That is, the conclusion holds exactly under the first-order Taylor approximation; the only source of error is the approximation itself.

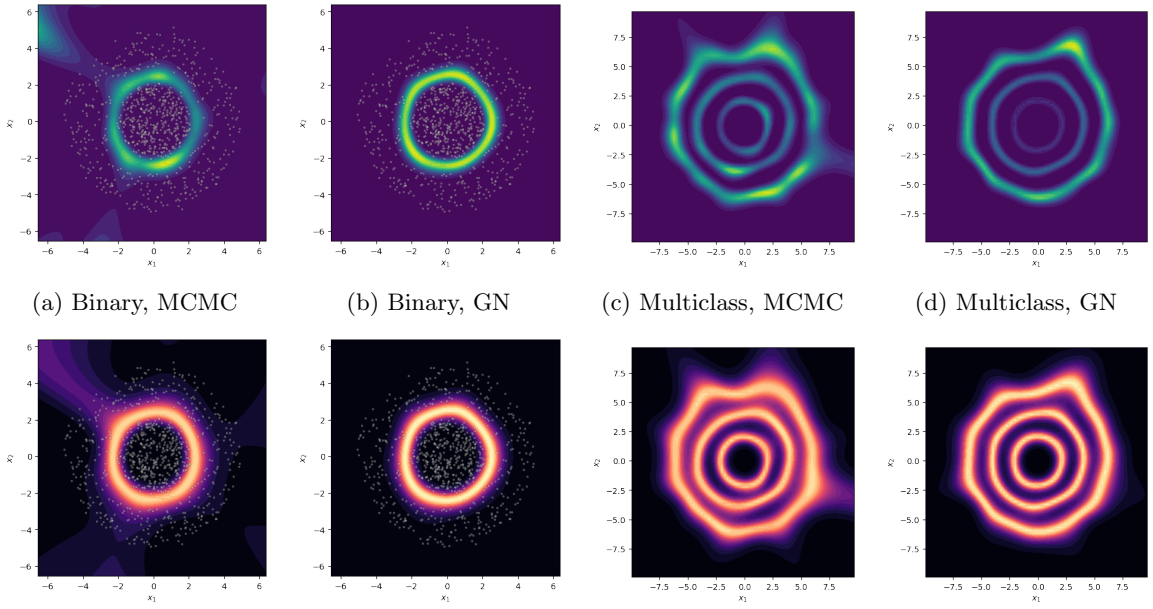
D. Additional Classification Results

Fig. 5 shows the uncertainty maps for the binary XOR problem, and Fig. 6 for the concentric rings problem, both omitted from the main text for space. The binary rings problem ($\rho = 0.44$) represents the most challenging setting for the gradient norm approximation in classification.



(a) Epistemic, MCMC (b) Epistemic, GN (c) Aleatoric, MCMC (d) Aleatoric, point est.

Figure 5: Binary XOR uncertainty maps. Left two panels: epistemic uncertainty (MCMC vs. $\|g\|^2$). Right two panels: aleatoric uncertainty (MCMC vs. point estimate). All maps are individually normalized to $[0, 1]$.



(a) Binary, MCMC (b) Binary, GN (c) Multiclass, MCMC (d) Multiclass, GN

(e) Binary, MCMC (f) Binary, point est. (g) Multiclass, MCMC (h) Multiclass, point est.

Figure 6: Concentric rings uncertainty maps. Top row: epistemic uncertainty (MCMC vs. $\|g\|^2$) for binary (left) and multiclass (right). Bottom row: aleatoric uncertainty (MCMC vs. point estimate). All maps are individually normalized to $[0, 1]$.

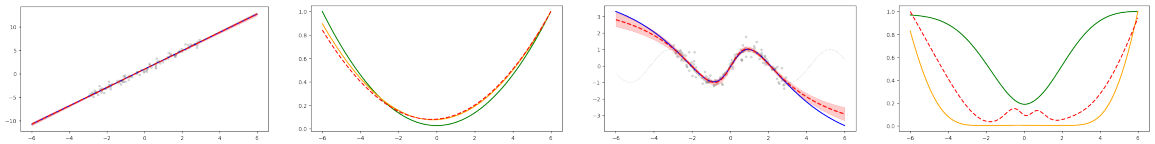
E. Regression Uncertainty

F. Validation Experiment Details

This appendix provides additional experimental details and per-problem results for the synthetic validation experiments in Section 4.1.

F.1. Setup

Binary classification. We train logistic regression and two-hidden-layer MLPs (with tanh activations) on three 2D binary classification problems: a linearly separable boundary, an XOR pattern, and concentric rings. These problems span a range of decision boundary complexity, from a setting where the linear model is correctly specified to ones requiring nonlinear capacity.



(a) Linear, data (b) Linear, uncertainty (c) Nonlinear, data (d) Nonlinear, uncertainty

Figure 7: Regression problems (columns 1 and 3) and normalized epistemic uncertainty (columns 2 and 4). In the data plots: gray dots are training samples, the dotted curve is the true data-generating function, the red dashed line is the MAP prediction, and the shaded band is the MCMC posterior predictive interval. In the uncertainty plots: the green solid line is $\|g\|^2$, the orange solid line is the Laplace approximation, and the red dashed line is the MCMC, all normalized to $[0, 1]$.

Multiclass classification. We use softmax regression and two-hidden-layer MLPs on three 2D datasets: well-separated Gaussian clusters (3 classes), interleaved spirals (3 classes), and concentric rings (3 classes). For multiclass models, we evaluate per-class epistemic uncertainty $\mathrm{Var}_\theta[p(y_c \mid x, \theta)]$ for each class c and report correlations aggregated across classes.

Regression. We use a single-hidden-layer MLP with tanh activations (97 parameters) on two 1D problems: a linear function and a nonlinear function.

Scaling. We train a sequence of models of increasing width on the binary concentric rings problem, ranging from logistic regression (12 parameters) to a two-hidden-layer MLP with 1,028 units per layer (approximately 1.07×10^6 parameters).

F.2. Additional Results

F.2.1. Binary Classification

Table 5 reports the full epistemic correlation results for binary classification, including gradient norm (GN), Laplace approximation (LA), and the correlation between the two. On all three problems the GN–LA correlation exceeds 0.97, meaning the Hessian correction provides almost no additional information beyond what the gradient norm already captures. This is consistent with the role of the sigmoid nonlinearity discussed in Section 4.1: the output compression attenuates the gradient components that would otherwise expose posterior anisotropy, so the identity and the inverse Hessian produce nearly identical uncertainty maps.

Fig. 8b shows the epistemic uncertainty maps for the binary XOR problem, and Fig. 8a for the linear problem.

	GN vs MCMC		LA vs MCMC		GN vs LA	
Problem	r	ρ	r	ρ	r	ρ
Linear (LogReg)	0.95	0.99	0.95	0.99	1.00	1.00
XOR (MLP)	0.65	0.68	0.68	0.70	0.94	0.98
Rings (MLP)	0.86	0.44	0.86	0.46	0.97	1.00

Table 5: Binary classification: Pearson (r) and Spearman (ρ) correlation between epistemic uncertainty estimates and MCMC estimates. GN: gradient norm; LA: Laplace approximation; GN–LA: correlation between gradient norm and Laplace.

Table 6 reports aleatoric correlations. The point estimate $p(y_c \mid x, \theta^*)(1 - p(y_c \mid x, \theta^*))$ tracks the MCMC estimates well on the linear problem ($r = 0.99$) and on the rings problem ($r = 0.95$), but less so on XOR ($r = 0.76$). The Laplace-based aleatoric estimate performs poorly on the MLP problems,

with negative correlations on XOR, suggesting that the Laplace posterior is a poor approximation to the true posterior in these settings.

	PE vs MCMC		LA vs MCMC		PE vs LA	
Problem	r	ρ	r	ρ	r	ρ
Linear (LogReg)	0.99	1.00	0.92	0.97	0.88	0.95
XOR (MLP)	0.76	0.74	0.08	0.12	−0.10	−0.11
Rings (MLP)	0.95	0.58	0.19	0.11	0.19	0.10

Table 6: Binary classification: aleatoric uncertainty correlations. PE: point estimate at MAP; LA: Laplace posterior mean. Both compared against MCMC posterior mean of $p(y_c \mid x, \theta)(1-p(y_c \mid x, \theta))$.

F.2.2. Multiclass Classification

Table 7 reports per-class epistemic correlations for the three multiclass problems; correlations are consistent across classes within each problem, confirming that the aggregate results are not driven by averaging over heterogeneous per-class performance. Table 8 reports per-class aleatoric correlations; the point estimate achieves consistently high correlations ($r \geq 0.95$, $\rho \geq 0.89$) across all problems and classes.

Problem	Class	r	ρ
Clusters (Softmax)	0	0.88	0.98
	1	0.86	0.96
	2	0.85	0.96
	3	0.87	0.98
	Overall	0.86	0.97
Spirals (MLP)	0	0.82	0.92
	1	0.79	0.91
	2	0.72	0.88
	3	0.72	0.93
	Overall	0.76	0.91
Rings (MLP)	0	0.92	0.88
	1	0.90	0.96
	2	0.88	0.95
	3	0.89	0.95
	Overall	0.88	0.97

Table 7: Multiclass: per-class Pearson (r) and Spearman (ρ) correlation between gradient norm and MCMC epistemic uncertainty.

Problem	Class	r	ρ
Clusters (Softmax)	0	0.95	1.00
	1	0.95	0.99
	2	0.95	0.99
	3	0.95	0.99
	Overall	0.95	0.99
Spirals (MLP)	0	0.96	0.98
	1	0.95	0.98
	2	0.97	0.96
	3	0.96	0.97
	Overall	0.96	0.97
Rings (MLP)	0	0.99	0.89
	1	0.98	0.97
	2	0.95	0.97
	3	0.95	0.97
	Overall	0.96	0.98

Table 8: Multiclass: per-class Pearson (r) and Spearman (ρ) correlation between point-estimate and MCMC aleatoric uncertainty.

F.2.3. Regression

Table 9 reports the full regression results, including Hessian eigenvalue ranges that illustrate the degree of posterior anisotropy. On the linear problem (2 parameters), the Hessian eigenvalues span a factor of 3.2×, and all three methods—gradient norm, Laplace, and MCMC—nearly coincide ($r \geq 0.98$). On the nonlinear problem (97 parameters), the eigenvalue range spans a factor of 1.5×10^4, reflecting severe posterior anisotropy. Here the Laplace approximation achieves $r = 0.93$ ($\rho = 0.97$) while the gradient norm drops to $r = 0.73$ ($\rho = 0.81$), confirming that the isotropy assumption is the primary source of error.

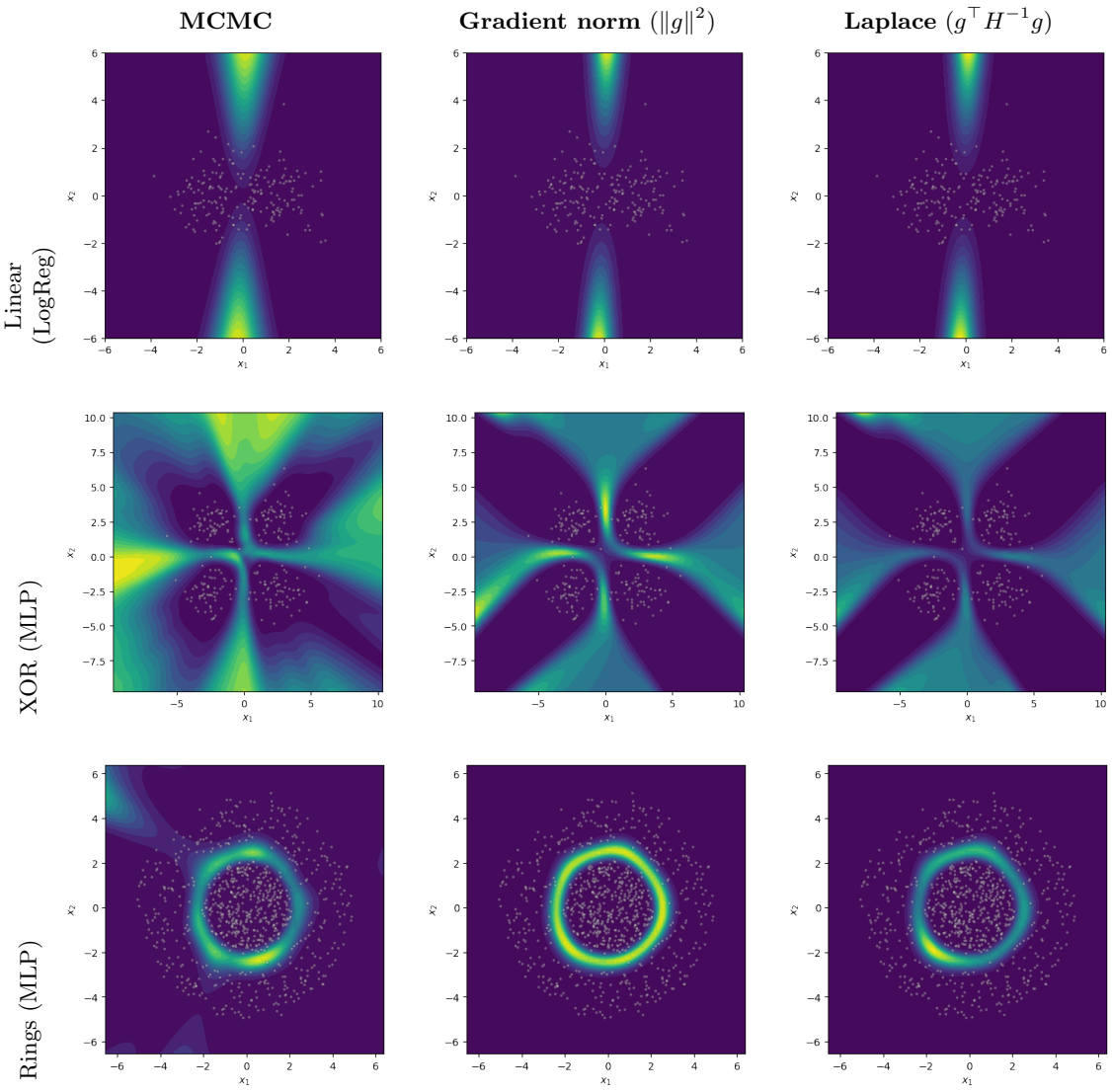


Figure 8: Epistemic uncertainty for all three binary classification problems. Each row shows a different problem; columns show MCMC, gradient norm, and Laplace approximation. On the linear problem all three methods are nearly identical; on XOR and rings the gradient norm and Laplace both recover the correct structure, while the Laplace approximation degrades on the nonlinear MLP problems.

	GN vs MCMC		LA vs MCMC		Hessian eigenvalues	
Problem	r	ρ	r	ρ	Range	Ratio
Linear (2 params)	0.98	0.99	1.00	1.00	$[395, 1281]$	$3.2\times$
Nonlinear (97 params)	0.73	0.81	0.93	0.97	$[1.0, 14648]$	1.5×10^4

Table 9: Regression: epistemic uncertainty correlations and Hessian eigenvalue ranges. The eigenvalue ratio $\lambda_{\max}/\lambda_{\min}$ quantifies posterior anisotropy.

F.2.4. Scaling

Table 10 reports the full scaling results for all nine model sizes. The epistemic Spearman correlation follows the U-shaped trajectory described in Section 4.1, reaching a minimum at intermediate scales before recovering at the largest model sizes.

		Epistemic		Aleatoric	
Model	D	r	ρ	r	ρ
LogReg	12	0.86	0.97	0.95	0.99
MLP(8,8)	132	0.51	0.84	0.83	0.87
MLP(16,16)	388	0.48	0.89	0.83	0.92
MLP(32,32)	1 284	0.43	0.68	0.74	0.81
MLP(64,64)	4 612	0.49	0.82	0.76	0.90
MLP(128,128)	17 412	0.60	0.66	0.79	0.84
MLP(256,256)	67 588	0.68	0.78	0.82	0.89
MLP(512,512)	266 244	0.65	0.83	0.88	0.89
MLP(1028,1028)	1 065 012	0.74	0.87	0.89	0.93

Table 10: Scaling study: correlation between gradient norm estimates and MCMC estimates as a function of model size (number of parameters D) on the concentric rings problem.

G. Question Answering Experiment Details

This appendix provides the full experimental details for the downstream question answering experiments in Section 4.2.

G.1. Models

We evaluate four language models spanning a range of architectures and scales:

- Llama 2 7B (Touvron et al., 2023), using a pre-quantized AWQ variant (Lin et al., 2024; TheBloke, 2023);
- Llama 3.2 3B (Grattafiori et al., 2024);
- OLMo 1B (Groeneveld et al., 2024);
- Phi-4 (Abdin et al., 2024).

All models except Llama 2 are quantized to 4-bit precision using bitsandbytes (Dettmers et al., 2023) for computational efficiency.

G.2. Baselines

Following Farquhar et al. (2024), we compare against three baselines:

- **Naïve entropy** (Farquhar et al., 2024): multiple completions are sampled and entropy is computed over the length-normalised token-sequence log-probabilities, treating lexically distinct but semantically equivalent outputs as different.
- **Semantic entropy** (Kuhn et al., 2022): multiple completions are sampled, clustered by semantic equivalence, and entropy is computed over the cluster probabilities.
- **P(True)** (Kadavath et al., 2022): the model is prompted to assess whether its own answer is correct, and the probability assigned to "True" is used as a confidence score.

G.3. Evaluation

Correctness is determined using the semantic equivalence criterion of Farquhar et al. (2024): an LLM judge checks whether the generated answer means the same as one of the reference answers in the context of the question. For each uncertainty method, we train a logistic regression classifier to predict correctness from the uncertainty score and report the AUROC over 300 bootstrap runs per model–dataset configuration. For the combined estimate (Epi. & Alea.), the logistic regression uses both scores as features, making it a two-feature model versus one feature for all other methods.

G.4. Datasets

TriviaQA (Joshi et al., 2017) tests factual recall with unambiguous answers, e.g. "What was the name of Michael Jackson's autobiography written in 1988?" A model may be highly confident in a correct answer (it has memorized the fact) or highly confident in a wrong one (it has memorized a distortion), so uncertainty and correctness are largely independent.

TruthfulQA (Lin et al., 2022) targets common misconceptions and genuinely open questions, e.g. "What happens to you if you eat watermelon seeds?" The popular answer (they grow in your stomach) is false, while the set of accepted truthful answers is broad, ranging from "nothing happens" to "they pass through your digestive system", reflecting genuine variability in how the question can be correctly addressed. The model therefore faces both epistemic conflict between what is commonly said and what is factually correct, and inherent ambiguity in what constitutes a complete answer.

G.5. Per-Model Results

	TruthfulQA					
Model	Naïve Ent.	Sem. Ent.	P(True)	Alea.	Epi.	Epi. & Alea.
Llama 2 (AWQ)	0.57	0.52	0.57	**0.61**	0.51	0.58
Llama 3.2 3B	0.53	0.58	0.50	**0.69**	0.53	0.68
OLMo 1B	0.47	0.47	0.54	0.51	0.53	**0.57**
Phi-4 (4-bit)	0.47	0.58	0.58	0.59	0.63	**0.69**
	TriviaQA					
Model	Naïve Ent.	Sem. Ent.	P(True)	Alea.	Epi.	Epi. & Alea.
Llama 2 (AWQ)	0.58	0.56	**0.75**	0.61	0.60	0.61
Llama 3.2 3B	0.53	0.51	0.59	0.64	0.51	**0.67**
OLMo 1B	0.48	0.58	**0.72**	0.55	0.55	0.55
Phi-4 (4-bit)	0.50	0.55	**0.69**	0.61	0.42	0.60

Table 11: Per-model AUROC for all methods (mean over 300 bootstrap runs). Best per row in bold. On TruthfulQA, at least one uncertainty estimate matches or exceeds all baselines for every model; on TriviaQA, baselines dominate for three of four models.

G.6. Wall-Clock Timing

To measure computational cost, we benchmark all methods on TruthfulQA for each model on a single NVIDIA H100 GPU, using default settings from their respective papers. Table 12 reports the mean and standard deviation of per-sample wall-clock time (excluding the shared generation step) for each model.

Model	Gradient norm (ours)	P(True)	Naïve Ent.	Sem. Ent.
Llama 2 (AWQ)	0.12 ± 0.01	7.14 ± 0.96	10.67 ± 1.87	11.29 ± 2.03
Llama 3.2 3B	0.08 ± 0.00	5.37 ± 0.39	8.18 ± 0.77	8.94 ± 0.76
OLMo 1B	0.06 ± 0.00	3.06 ± 0.64	4.52 ± 1.08	5.24 ± 1.18
Phi-4 (4-bit)	0.15 ± 0.00	6.73 ± 0.98	10.26 ± 1.50	10.93 ± 1.67

Table 12: Wall-clock time per sample (mean ± std in seconds) on a single H100 GPU, excluding shared generation cost. The gradient norm requires only a single backward pass after generation, yielding a 46–107× speedup over the baselines.

G.7. Statistical Significance

To assess whether the AUROC differences in Table 2 are statistically significant, we perform paired t-tests across 10 random train/test splits (80/20), with AUROC values averaged over models within each split so that the 10 splits provide paired observations. With 18 simultaneous tests (3 gradient methods × 3 baselines × 2 datasets), running each at $\alpha = 0.05$ would be expected to produce roughly one false positive by chance; we therefore apply the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg, 1995), which adjusts each threshold to limit the expected *fraction* of false discoveries among those declared significant, rather than requiring that every single test be free of error. Table 13 reports BH-corrected p-values; bold entries are significant, and arrows indicate whether the gradient-based method is better (↑) or worse (↓) than the baseline.

	vs Naïve Ent.	vs Sem. Ent.	vs P(True)
TruthfulQA			
Aleatoric	$<$ **.001** ↑	**.005** ↑	**.005** ↑
Epistemic	**.005** ↑	.575	.985
Epi. & Alea.	$<$ **.001** ↑	$<$ **.001** ↑	$<$ **.001** ↑
TriviaQA			
Aleatoric	$<$ **.001** ↑	$<$ **.001** ↑	$<$ **.001** ↓
Epistemic	.712	$<$ **.001** ↓	$<$ **.001** ↓
Epi. & Alea.	$<$ **.001** ↑	$<$ **.001** ↑	$<$ **.001** ↓

Table 13: BH-corrected p-values from paired t-tests comparing each gradient-based method against each baseline, per benchmark. Bold = significant at $\alpha = 0.05$; ↑ = gradient method better, ↓ = baseline better. Tests use the Farquhar correctness criterion, 10 random splits, AUROC averaged over 4 LLMs per split.

On TruthfulQA, the combined estimate significantly outperforms all three baselines ($p < 0.01$), as does the aleatoric estimate alone. The epistemic estimate significantly exceeds naïve entropy ($p = .005$) but is statistically indistinguishable from semantic entropy and P(True). On TriviaQA, P(True) significantly outperforms all gradient-based methods ($p < 0.001$); the epistemic estimate is significantly *worse* than semantic entropy ($p < 0.001$) and indistinguishable from naïve entropy.

A paired bootstrap test (10 000 resamples, paired by model and split) on the key comparison—Epi. & Alea. vs. Aleatoric—yields $\Delta\text{AUROC} = +0.027$, 95% CI $[0.002, 0.054]$, $p = 0.018$ on TruthfulQA, confirming that the epistemic term provides a statistically significant lift on this benchmark. On TriviaQA, the lift is not significant ($\Delta\text{AUROC} = +0.006$, 95% CI $[-0.003, 0.016]$, $p = 0.106$).

G.8. Correlation Between Epistemic Uncertainty and P(True)

To assess whether the gradient-based epistemic estimate and P(True) capture redundant or complementary signal, we compute Spearman rank correlations on the raw per-sample values (Table 14).

Model	Dataset	n	ρ	p
Llama 2 (AWQ)	TruthfulQA	811	−0.17	$< .001$
Llama 3.2 3B	TruthfulQA	794	−0.11	.002
OLMo 1B	TruthfulQA	501	−0.12	.007
Phi-4 (4-bit)	TruthfulQA	455	−0.20	$< .001$
Llama 2 (AWQ)	TriviaQA	7 885	−0.22	$< .001$
Llama 3.2 3B	TriviaQA	7 408	−0.27	$< .001$
OLMo 1B	TriviaQA	4 670	−0.19	$< .001$
Phi-4 (4-bit)	TriviaQA	3 270	−0.10	$< .001$
Pooled	TruthfulQA	2 561	−0.23	$< .001$
Pooled	TriviaQA	23 233	−0.21	$< .001$

Table 14: Spearman rank correlation (ρ) between the epistemic uncertainty estimate ($\|g\|^2$) and P(True) on raw per-sample values.

All correlations are negative and significant: higher epistemic uncertainty (larger gradient norm) corresponds to lower self-assessed confidence, as expected. However, the magnitudes are weak ($|\rho| \approx 0.10$–0.27), indicating approximately 4% shared variance at the pooled level. This confirms that the gradient-based estimate captures information largely complementary to the model's self-assessed confidence, consistent with the observation that the two measures are most useful on different benchmarks (Section 4.2).

G.9. Cross-Model Transfer

To test whether the gradient-based epistemic estimate generalizes across architectures, we run a leave-one-model-out (LOMO) experiment: for each of the four LLMs, we train a logistic regression classifier on the remaining three and evaluate on the held-out model. The raw $\|g\|^2$ depends on the absolute scale of the parameters, which varies across model families and quantization schemes; dividing by $\|\theta^*\|^2$ removes this dependence and should in principle yield a relative measure that is comparable across architectures. We compare the raw $\|g\|^2$ against this parameter-norm-normalized variant ($\|g\|^2/\|\theta^*\|^2$), both alone and combined with the aleatoric estimate.

	Held-out	Raw	Normalized	Raw & Alea.	Norm & Alea.
TruthfulQA	Llama 3.2 3B	0.53	0.47	0.33	0.33
	Llama 2 (AWQ)	0.55	0.55	0.51	0.53
	OLMo 1B	0.47	0.47	0.49	0.49
	Phi-4 (4-bit)	0.37	0.37	0.61	0.61
	Mean (Std)	0.48 (0.08)	0.46 (0.08)	0.48 (0.12)	0.49 (0.12)
TriviaQA	Llama 3.2 3B	0.49	0.49	0.35	0.35
	Llama 2 (AWQ)	0.40	0.40	0.38	0.38
	OLMo 1B	0.45	0.55	0.44	0.44
	Phi-4 (4-bit)	0.40	0.40	0.61	0.61
	Mean (Std)	0.44 (0.05)	0.46 (0.07)	0.45 (0.11)	0.45 (0.11)

Table 15: Leave-one-model-out AUROC: train on 3 models, evaluate on held-out 4th. Values are at or below chance (0.50), and the relationship occasionally inverts across architectures. Normalization by the parameter norm does not improve transfer.

All configurations produce chance-or-below AUROC on held-out models (Table 15), and the relationship occasionally inverts—Phi-4's raw epistemic score drops to 0.37–0.40, meaning higher gradient norm predicts *correct* answers when trained on other models—confirming that the mapping between gradient magnitude and correctness is architecture-specific with no consistent direction

across models. Parameter-norm normalization does not improve transfer: the mean AUROC and cross-model variance are essentially unchanged. The one apparent exception is Phi-4 under the combined feature set (0.61 on both benchmarks), but inspection shows this lift comes entirely from the aleatoric term: the raw epistemic score alone is 0.37–0.40 for Phi-4, below chance, while the aleatoric score transfers because it is architecture-agnostic (depending only on output probabilities, not gradient magnitudes). The high cross-model standard deviation (0.11–0.12) for the combined features is driven by this single model; without Phi-4, all means drop further below chance. This is consistent with the high per-model variance observed in Table 11 and indicates that the epistemic estimate should be calibrated per model rather than applied with a universal threshold.