GENERATIVE MODELLING UNDER EPISTEMIC UNCERTAINTY

by

MUHAMMAD MUBASHAR

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing

School of Engineering, Computing, and Mathematics

OXFORD BROOKES UNIVERSITY

30 January 2026

THESIS COMMITTEE

Supervisors

Prof. Fabio Cuzzolin

Professor of Artificial Intelligence

Director of AIDAS Institute

Oxford Brookes University

Dr. Andrew Bradley

Reader in Computer Science

Lead, Autonomous Driving and Intelligent Transport

Oxford Brookes University

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work, except for contributions explicitly specified in the text, Acknowledgements, and any mentioned collaborations. This thesis contains less than 40,000 words excluding bibliography, footnotes, tables, and equations and has fewer than 40 figures.

Muhammad Mubashar

January 2026

ACKNOWLEDGEMENTS

I would like to offer my most sincere gratitude to my supervisor, Fabio. His mentorship has been nothing short of transformative; his patience, wisdom, and unflagging support have served as the bedrock of this thesis. I am deeply appreciative of the intellectual autonomy he afforded me and his steadfast confidence in my abilities, both of which were pivotal in shaping my identity as a scholar. The profound impact he has had on my career cannot be overstated, and his guidance will remain a lasting influence on my life.

I am equally indebted to my second supervisor, Andy, for his invaluable counsel and constant encouragement. His constructive feedback contributed greatly to my professional growth. My heartfelt thanks also go to PGR tutor, Matthias. His direction and clarity were essential in helping me navigate the more arduous phases of my doctoral journey.

It has been a privilege to conduct this research alongside such brilliant colleagues. I extend a special note of appreciation to Shireen, Salman, Maryam, and wider Epistemic AI project team for their crucial collaboration and support on our shared publications. I am also deeply grateful to my mentor Naeemullah for his guidance and support.

Finally, and most importantly, I dedicate my deepest thanks to my family—my parents, Shazia and Fayyaz, as well as Ejaz, Nagina, Asma, Umair, Ali, Seemal and Qurat ul Ain. To my wife and partner, Afnan: you have been my rock. Your unconditional love, patience, and strength have been the anchor that sustained me through every challenge and triumph of this journey.

This thesis is based on work conducted as part of the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 964505 (E-pi), which has also funded my PhD and supported this research throughout.

TABLE OF CONTENTS

Acknowledgements iv
List of Tables xi
List of Figures xiv
List of Abbreviations xviii
Abstract xxi
Chapter I: Introduction 1
1.1 Motivation 2
1.2 Research Questions 3
1.3 Thesis Structure 3
1.4 Contributions 5
1.5 Dissemination 6
Chapter II: Background 8
2.1 Introduction 8
2.2 Uncertainty in Machine Learning 8
2.2.1 Aleatoric Uncertainty 8
2.2.2 Epistemic Uncertainty 9
2.3 Dempster-Shafer Theory and Random Sets 9
2.3.1 Random Sets 10
2.3.2 Mass Functions (Basic Probability Assignment) 10
2.3.3 Belief and Plausibility 11
2.3.4 Möbius Inversion 11
2.3.5 Example 12
2.4 From Beliefs to Decisions 13
2.4.1 Credal Sets 13
2.4.2 Pignistic Transformation 13
2.5 Existing Approaches to Uncertainty in Deep Learning 13

2.5.1 Bayesian Neural Networks (BNNs) 14
2.5.2 Deep Ensembles 14
2.5.3 Evidential Deep Learning (EDL) 14
2.6 Standard Generative Models 15
2.6.1 Generative Adversarial Networks (GANs) 15
2.6.2 Diffusion Models 15
2.6.3 Large Language Models 15
Chapter III: Related Work 17
3.1 Uncertainty Quantification in Classification 17
3.1.1 Probabilistic and Bayesian Approaches 17
3.1.2 Ensemble and Deterministic Methods 18
3.1.3 Evidential and Imprecise Probability Models 19
3.2 Generative Adversarial Networks (GANs) 19
3.2.1 Addressing Mode Collapse 19
3.3 Diffusion Models 21
3.3.1 Categories of Diffusion Models 21
3.3.2 Challenges: Inference Speed and Diversity 21
3.4 Uncertainty in Large Language Models (LLMs) 22
3.4.1 Current Approaches 22
Chapter IV: Budgeting of Focal Sets 23
4.1 Motivation and Background 23
4.2 The Problem of the Power Set 23
4.3 Budgeting Algorithm 24
4.4 Case Study: CIFAR-10 26
4.5 Ablation Study on Budget Size ( ) 27
4.6 Conclusion 28
Chapter V: Random-Set Neural Networks 30
5.1 Introduction 30

5.2 The Random-Set Neural Network 32
5.2.1 Architecture and Budgeting 32
5.2.2 Loss Function 33
5.2.3 Measuring Uncertainty 34
5.3 Experiments 34
5.3.1 Experimental Setup and Implementation 34
5.3.2 Standard Predictive Performance 36
5.4 Out-of-Distribution (OoD) Detection 37
5.4.1 Robustness to Distributional Shifts 39
5.4.2 Hyperparameter Sensitivity Analysis 40
5.4.3 Scalability to Large-Scale Architectures 41
5.5 Conclusion 41
Chapter VI: Evaluation Under Uncertainty 44
6.1 Introduction 44
6.2 Classes of Epistemic Predictions 44
6.2.1 Point-Estimate Models 45
6.2.2 Probabilistic and Ensemble Models 45
6.2.3 Set-Valued Models 46
6.3 Mapping Predictions to Credal Sets 46
6.3.1 Construction Procedure 46
6.4 Evaluation of Epistemic Predictions 47
6.4.1 The Unified Metric 47
6.4.2 Rationale and Scenarios 48
6.5 Experiments 49
6.5.1 Experimental Setup 49
6.5.2 Analysis of the Evaluation Metric 49
6.5.3 Evaluation of the Trade-off Parameter 51
6.5.4 Model Selection 51

6.6 Conclusion 52
Chapter VII: Random-Set Large Language Models 54
7.1 Introduction 54
7.2 Introduction and Motivation 54
7.3 Methodology 56
7.3.1 Budgeting: Constructing the Frame of Discernment 56
7.3.2 Architecture and Training 57
7.3.3 Uncertainty Estimation 59
7.4 Experiments 59
7.4.1 Experimental Setup 59
7.4.2 Generative Performance 61
7.4.3 Comparison with Uncertainty Baselines 61
7.4.4 Uncertainty quantification 62
7.4.5 Hallucination Detection 63
7.4.6 Align Score 65
7.4.7 Ablation on Hyperparameters α & βK. 66
7.4.8 Ablation on Number of Focal Sets 67
7.5 Analysis of Budgeted Focal Sets 68
7.5.1 Qualitative Assessment of Focal Sets 68
7.5.2 Quantitative Semantic Analysis: Centroid Distance 69
7.5.3 Distribution of Focal Set Sizes 69
7.5.4 Morphology and synonymity analysis 70
7.5.5 Impact of Distance Metrics on Performance 71
7.6 Conclusion 72
Chapter VIII: Epistemic Diffusion Models 74
8.1 Introduction 74
8.2 Background: Denoising Diffusion Probabilistic Models 74
8.2.1 The Forward Diffusion Process 75

8.2.2 The Reverse Denoising Process 76
8.3 Problem Statement and Motivation 77
8.4 Epistemic Diffusion Models 78
8.4.1 Methodology: The Normal-Inverse-Gamma Prior 79
8.4.2 Loss Function 80
8.5 Bayesian Formulation and Challenges 81
8.6 Evaluation metrics 81
8.7 Experimental Results 82
8.7.1 Fixed Variance Baseline 82
8.7.2 Full Bayesian Learning 83
8.7.3 Manifold-Aware Diffusion 84
8.8 Conclusion 85
Chapter IX: Epistemic Generative Adversarial Networks 87
9.1 Introduction 87
9.1.1 The Problem of Mode Collapse 87
9.1.2 Our Contribution: Epistemic GANs 88
9.2 Theoretical Framework 89
9.2.1 Dempster-Shafer Theory in GANs 89
9.3 Architecture 89
9.3.1 Discriminator with Belief Outputs 89
9.3.2 Generator with Pixel-Wise Mass Prediction 90
9.4 Loss Function 91
9.4.1 Discriminator Loss 92
9.4.2 Generator Loss 92
9.5 Experiments 93
9.5.1 Experimental Setup 93
9.5.2 Results and Analysis 93
9.5.3 Ablation Studies 94

9.6 Application: Synthetic Data for Autonomous Driving 95
9.7 Conclusion 96
Chapter X: Application in Autonomous Driving 99
10.1 Introduction 99
10.2 The ROAD-INTENT Dataset 100
10.2.1 Data Collection at the RACE Facility 100
10.2.2 Scenario Design and Complexity 101
10.2.3 Annotation Protocol: Intent vs. Action 102
10.3 Random-Set Vision Language Models (RS-VLMs) 103
10.3.1 Methodology and Architecture 103
10.3.2 Experiments 104
10.4 Annotating ROAD-INTENT with RS-VLM 105
10.4.1 Qualitative Results 105
10.4.2 Impact on Inverse Reinforcement Learning 106
10.5 Conclusion 107
Chapter XI: Conclusion and Future Work 109
11.1 Summary of Contributions 109
11.2 Limitations 110
11.3 Future Work 111
11.3.1 Dynamic and End-to-End Budgeting 111
11.3.2 Manifold-Aware Diffusion 111
11.3.3 Towards Random-Set Concept Models 111
11.4 Final Words 112

LIST OF TABLES

Number *Page*

5.1 Test accuracies (%) and inference time (ms) for uncertainty estimation over 5 consecutive runs across methods and datasets. Average and standard deviation are shown for each experiment. 36

5.2 OoD detection performance and uncertainty estimation for models trained on ResNet50 on CIFAR-10 vs SVHN/Intel Image, MNIST vs F-MNIST/K-MNIST and ImageNet vs ImageNet-O. Evaluation metrics include AUROC/AUPRC (OoD); Entropy of predictions (uncertainty) and Expected Calibration Error (ECE). 37

5.3 Credal set width for RS-NN on iD vs OoD datasets: CIFAR10 vs SVHN/Intel Image, MNIST vs F-MNIST/K-MNIST and ImageNet vs ImageNet-O. 38

5.4 Test accuracies (%) for RS-NN and standard CNN on Rotated MNIST out-of-distribution (OOD) samples. Rotation angle is randomized within the specified intervals. 39

5.5 Adaptability to large-scale model architectures with test accuracy (%) and parameters (in million) reported on CIFAR10. 41

5.6 Comparison of predictions on Out-of-Distribution (OoD) noisy and rotated MNIST samples. Standard CNNs frequently predict incorrect classes with high confidence, whereas the RS-NN distributes belief mass to correctly encompass the true label or exhibit lower pignistic confidence. 43

6.1 Comparison of Kullback-Leibler divergence (KL), Non-Specificity (NS) and Evaluation Metric ($\mathcal{E}$) for uncertainty-aware classifiers (trade-off λK= 1). Mean and standard deviation are shown for CIFAR-10, MNIST and CIFAR-100 datasets. 50

6.2 Model Rankings Based on KL and NS on the CIFAR-10 dataset for different values of trade-off λ. Model selection is based on the mean of Evaluation Metric ($\mathcal{E}$) with models with the lowest $\mathcal{E}$ ranking first. 53

7.1 Performance comparison on CoQA (Cosine Similarity) and OBQA (Accuracy). RS-LLMs consistently outperform standard baselines, indicating that the belief-function representation captures richer semantic information. 62

7.2 Accuracy comparison on ARC-E, BoolQ, and OBQA. RS-Llama2 surpasses both Bayesian and Ensemble approaches. 62

7.3 Hallucination detection performance (AUROC) on BoolQ using Entropy. 64

7.4 Hallucination detection performance (AUROC) on OBQA using Entropy. 64

7.5 Uncertainty evaluation using entropy for Llama2 and RS-Llama2 for CoQA on correct and incorrect context. 64

7.6 Uncertainty evaluation using credal set width for RS-Llama2 on correct and incorrect context. 65

7.7 AlignScore on CoQA Dataset for Llama2 and RS-Llama2 65

7.8 Cosine Similarity values for different $\alpha = \beta$ settings 66

7.9 Cosine Similarity values across different budget sizes 67

7.10 Cosine Similarity values for different distance metrics in hierarchical clustering 71

7.11 Uncertainty evaluation using pignistic entropy of budgeting with a stand-alone metric and with combined metrics. 72

8.1 Performance of Standard GAN and different modes of epistemic diffusion model on Celeb-A and Cifar-10 dataset. Reference represents the Vendi score (diversity) of the training data. 84

9.1 Performance of Standard GAN and Epistemic GAN on CelebA, CIFAR-10, and Food-101. Reference represents the Vendi score (diversity) of the training data. Epistemic GAN achieves lower FID and higher Vendi Scores. 93

9.2 Ablation for γK and β on CelebA. [FID (Vendi Score)]. The best balance is achieved when both terms are active (≈ 1). 94

9.3 Ablation study for λK on the CelebA dataset. Results are reported in **FID (Vendi Score)** format. 95

9.4 Ablation for architecture components on CelebA. 95

9.5 Training time per epoch on CelebA (5 runs). 95

10.1 Performance (Accuracy) of Standard VLM and RS-VLM on MSVD-QA and TGIF-QA. 105

10.2 Hallucination Detection: Uncertainty evaluation (Entropy) on Correct (CC) vs Incorrect (IC) video contexts. 105

LIST OF FIGURES

Number *Page*

2.1 A belief function measures the total belief (sum of masses of its subsets) for a set. 12

4.1 2D visualization of the clusters of 10 classes of datapoints belonging to the CIFAR10 dataset, together with the ellipses formed by RS-CNN using Gaussian Mixture Models. 25

4.2 Ablation study on number of non-singleton focal sets on Cifar-10 dataset. The maximum value of K can be 1013 for 10 classes (after excluding the singletons and empty set). 28

5.1 Inference in a Bayesian Neural Network (top) compared to a Random-Set Neural Network (bottom). While BNNs sample from a distribution of weights to get predictions, RS-NN maps the input directly to a belief function over a budget of focal sets. 31

5.2 Confidence scores of RS-NN and CNN for FGSM adversarial attack on MNIST. The RS-NN correctly drops its confidence when attacked, while the CNN stays overconfident. 31

5.3 RS-NN model architecture. (a) *Budgeting*: We select the top relevant sets of classes using clustering. (b) *Training*: The network predicts belief functions. The grey layers are the ones we train, while the blue layers can be any pre-trained backbone. 33

5.4 Width of credal predictions for CIFAR-10 test data (correctly classified, blue; incorrectly classified, red). 38

5.5 Ablation study of hyperparameters α and βKon CIFAR-10 test accuracy. The model exhibits optimal performance when α and β are minimized, suggesting that while regularization is necessary for validity, excessive constraints can hamper optimization. 40

6.1 Different types of uncertainty-aware model predictions, shown in a unit simplex of probability distributions defined on the list of classes $\mathbf{Y} = \{a, b, c\}$. Our proposed evaluation framework uses a metric which combines, for each input $\mathbf{x}$, a distance (arrows) between the corresponding ground truth (e.g., $(0, 1, 0)$) and the *epistemic predictions* generated by the various models (in the form of credal sets), and a measure of the extent of the credal prediction (*non-specificity*). 48

6.2 Measures of KL divergence (top left), Non-specificity (top right), Evaluation Metric (bottom left) for both Correctly (CC) and Incorrectly Classified (ICC) samples from CIFAR-10. 51

6.3 Evaluation metric vs trade-off parameter 52

7.1 The RS-LLM Budgeting Pipeline. Token embeddings are extracted from the base model and processed via hierarchical clustering to form semantically meaningful focal sets (e.g., synonyms, morphological variants). 57

7.2 Training and Generation flow of RS-LLM. During training (teacher forcing), the model predicts belief functions for the next token. During generation, the predicted belief function is converted to a pignistic probability distribution for sampling. 57

7.3 Training examples from CoQA and OBQA datasets. The text in black highlights the actual question, while the blue text represents prompt instructions. The model is trained to predict the text in green. 61

7.4 Behavior of entropy measures for all models with respect to the correctness on OBQA dataset. 63

7.5 (a) Frequency distribution of centroid distances of the obtained budgeted focal sets. (b) Frequency distribution of sizes of obtained budgeted focal sets. Note: There are 8 focal sets with sizes > 30. They are excluded here for better visualisation. 70

8.1 Diversity analysis on images generated by the Stable Diffusion Model [122]. Top: "an oil painting of a horse sitting in an airplane" yields samples with nearly identical composition. Middle: "a picture of a girl eating dinner" produces subjects with similar demographics. Bottom: replacing "girl" with "female" results in identical poses, highlighting a lack of generative diversity. 78

8.2 Left: Standard diffusion generation using a fixed Gaussian. Right: Epistemic diffusion generation. At each step, we sample a Gaussian from a learned second-order distribution (depicted by the shaded region representing the variance of the cloud of Gaussians), enhancing diversity. 79

8.3 Left: Standard diffusion models capture data variability using a single Gaussian. Right: Epistemic diffusion models model this variability as a second-order distribution over possible Gaussian PDFs, allowing for more flexible data modelling. 80

8.4 Images generated by an Epistemic Diffusion Model trained on CIFAR-10. In this experiment, the variance was fixed, resulting in high-quality but standard diversity outputs. 83

8.5 Images generated by an Epistemic Diffusion Model with image manifold trained on CIFAR-10. 85

9.1 Discriminator architecture comparison. The Epistemic Discriminator outputs two belief values, allowing for a residual "ignorance" mass. 90

9.2 Generator Architecture and flow for Epistemic GANs. The generator predicts mass functions (Dirichlet parameters) for regions, samples intervals, and then constructs the image. 91

9.3 Representation of a belief function with Borel intervals used in the generator's intermediate representation. 91

9.4 Generations for Standard GAN (left) and Epistemic GAN (right) on Celeb-A. 94
9.5 Epistemic GAN generations for road scenarios on Citiscape dataset. . 97
10.1 RADAR-realistic roadside infrastructure at the RACE facility, allowing for reconfigurable urban layouts. 100
10.2 Illustration of the Giving Way scenario across three levels of complexity. (a) Expected behaviour. (b) Ambiguous behaviour. (c) Unexpected behaviour causing a near-miss. 102
10.3 Architecture flow of RS-VLM. Visual tokens from the video encoder are concatenated with text prompts. The model predicts a belief function over a budget of token sets, which is converted to a mass function for uncertainty estimation. 104
10.4 Sample frames from the ROAD-INTENT dataset used for automated annotation. 107

LIST OF ABBREVIATIONS

AE Aleatoric Uncertainty.

AGI Artificial General Intelligence.

AI Artificial Intelligence.

AUC Area Under the Curve.

AUPRC Area Under the Precision–Recall Curve.

AUROC Area Under the Receiver Operating Characteristic Curve.

BNN Bayesian Neural Networks.

BPA Basic Probability Assignment.

CNN Convolutional Neural Networks.

CP Conformal Prediction.

DDU Deep Deterministic Uncertainty.

DIR Dirichlet Distribution.

DL Deep Learning.

DP Dirichlet Process.

DST Dempster–Shafer Theory.

ECE Expected Calibration Error.

EDL Evidential Deep Learning.

EU Epistemic Uncertainty.

FPR False Positive Rate.

GAN Generative Adversarial Network.

GMM Gaussian Mixture Model.

KL Kullback–Leibler Divergence.

LAPLACE Laplace Approximation for Bayesian Inference.

LLM Large Language Model.

LSTM Long Short-Term Memory.

MAP Maximum A Posteriori.

MC Monte Carlo Sampling.

MCD Monte Carlo Dropout.

MCMC Markov Chain Monte Carlo.

ML Machine Learning.

MLE Maximum Likelihood Estimation.

MSE Mean Squared Error.

NS Non-Specificity.

OoD Out-of-Distribution Detection.

ROC Receiver Operating Characteristic.

RS Random Sets.

RS-LLM Random-Set Large Language Models.

RS-NN Random-Set Neural Networks.

RS-VLM Random-Set Vision Language Models.

TPR True Positive Rate.

UQ Uncertainty Quantification.

VAE Variational Autoencoder.

VI Variational Inference.

Generative Modelling under Epistemic Uncertainty

Abstract

by

MUHAMMAD MUBASHAR

The deployment of Deep Learning in safety-critical domains is hindered by pathological overconfidence and an inability to distinguish between aleatoric uncertainty (data ambiguity) and epistemic uncertainty (lack of knowledge). This thesis addresses these limitations by establishing a rigorous framework for Random-Set Deep Learning, shifting from point-estimate probabilities to belief functions over the power set of outcomes. Enabled by a novel Budgeting strategy that ensures scalability, this framework allows models to explicitly represent ignorance.

While we validate these principles through Random-Set Neural Networks (RS-NN) for classification and a Unified Evaluation Framework, the primary contribution of this work lies in re-imagining Generative AI under epistemic uncertainty. We introduce Random-Set Large Language Models (RS-LLMs), which predict belief functions over token sets to quantify second-order uncertainty, thereby providing a robust mechanism for hallucination detection. Furthermore, we propose Epistemic Generative Adversarial Networks (GANs) and Epistemic Diffusion Models. By modeling the uncertainty of the generation process itself via second-order distributions, these architectures significantly mitigate mode collapse and enhance sample diversity compared to standard baselines.

Finally, we demonstrate the practical utility of these generative capabilities in Autonomous Driving. We introduce the ROAD-INTENT dataset and utilize Random-Set Vision Language Models (RS-VLMs) to automate the annotation of actor intent with calibrated uncertainty scores, paving the way for safer, self-aware AI systems.

Chapter 1

INTRODUCTION

Artificial Intelligence (AI) has seen massive growth in the last decade. Deep learning models can now drive cars, diagnose diseases, and write code. However, despite these successes, there is a big problem: most of these models are ‘‘overconfident’’. They tend to be sure of their predictions even when they are wrong. For example, a standard neural network might classify a blurry image of a dog as a cat with 99% confidence, or a Large Language Model (LLM) might hallucinate a fact while sounding perfectly plausible.

This lack of reliable uncertainty quantification is a major barrier to deploying AI in safety-critical systems. If an autonomous vehicle is unsure about a pedestrian’s location, it should slow down, not drive ahead with false confidence. This thesis argues that the solution lies in modelling *epistemic uncertainty*—the uncertainty that comes from a lack of knowledge or data—using the mathematical framework of Random Sets and Belief Functions.

In this work, I explore how to integrate epistemic uncertainty into both discriminative and generative models. I start by building a foundation in unsupervised learning and clustering, proposing a method to handle the computational complexity of random sets. I then apply this to create Random-Set Neural Networks (RS-NNs) for classification. To measure success, I develop a unified evaluation framework. Finally, I expand these ideas into the frontier of Generative AI—creating Random-Set LLMs, Epistemic GANs, and Epistemic Diffusion models—and demonstrate their real-world utility in autonomous driving.

1.1 Motivation

Current deep learning methods, primarily based on Bayesian probability or standard softmax outputs, have limitations. Bayesian methods, while theoretically sound, are often computationally expensive (e.g., requiring sampling) and require defining prior distributions which can introduce bias. On the other hand, ensemble methods are costly to train.

The core motivation of this thesis is to move beyond single-point probabilities. When a model faces an input it has never seen before (Out-of-Distribution or OoD data), or when the training data was scarce, the model should be able to say ”I don’t know” or output a set of possibilities rather than a single guess.

This is where *Random Sets* (or Belief Functions) come in. Unlike a probability distribution that must sum to one across mutually exclusive outcomes, a belief function assigns mass to *sets* of outcomes. This allows us to distinguish between ‘‘conflicting evidence’’ (equal evidence for cat and dog) and ‘‘lack of evidence’’ (I haven’t seen this animal before).

However, applying random sets to deep learning is hard. The size of the power set (all possible subsets of classes) grows exponentially. If we have 100 classes, we theoretically have 2^{100} possible subsets, which is impossible to compute. A significant part of my motivation was to solve this *scalability* problem so that epistemic uncertainty can be used in large-scale models like LLMs and Vision Transformers.

Furthermore, in Generative AI, we face issues like *mode collapse* in GANs (where the model only generates a few types of images) and *hallucinations* in LLMs. I am motivated to show that by modeling the uncertainty in the generation process itself—using epistemic noise or random-set outputs—we can generate more diverse images and more truthful text.

1.2 Research Questions

To address the challenges outlined above, this thesis tackles the following key research questions:

1. **How can we make epistemic uncertainty modelling computationally feasible for deep learning?** specifically, how can we efficiently select a ”budget” of focal sets to avoid the exponential explosion of the power set?

2. **Can Random-Set Neural Networks (RS-NNs) outperform state-of-the-art Bayesian and Ensemble methods?** We investigate if RS-NNs can provide better accuracy and uncertainty quantification, particularly for Out-of-Distribution (OoD) detection and robustness against adversarial attacks.

3. **How should we evaluate models that output set-valued or imprecise predictions?** Since standard metrics like accuracy or negative log-likelihood don't fully capture the quality of a belief function, we need a new, unified evaluation framework.

4. **How can epistemic uncertainty improve Generative AI?** Can we reduce hallucinations in Large Language Models (LLMs) and mitigate mode collapse in Generative Adversarial Networks (GANs) by incorporating evidential theory?

5. **How can these methods be applied to real-world problems like Autonomous Driving?** Specifically, can we use these models to better understand agent intent and automate the labeling of complex road scenarios?

1.3 Thesis Structure

The thesis is organised into eleven chapters, telling the story from foundational clustering methods to advanced generative applications.

- **Chapter 1: Introduction**. This chapter, outlining the motivation and structure.
- **Chapter 2: Background**. Reviews the mathematical foundations of Dempster-Shafer Theory, Random Sets, existing uncertainty estimation methods (Bayesian NN, Ensembles) and generative models.
- **Chapter 3: Related Work**. Surveys the state-of-the-art in uncertainty quantification, generative modelling, and their applications.
- **Chapter 4: Budgeting of Focal Sets**. This chapter introduces a novel *budgeting* strategy. This strategy uses traditional clustering (GMMs) to intelligently select a limited number of focal sets, solving the exponential complexity issue.
- **Chapter 5: Random-Set Neural Networks**. This chapter introduces the RS-NN. It is a wrapper approach that predicts belief functions over the budget of sets defined in Chapter 4. We show it scales to large architectures (ViT, EfficientNet) and beats baselines on CIFAR-10 and ImageNet.
- **Chapter 6: Evaluation Under Uncertainty**. This proposes a unified evaluation framework. It maps predictions from different models (Bayesian, Evidential, Deterministic) into a common ‘‘credal set’’ representation. It define a new metric that trades off accuracy (distance to ground truth) against imprecision (size of the set).
- **Chapter 7: Random-Set Large Language Models**. Moving to Generative AI, this chapter describes RS-LLMs. By predicting belief functions over sets of tokens, we aim to reduce hallucinations and provide a measure of uncertainty for generated text.
- **Chapter 8: Epistemic Diffusion Models**. This chapter proposes a ‘‘two-level’’ diffusion process that models second-order distributions over Gaussian noise, aiming to improve the diversity and quality of generated images.

- **Chapter 9: Epistemic Generative Adversarial Networks**. This chapter presents Epistemic GANs. We modify the discriminator to output belief functions (Real, Fake, or Uncertain) and the generator to predict mass functions. This forces the generator to explore diverse modes to fool the epistemic discriminator.
- **Chapter 10: Application in Autonomous Driving**. Finally, we apply these techniques to the real world. We describe the creation of the ROAD-Intent dataset and the development of Random-Set Vision Language Models (RS-VLMs) for automated, uncertainty-aware labeling of road scenes.
- **Chapter 11: Conclusion and Future Work.** Summary of findings and future directions.

1.4 Contributions

The main contributions of this thesis are as follows:

1. **A Scalable Budgeting Method for Random Sets**: We developed a method using GMM clustering and t-SNE to select the most relevant subsets of classes. This reduces the complexity from $2^{NK}$to a manageable number , making random-set methods feasible for large-scale problems.
2. **Random-Set Neural Networks (RS-NN)**: We proposed a novel architecture that predicts belief functions. It achieves superior performance in uncertainty estimation and OoD detection compared to Deep Ensembles and Bayesian methods, while being robust to adversarial attacks.
3. **Unified Evaluation Metric**: We introduced a new metric that combines distance-to-ground-truth (accuracy) with non-specificity (precision). This allows for a fair comparison of diverse uncertainty-aware models.
4. **Generative Models with Epistemic Uncertainty**:

- **RS-LLMs**: Extending random sets to language modelling to improve truthfulness.
- **Epistemic Diffusion Models**: A novel diffusion model formulation leveraging second-level distributions to improve diversity.
- **Epistemic GANs**: A new GAN formulation where the discriminator predicts belief values, significantly improving sample diversity (Vendi score) and reducing mode collapse.

5. **ROAD-Intent Dataset & RS-VLM**: We put forward a unique autonomous driving dataset with explicit intent labels. We also developed Random-Set Vision Language Models to automatically annotate this data with uncertainty scores, aiding in the training of robust Inverse Reinforcement Learning agents.

1.5 Dissemination

The research presented in this thesis has led to several publications and submissions:

- Manchingal, S., **Mubashar, M.**, Wang, K., Shariatmadar, K., & Cuzzolin, F. (2025). *Random-Set Neural Networks*. Accepted at ICLR 2025.
- Manchingal, S., **Mubashar, M.**, Wang, K., & Cuzzolin, F. (2025). *A Unified Evaluation Framework for Epistemic Predictions*. Accepted at AISTATS 2025.
- **Mubashar, M.**, Manchingal, S., & Cuzzolin, F.*Random-Set Large Language Models*. Under review at ICML 2026.
- **Mubashar, M.**, & Cuzzolin, F. *Epistemic Generative Adversarial Networks*. Under review at ICML 2026.

This body of work represents a step towards AI systems that are not just smart, but also self-aware of their own limits, paving the way for safer and more trustworthy technology.

Chapter 2

BACKGROUND

2.1 Introduction

Before introducing novel architectures for uncertainty quantification, it is essential to establish a rigorous understanding of what ‘‘uncertainty’’ means in the context of machine learning. Standard Deep Learning models typically rely on frequentist or Bayesian probability theory to model outputs. However, as argued throughout this thesis, classical probability distributions (specifically the categorical distribution produced by the Softmax operator) are often insufficient for distinguishing between different *sources* of uncertainty.

This chapter provides the necessary theoretical scaffolding. We begin by defining the types of uncertainty (Aleatoric vs. Epistemic). We then introduce the primary mathematical framework of this thesis: Dempster-Shafer Theory (DST), also known as the Theory of Belief Functions, and its connection to Random Sets. Finally, we review the standard generative models (GANs, Diffusion Models and LLMs) to set the baseline for the modifications proposed in later chapters.

2.2 Uncertainty in Machine Learning

In the context of predictive modelling, uncertainty is conventionally categorised into two distinct types,.

2.2.1 Aleatoric Uncertainty

Aleatoric uncertainty refers to the inherent randomness or noise in the data generation process. It is a property of the data itself and cannot be reduced by collecting more training data.

- **Example**: In an image of a coin flip that is blurry due to motion, the outcome is inherently ambiguous. No amount of training will allow the model to know whether it is Heads or Tails with certainty.
- **Modelling**: This is typically modelled by the probability distribution $P(y|x)$ output by a standard network.

2.2.2 Epistemic Uncertainty

Epistemic uncertainty refers to uncertainty caused by a lack of knowledge on the part of the model. It arises from limited data, model misspecification, or domain shift. Crucially, epistemic uncertainty *can* be reduced by observing more data.

- **Example**: A model trained on images of dogs and cats is presented with an image of a helicopter. The model is ignorant of this class. Ideally, it should output a state of ”I don’t know,” rather than forcing a probability assignment to ”Dog” or ”Cat”.
- **The Problem with Softmax**: A standard neural network is forced to normalize its output to sum to 1. In the helicopter example, it might output $P(\text{Dog}) = 0.5, P(\text{Cat}) = 0.5$. This distribution is mathematically indistinguishable from the aleatoric ambiguity of a blurry dog/cat image. This conflation is the root cause of the ”overconfidence” problem [40].

To resolve this, we require a calculus that can represent ”ignorance” explicitly. This brings us to the Theory of Belief Functions.

2.3 Dempster-Shafer Theory and Random Sets

The Dempster-Shafer Theory (DST) [42, 129] is a generalisation of Bayesian probability theory. While probability theory assigns mass to individual outcomes, DST assigns mass to *sets* of outcomes. This flexibility allows for the explicit modelling of ignorance.

2.3.1 Random Sets

A Random Set is a random variable taking values in the power set of a space, rather than the space itself [110, 105]. Consider a meteorological station recording a 2D environmental vector, $\mathbf{e} = (T, H)$, where TKdenotes temperature and HKdenotes humidity, each modelled by a data probability distribution. Under normal operation the station reports exact values for both variables. Suppose, however, that the humidity sensor malfunctions and fails to return a numerical reading.

Rather than imputing an arbitrary humidity value, we represent the unknown component HKby the set $\mathcal{H}$ of all plausible humidity levels (e.g., $0\% \leq h \leq 100\%$). The underlying random process generating environmental conditions remains intact, but the observation becomes a set-valued random variable:

$$\mathbf{E} = (T, \mathcal{H}), K \tag{2.1}$$

where $\mathcal{H}$ encapsulates the missing information. This model, in which random outcomes may be sets rather than points, defines a *random set*. Random sets extend classical probability by encoding uncertainty through collections of values instead of precise realizations, making them ideal for modeling imprecise or incomplete measurements [98, 110, 105].

2.3.2 Mass Functions (Basic Probability Assignment)

In the finite discrete case, a random set is characterized by a Mass Function (or Basic Probability Assignment - BPA) $m : 2^{\Theta} \rightarrow [0, 1]$. It must satisfy the following axioms:

$$m(\emptyset) = 0, \quad \sum_{A \subseteq \Theta} m(A) = 1 \tag{2.2}$$

The mass $m(A)$ represents the portion of evidence that supports the proposition that the truth lies exactly in set A, and not in any specific subset of A.

- **Bayesian Case**: If all focal sets A are singletons (cardinality $|A| = 1$), the mass function collapses to a standard Probability Mass Function (PMF).

- **Total Ignorance**: If $m(\Theta) = 1$, we represent total ignorance. We know the truth is *somewhere* in the domain, but we have zero evidence favoring any specific element.

2.3.3 Belief and Plausibility

From the mass function, we can derive two non-additive measures for any set A: Belief (Bel) and Plausibility (Pl).

Belief represents the total evidence that fully supports A. It is the sum of masses of all subsets of A:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2.3}$$

Plausibility represents the total evidence that does not contradict A (i.e., is consistent with A). It is the sum of masses of all sets that intersect with A:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(A^c) \tag{2.4}$$

The interval $[Bel(A), Pl(A)]$ represents the range of uncertainty for hypothesis A. In a standard probability distribution, $Bel(A) = Pl(A) = P(A)$, collapsing the uncertainty interval to zero. In epistemic AI, the width of this interval ($Pl - Bel$) is a key metric for ignorance [146].

2.3.4 Möbius Inversion

Crucially for our neural network implementations, the mass function can be recovered from the Belief function via the Möbius Inversion formula [129]:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bel(B) \tag{2.5}$$

This relationship allows us to train networks to predict Belief values (which are monotonic and easier to constrain) and mathematically derive the underlying masses for uncertainty estimation.

2.3.5 *Example*

Consider a classification problem (Figure 2.1) where an object belongs to one of three possible categories, $\Theta = \{c_1, c_2, c_3\}$. A belief function might express uncertainty by assigning mass as follows:

$$m(\{c_1\}) = 0.4, m(\{c_3\}) = 0.2, m(\{c_1, c_2\}) = 0.4.K \tag{2.6}$$

Here, 40% of the belief supports c_1, 20% supports c_3, and 40% supports the composite hypothesis that the object belongs to either c_1 or c_2 but not c_3, without being able to specify which. The belief value $Bel(A)$ of a set of classes AKaccumulates mass from all the subsets of A:

$$Bel(\{c_1, c_2\}) = m(\{c_1\}) + m(\{c_1, c_2\}) = 0.4 + 0.4 = 0.8.K \tag{2.7}$$

This means we have an 80% belief that the object belongs to either c_1 or c_2, reflecting epistemic uncertainty rather than simply assigning a probability to each class. Belief functions generalise Bayesian probability by allowing explicit representation of uncertainty, making them useful in applications where knowledge is incomplete or ambiguous [130, 17].

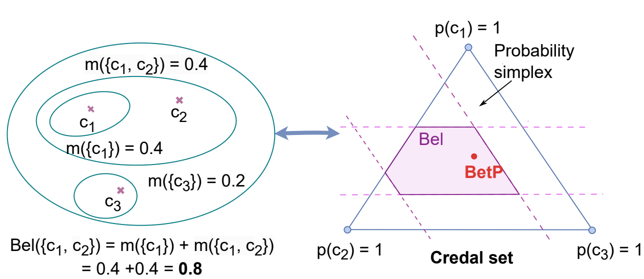


Figure 2.1: A belief function measures the total belief (sum of masses of its subsets) for a set.

2.4 From Beliefs to Decisions

While random sets capture rich uncertainty information, practical applications (like classifying an image) often require a specific decision or a point-estimate probability.

2.4.1 Credal Sets

A belief function Bel defines a Credal Set $\mathcal{P}_{Bel}$, which is the convex set of all probability distributions P consistent with the belief evidence [84, 32]:

$$\mathcal{P}_{BelK} = \{P \in \Delta^{-1} \mid P(A) \geq Bel(A), \forall A \subseteq \Theta\} \tag{2.8}$$

Geometrically, this is a polytope inside the probability simplex. A large polytope indicates high epistemic uncertainty; a single point indicates certainty.

2.4.2 Pignistic Transformation

To make a decision, we often project the random set onto a single probability distribution. The most common method is the Pignistic Transformation [131], denoted as $BetP$. It distributes the mass of each focal set equally among its elements:

$$BetP(\omega) = \sum_{A \ni \omega K} \frac{m(A)}{|A|} \quad \forall \omega \in \Theta \tag{2.9}$$

Smets [131] originally proposed to use the pignistic probability for decision making using belief functions, by applying expected utility to it. Notably, the pignistic probability is geometrically the centre of mass of the credal set associated with a belief function [32, 37, 38]. This $BetPK$is used in our RS-NN and RS-LLM architectures to calculate the final classification accuracy (or next-token probability), while the underlying mass structure is retained for uncertainty quantification.

2.5 Existing Approaches to Uncertainty in Deep Learning

There are already several ways to estimate uncertainty in neural networks. However, they have some drawbacks which motivates the new methods we propose later.

2.5.1 Bayesian Neural Networks (BNNs)

BNNs are arguably the most theoretically grounded approach [15]. Instead of having fixed weights (parameters), the weights in a BNN are probability distributions. During inference, we marginalise over the weights to get the prediction.

$$P(y|x,D) = \int P(y|x,w)P(w|D)dwK$$

The problem is that this integral is usually intractable (impossible to calculate exactly). We have to use approximations like Variational Inference (VI), which can be computationally expensive and hard to tune. Also, choosing the ‘‘prior’’ distribution $P(w)$ is difficult and can bias the results.

2.5.2 Deep Ensembles

This method is very popular because it is simple [82]. You train the same neural network MKtimes (e.g., 5 or 10 times) with different random initializations. To predict, you average their outputs.

If all models agree, uncertainty is low. If they disagree, uncertainty is high. While effective, training 10 deep networks is huge computational burden. It’s not suitable for real-time applications like autonomous driving where we have limited resources.

2.5.3 Evidential Deep Learning (EDL)

EDL is closer to what we do [126]. It places a Dirichlet distribution over the class probabilities. It predicts the parameters $\alpha K$of the Dirichlet distribution. However, standard EDL methods often struggle to distinguish between conflicting evidence and lack of evidence effectively in high-dimensional spaces, and they don’t fully utilize the power of Random Sets math, specifically the use of arbitrary focal sets beyond just singletons and the whole set .

2.6 Standard Generative Models

In later chapters, we apply Epistemic principles to generative models. Here, we briefly review the standard formulations.

2.6.1 Generative Adversarial Networks (GANs)

A GAN consists of a Generator G and a Discriminator D playing a minimax game [56]:

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))] \tag{2.10}$$

Standard GANs are prone to Mode Collapse, where G produces limited varieties of samples. In Chapter 9, we will show how replacing the scalar discriminator output with a belief function mitigates this.

2.6.2 Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [64] learn to reverse a gradual noising process. The reverse process is modeled as a Markov chain with learned Gaussian transitions:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2.11}$$

Typically, Σ_θ is fixed. In Chapter 8, we will explore Epistemic Diffusion, where we learn a distribution over these Gaussian parameters to capture the uncertainty of the generation process itself.

2.6.3 Large Language Models

Large Language Models (LLMs) [1, 3, 139] represent a paradigm shift in Natural Language Processing. At their core, these models are probabilistic systems trained to model the conditional probability of the next token x_t given a sequence of preceding tokens $x_{<t}$:

$$P(x_t|x_{<t}) = \text{Softmax}(f_\theta(x_{<t})) \tag{2.12}$$

where f_θ is a neural network (typically a Transformer) parameterized by θ. LLMs are prone to hallucination. In Chapter 7, we will explore RS-LLMs, where we a predict a beleif function over token space gaining the advantage of uncertainty quantification hallucination detection.

Chapter 3

RELATED WORK

This chapter reviews the landscape of research relevant to this thesis. We begin by examining the broad field of uncertainty quantification in deep learning, focusing on classification. We then transition to generative modelling, detailing the evolution of Generative Adversarial Networks (GANs) and Diffusion Models, with a specific focus on the challenges of diversity and mode collapse. Finally, we survey the nascent field of uncertainty in Large Language Models (LLMs), identifying the gaps that our Random-Set approach aims to fill.

3.1 Uncertainty Quantification in Classification

The machine learning community has long recognised the challenge of estimating uncertainty in model predictions. Awareness of uncertainty enables models to offer more reliable and interpretable predictions, fostering trust and transparency. This is crucial in safety-critical domains such as medical diagnosis [143], autonomous vehicles [66], and finance [145], where inaccurate predictions may lead to adverse consequences.

3.1.1 Probabilistic and Bayesian Approaches

Bayesian approaches, pioneered by [19] and others [93, 107], are dominant in uncertainty estimation. In Bayesian Neural Networks (BNNs) [74, 76], uncertainty is explicitly represented through posterior predictive distributions over the parameter space.

However, exact Bayesian inference is often intractable. This has led to numerous approximations:

- Dropout Variational Inference: [51] proposed using dropout at test time (Monte Carlo Dropout) as a Bayesian approximation.
- Variational Inference (VI): Methods like Bayes by Backprop [15] approximate the posterior using a simpler distribution.
- Laplace Approximation: Recent works like the Laplace Bridge Bayesian Neural Network (LB-BNN) [65] use the Laplace Bridge to map between Gaussian and Dirichlet distributions, addressing computational costs [41].
- Function-Space Inference: Approaches like Function-Space Variational Inference (FSVI) [123] perform inference directly on the function outputs rather than weights.

Despite their advantages, Bayesian models face challenges when the model prior is misspecified [140] and often require computationally expensive sampling at inference time.

3.1.2 Ensemble and Deterministic Methods

Ensemble-based approaches, such as Deep Ensembles (DE) [82], efficiently estimate uncertainty by training multiple independent models and aggregating their predictions. While robust, the computational cost of training and maintaining multiple large models is often impractical. Epistemic Neural Networks (ENN) [113] attempt to mitigate this by using a base network with multiple heads (epinet), but still incur significant overhead.

Deep Deterministic Uncertainty (DDU) [106] and other deterministic methods attempt to estimate uncertainty via a single forward pass, often by analyzing the feature density, but they struggle to capture the full richness of epistemic uncertainty compared to ensemble or Bayesian methods.

3.1.3 Evidential and Imprecise Probability Models

Evidential Deep Learning (EDL) [127, 52] places a Dirichlet distribution over the class probabilities, interpreting the network outputs as evidence counts. However, some scholars [11, 68] argue that classical probability is not equipped to model 'second-level' uncertainty on the probabilities themselves. This has led to the formulation of numerous uncertainty calculi [38], including:

- Possibility Theory [46].
- Probability Intervals [58].
- Credal Sets [84, 144]: Convex sets of probability distributions.
- Random Sets [110]: Assigning mass to subsets of outcomes.

In imprecise-probabilistic models [21, 146], predictions correspond to credal sets or random sets [138, 95], allowing for a rigorous distinction between *aleatoric* (data) and *epistemic* (model) uncertainty [75, 94].

Random-set Neural Networks (see Chapter 5) builds on this by predicting belief functions over a budgeted set of classes, eliminating the need for inference-time sampling or prior selection.

3.2 Generative Adversarial Networks (GANs)

Generative models have achieved remarkable progress, with GANs standing out for their sample quality [18, 117]. However, lack of output diversity and mode collapse remain persistent challenges [120, 92].

3.2.1 Addressing Mode Collapse

Existing methods to encourage diversity in GANs can be categorized into three groups:

1. Loss Function Modifications:

- Wasserstein GAN (WGAN) [6]: Replaces Jensen-Shannon divergence with Wasserstein distance to improve stability.
- Least Squares GAN (LSGAN) [96]: Uses least squares loss to smooth gradients.
- Relativistic GAN (RGAN) [73]: Considers relative realism between real and generated samples.

2. Architectural Modifications:

- InfoGAN [23]: Maximizes mutual information between latent codes and outputs.
- Multi-Generator/Discriminator: MAD-GAN [53] and GMAN [47] use multiple components to cover different modes.
- PacGAN [87]: Feeds multiple samples to the discriminator to detect lack of diversity.
- BigGAN [18]: Scales up training with large batches and orthogonal regularization.

3. Training Strategies:

- Unrolled GAN [100]: Unrolls discriminator steps to stabilize training.
- UniGAN [115]: Introduces uniformity regularization within a normalizing flow framework.

Despite these advances, most solutions lack a principled theoretical framework for modeling the uncertainty inherent in the generative process. This work introduces Epistemic GANs (see Chapter 9) to fill this gap using Dempster-Shafer theory.

3.3 Diffusion Models

Diffusion Models [133, 64] represent the current state-of-the-art in visual generation. They incrementally add noise to data (forward process) and learn to reverse it (denoising process).

3.3.1 Categories of Diffusion Models

1. Denoising Diffusion Probabilistic Models (DDPMs) [64]: The primary focus of our work.

2. Noise-Conditioned Score Networks (NCSNs) [135]: Learn gradients of the data density.

3. Stochastic Differential Equations (SDEs) [136]: Model diffusion via continuous-time SDEs.

3.3.2 Challenges: Inference Speed and Diversity

While Diffusion models solve GAN stability issues, they suffer from slow inference [134]. Many methods address this, such as progressive distillation [125] and latent diffusion [122].

However, a critical and less-discussed issue is Diversity. Quantitative metrics like the Vendi Score reveal that diffusion samples are often *less* diverse than the training data [49]. Furthermore, standard DDPMs often fix the variance schedule or learn it within a rigid range [111], limiting the model's ”creativity.” The Epistemic Diffusion (see Chapter 8) approach addresses this by learning a second-order distribution over the noise parameters.

3.4 Uncertainty in Large Language Models (LLMs)

LLMs [139, 1] have revolutionized NLP but suffer from hallucinations [99]. Quantifying uncertainty in LLMs is an active research area.

3.4.1 Current Approaches

- Entropy-based: Using Softmax entropy [118] or Semantic Entropy [81] over multiple generations.
- Calibration: Measuring calibration errors to assess trustworthiness [55].
- Hidden States: probing internal representations [22].
- Bayesian/Ensemble: Applying Monte-Carlo Dropout [51], Bayes by Backprop, or Ensembles [8] to LLMs.
- Fine-tuning: Training models to verbally express their uncertainty [86].

Most of these methods are either post-hoc (prompting, sampling) or computationally heavy (ensembles). The Random-Set LLM (see Chapter 7) proposes a fundamental architectural change: predicting belief functions over token sets directly, providing an intrinsic measure of epistemic uncertainty without requiring multiple inference passes.

Chapter 4

BUDGETING OF FOCAL SETS

4.1 Motivation and Background

In the previous chapters, we established that random set [37] over a frame of discernment $\Theta = \{c_1, c_2, \ldots, c_n\}$ are a great tool for modelling ignorance. However, we also hit a wall: the computational complexity. If we have a classification problem with classes, the number of possible subsets is 2 . For a small dataset like CIFAR-10, that is $2^{10} = 1024$ sets. Manageable. But for ImageNet (= 1000), it is 2^{1000}, which is more than the number of atoms in the universe.

Existing solutions often limit focal sets arbitrarily by cardinality [44] or restrict the method to small datasets [126]. To address this, we propose a strategic budgeting method based on traditional clustering to identify and select only the most relevant ''focal sets,'' reducing the output space from $2^{NK}$to a manageable size. The core idea is to use clustering to discover which classes are naturally confused with each other. If the data for "Dog" and "Cat" overlaps significantly in the feature space, then the set $\{Dog, Cat\}$ is a meaningful focal set. But the set $\{Dog, Airplane\}$ probably isn't. By finding these natural groupings, we can build a ''budget'' of sets that covers the epistemic uncertainty without wasting resources.

4.2 The Problem of the Power Set

Let $= \{\omega_1, \ldots, \omega K\}$ be our set of classes. A standard classifier maps an input x to one of these classes. An epistemic classifier maps x to a mass function m over 2 .

The challenge is two-fold:

1. **Memory**: Storing a vector of size 2 is impossible for large .

2. **Training**: Ideally, we want the model to learn which sets are important. But if the output space is too huge, the model will never converge.

Existing methods like Evidential C-Means (ECM) [97] try to solve this by optimizing the centers of the focal sets. However, even ECM struggles when gets large. We need a heuristic, a pre-processing step that tells us: ”Here are the 100 sets you should care about. Ignore the rest.”

4.3 Budgeting Algorithm

Our strategy selects relevant subsets (focal sets) out of the 2^{NK}possibilities to serve as the output nodes of the neural network. This process relies on analyzing the semantic and feature-space overlap between classes in the training data.

The methodology proceeds as follows:

1. **Feature Extraction:** Given a training set with NKclasses, we extract feature vectors for each sample from the penultimate layer of a standard, pre-trained CNN.

2. **Dimensionality Reduction:** To facilitate efficient clustering and visualization, we reduce the dimensionality of these feature vectors to 3 dimensions. While our experiments utilize the t-SNE (t-Distributed Stochastic Neighbor Embedding) algorithm [142], the approach is agnostic and compatible with other techniques such as autoencoders.

3. **Clustering and Ellipsoid Fitting:** We fit a Gaussian Mixture Model (GMM) to the reduced feature vectors for each class. For every class $c \in C$, we obtain a mean vector $\mu_{cK}$and a covariance matrix Σ_c.

As visualized in Figure 4.1, we define an ellipsoid for each class that covers 95% of its data points [137]. The geometry of these ellipsoids is derived from the eigen-decomposition of the covariance matrix Σ_c. The length of the principal axes is calculated as:

$$length_{c,iK} = 2\sqrt{7.815\,\lambda_{iK}} \tag{4.1}$$

where λ_{iK} is the i^{thK} eigenvalue and the scalar 7.815 corresponds to the 95% confidence interval for a chi-square distribution with 3 degrees of freedom.

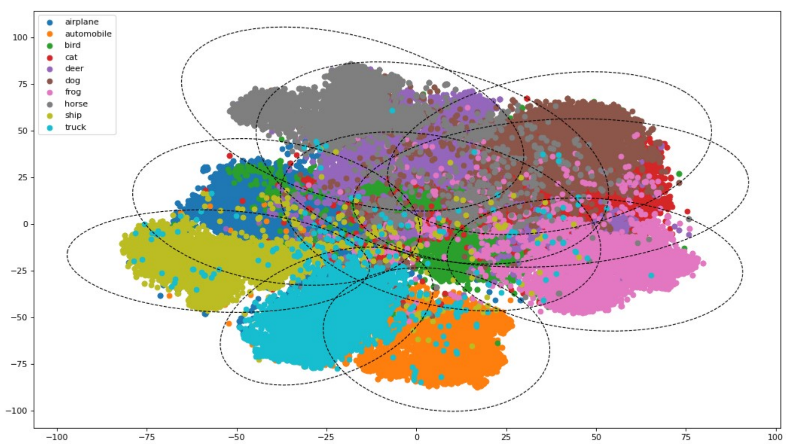


Figure 4.1: 2D visualization of the clusters of 10 classes of datapoints belonging to the CIFAR10 dataset, together with the ellipses formed by RS-CNN using Gaussian Mixture Models.

4. **Overlap Calculation:** The core of the budgeting strategy involves computing the geometric overlap between these class ellipsoids in the 3D space. We calculate the Intersection over Union (IoU) for subsets AK in the power set $\mathbb{P}(C)$:

$$\text{overlap}(A) = \frac{\cap_{i\in A} A_i}{\cup_{i\in A} A_{iK}},\ A \in \mathbb{P}(C).K \tag{4.2}$$

To maintain computational feasibility, we begin calculating overlaps for subsets of cardinality 2 and increment upwards. We employ early stopping when increasing cardinality no longer yields sets with significant overlap.

5. **Selection:** We select the top- subsets (AK) with the highest overlapping ratios. These non-singleton focal sets, combined with the original N singleton sets, constitute the final output budget $\mathcal{O}$.

Algorithm 1 provides a step-by-step illustration for our proposed budgeting strategy.

Algorithm 1 Budgeting Algorithm

Input: $\mathcal{D}$ – Training data with NKclasses, C – The set of classes, – Number of non-singleton focal sets
Output: $\mathcal{O}$ – Set containing $N+$ focal sets
Initialization
Extract feature vectors using a trained CNN
Apply t-SNE for dimensionality reduction to 3 dimensions
for each class c **do**
    Fit GMM to the reduced feature vectors for that class to obtain $\mu_{cK}$and Σ_{cK}
    Define an ellipsoid covering 95% data using $\mu_{cK}$and Σ_{cK}
end for
most_overlapping_sets ← Initialize an empty list for non-singleton focal sets $A_1, \ldots, AK$
Set $current\_cardinality \leftarrow 2$
while $current\_cardinality \leq NK$**do**
    Compute overlaps for subsets of cardinality $current\_cardinalityK$
    overlap$(A) = \frac{\cap_{c\in A} A^c}{\cup_{c\in A} A^c}$
    Select top- subsets with highest overlap
    Update most_overlapping_sets
    if no change in most_overlapping_sets **then**
        break
    end if
    $current\_cardinality \leftarrow current\_cardinality + 1$
end while
Combine selected non-singleton focal sets with NKsingleton sets
$\mathcal{O} \leftarrow C \cup \{A_1, \ldots, AK\}$
return $\mathcal{O}$

4.4 Case Study: CIFAR-10

To demonstrate the efficacy of this approach, we applied the algorithm to the CIFAR-10 dataset [79]. Computing predictions for all 1024 theoretical subsets is inefficient. Instead, we selected a budget of $= 20$ non-singleton focal sets.

The algorithm successfully identified semantically and visually similar groupings, such as:

- {`cat`, `dog`}
- {`horse`, `deer`}
- {`bird`, `airplane`}
- {`automobile`, `truck`}
- {`cat`, `horse`, `dog`}

These 20 focal sets, added to the 10 singletons, formed the output layer for the RS-CNN. While stochastic elements in t-SNE may cause slight variations across runs, robust pairs like {`bird`, `airplane`} appear consistently. This budgeting preprocessing is critical for scaling RS-NNs [95] to larger datasets, such as ImageNet..

4.5 Ablation Study on Budget Size ( )

The number of non-singleton focal sets, , is a crucial hyperparameter. A value of = 0 reduces the model to classical classification, while excessive increases computational complexity without necessarily adding informational value.

We conducted an ablation study on the CIFAR-10 dataset using Random-set Neural Networks (described in Chapter 5) to determine the optimal budget size. As shown in Figure 4.2, performance does not scale linearly with . We found that a small value of (comparable to the number of classes, N) yields the best performance, surpassing the baseline of = 0. This indicates that incorporating a carefully selected budget of sets improves model performance—measured via accuracy derived from Smets' Pignistic Transform [132]—while simultaneously enabling uncertainty quantification.

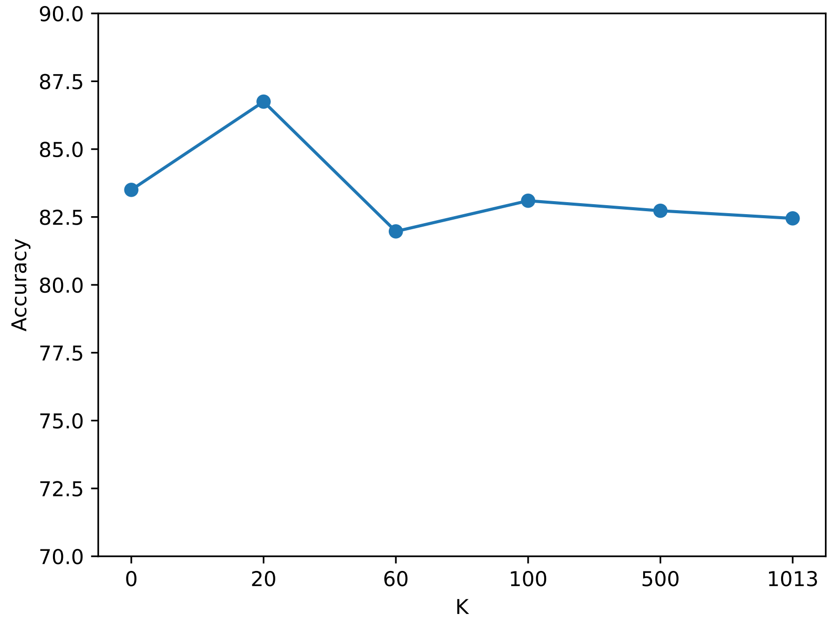


Figure 4.2: Ablation study on number of non-singleton focal sets on Cifar-10 dataset. The maximum value of K can be 1013 for 10 classes (after excluding the singletons and empty set).

4.6 Conclusion

This chapter addressed the critical bottleneck hindering the application of random-set theory to deep learning: the exponential explosion of the power set. By reframing the selection of focal sets as a Budgeting problem, we moved from an intractable output space of size 2 to a manageable subspace of size $+N+1$.

Our proposed methodology leverages unsupervised clustering (GMMs on t-SNE embeddings) to identify the ”natural confusion” between classes. Rather than arbitrarily selecting sets, we geometrically derive the subsets of classes that overlap significantly in the feature space. As demonstrated in our CIFAR-10 case study, this process successfully recovers semantically meaningful groupings (e.g., $\{bird, airplane\}$ or $\{cat, dog\}$) without any explicit semantic supervision.

Crucially, our ablation studies reveal that ”more is not always better.” A massive budget does not necessarily improve performance; instead, a carefully selected, small budget (comparable to the number of singleton classes) is sufficient to capture the necessary epistemic uncertainty. This insight is pivotal: it implies that we can equip standard neural networks with rigorous uncertainty quantification capabilities with negligible computational overhead.

This budgeting framework serves as the foundational pre-processing step for all the random-set architectures proposed in this thesis. Whether applied to classification (Chapter 5), language modelling (Chapter 7), or autonomous driving (Chapter 10), the ability to intelligently select *which* sets to believe in is what makes the theory practical. By solving the ”Problem of the Power Set,” we have cleared the path for scalable, uncertainty-aware deep learning.

Chapter 5

RANDOM-SET NEURAL NETWORKS

In the previous chapter, we solved the ‘‘explosion’’ problem. We used unsupervised clustering to find a manageable budget of focal sets that represent the most confusing classes in our dataset (like {Dog, Cat} or {Car, Truck}). This chapter presents an application of these budgeted focal set in Random-set Neural Networks.

5.1 Introduction

Machine learning is being used more and more in safety-critical fields like autonomous driving, healthcare, and finance. In these areas, making a wrong prediction can have serious consequences. Because of this, it is not enough for a model to just be accurate; it also needs to know when it might be wrong. Ideally, an AI system should be able to say ”I don’t know” when it sees data that is very different from what it was trained on [57, 114, 102].

The problem is that most current deep learning models are ”overconfident.” They tend to give high probability scores even when they are making a mistake or looking at something they have never seen before [116]. This happens mainly because standard neural networks force their outputs to sum up to 1 (using the Softmax function). This makes it hard for the model to distinguish between ”noise in the data” (aleatoric uncertainty) and ”lack of knowledge” (epistemic uncertainty) [75, 68, 94].

In this chapter, we propose a new solution: the **Random-Set Neural Network (RS-NN)**. Instead of predicting a simple vector of probabilities, this model predicts *belief functions* based on the mathematics of *random sets* [105, 104]. This allows the network to assign probability mass to *sets* of classes (like ”Class A or Class B”) rather than just single classes. This is a much more natural way to represent ignorance.

Unlike Bayesian Neural Networks (BNNs), which can be computationally heavy and rely on choosing the right "priors" [48], or Deep Ensembles, which require training multiple models [82], the RS-NN is a single network that is efficient to train and run.

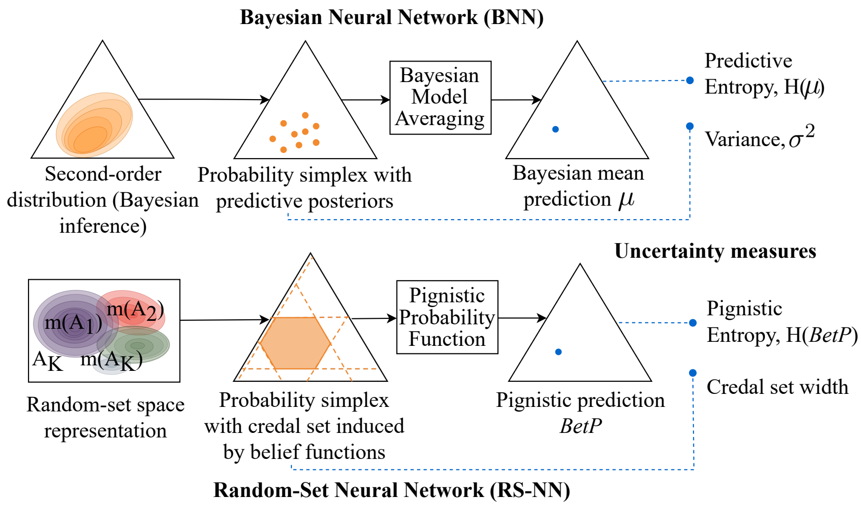


Figure 5.1: Inference in a Bayesian Neural Network (top) compared to a Random-Set Neural Network (bottom). While BNNs sample from a distribution of weights to get predictions, RS-NN maps the input directly to a belief function over a budget of focal sets.

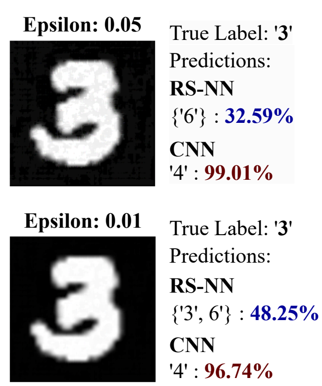


Figure 5.2: Confidence scores of RS-NN and CNN for FGSM adversarial attack on MNIST. The RS-NN correctly drops its confidence when attacked, while the CNN stays overconfident.

As shown in Figure 5.1, our approach works by predicting a belief function that corresponds to a convex set of probability distributions, known as a *credal set*. The size of this set tells us how uncertain the model is [5]. We also show that RS-NN is much more robust to attacks. For example, in Figure 5.2, when we attack the model with adversarial noise (FGSM), the confidence of the standard CNN stays very high (over 90%), while the RS-NN's confidence drops significantly, effectively warning us that something is wrong [56].

5.2 The Random-Set Neural Network

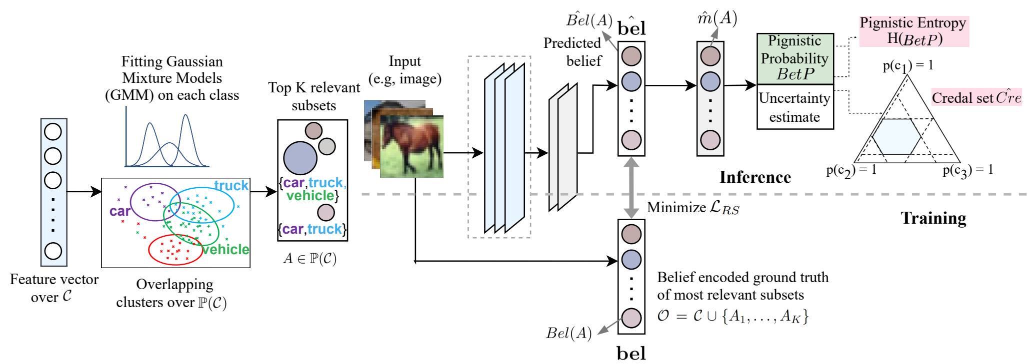
5.2.1 Architecture and Budgeting

The RS-NN is designed to be flexible. It acts as a "wrapper" that can be put on top of any standard deep learning model (like ResNet or EfficientNet). Instead of outputting scores for just the classes, it outputs scores for a "budget" of sets.

The main challenge with random sets is the sheer number of possible subsets. If we have NKclasses, there are 2^{NK}possible subsets. For a large dataset like ImageNet, this is impossible to compute. To solve this, we use the **Budgeting** method proposed in Chapter 4. We select a fixed number of "focal sets" that are most relevant. We do this by:

1. Taking a pre-trained model and extracting features from the data.
2. Reducing the dimensions of these features using t-SNE or UMAP.
3. Clustering the data using a Gaussian Mixture Model (GMM).
4. Finding which classes overlap in these clusters. If "Dog" and "Cat" often appear in the same cluster, we add the set $\{Dog, Cat\}$ to our budget.

This process is shown in Figure 5.3. By doing this, we reduce the problem to a manageable size. For example, for ImageNet, we used a budget of 3000 sets, which takes about 60 minutes to calculate once.



(a)

(b)

Figure 5.3: RS-NN model architecture. (a) *Budgeting*: We select the top relevant sets of classes using clustering. (b) *Training*: The network predicts belief functions. The grey layers are the ones we train, while the blue layers can be any pre-trained backbone.

5.2.2 Loss Function

Training the RS-NN is a bit different from standard networks. Since our output is a belief function, we treat the ground truth as a belief vector. If the true class is "Dog", then the belief in the set $\{Dog\}$ is 1, and the belief in $\{Dog, Cat\}$ is also 1 (because a Dog is indeed a "Dog or Cat").

A random-set prediction problem is mathematically similar to the multi-label classification problem, for in both cases the corresponds to a vector of 0s and 1s and the prediction is vector in $[0,1]$. They do have different semantics, as in the case of random sets, they indicate if the set contains true class or not; whereas in multi-label classification, it is the probability of input belonging to a particular class. Still, despite the different semantics, we use a Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{BCEK} = -\frac{1}{b_{sizeK}} \sum_{i=1}^{b_{size}} \frac{1}{|\mathcal{O}|} \sum_{A \in \mathcal{O}} \Big[Bel_i(A) \log(\hat{Bel}_i(A)) + (1 - Bel_i(A)) \log(1 - \hat{Bel}_i(A)) \Big] .K \tag{5.1}$$

We also need to make sure the output is a valid mass function (non-negative and sums to 1). So, we add two regularization terms, M_{rK} and M_s:

$$M_{rK} = \sum \max(0, -\hat{m}(A)),K \quad M_{sK} = \sum \hat{m}(A) - 1 \;.K \tag{5.2}$$

The final loss is $\mathcal{L}_{RSK} = \mathcal{L}_{BCEK} + \alpha M_{rK} + \beta M_s$, where we typically set α and β to small values like $1e-3$.

5.2.3 *Measuring Uncertainty*

The RS-NN gives us two good ways to measure uncertainty:

1. **Entropy**: We can calculate the Shannon entropy of the pignistic probability prediction.

2. **Credal Set Width**: We can measure the size of the credal set. We approximate this by looking at the difference between the maximum and minimum possible probability for the predicted class:

$$\text{Width}(c) = \overline{P}(c) - \underline{P}(c) \tag{5.3}$$

5.3 Experiments

5.3.1 *Experimental Setup and Implementation*

Datasets and Protocols. To ensure a comprehensive evaluation of the Random-Set Neural Network (RS-NN), we conducted experiments across a diverse array of multi-class image classification benchmarks, ranging from grayscale digits to high-resolution object recognition. Specifically, we utilized MNIST [83], CIFAR-10 [78], Intel Image [10], CIFAR-100 [77], and the large-scale ImageNet dataset [43].

A critical component of our evaluation framework involves Out-of-Distribution (OoD) detection, a proxy for measuring epistemic uncertainty. We established several In-Distribution (iD) versus OoD pairs to test the models' ability to flag unseen semantics:

- **CIFAR-10 (iD)** vs. SVHN [108] and Intel-Image [10] (OoD).

- **MNIST (iD)** vs. Fashion-MNIST [151] and K-MNIST [27] (OoD).

- **ImageNet (iD)** vs. ImageNet-O [61] (OoD).

Data partitioning adhered to standard protocols: 40,000/10,000/10,000 samples for training/testing/validation respectively for CIFAR datasets; 50,000/10,000/10,000 for MNIST; 13,934/3,000/100 for Intel Image; and 1,172,498/50,000/108,669 for ImageNet. For OoD evaluation, we utilized 10,000 testing samples, with the exception of Intel Image (3,000) and ImageNet-O (2,000). All input images were resized to a standardized resolution of 224×224 pixels to ensure architecture consistency.

Baselines and Architectures. We benchmarked the RS-NN against a suite of state-of-the-art uncertainty quantification methods, ensuring coverage of both Bayesian and Frequentist paradigms:

1. **Bayesian Methods:** Laplace Bridge Bayesian Neural Networks (LB-BNN) [65] and Function-Space Variational Inference (FSVI) [123].

2. **Ensemble Methods:** Deep Ensembles (DE) [82] (comprising 5 models) and Epistemic Neural Networks (ENN) [113] (comprising 3 models).

3. **Standard Deterministic:** A standard CNN trained with Cross-Entropy loss, serving as a baseline for accuracy and overconfidence comparisons.

These baselines represent the current gold standard in the field, aligning with recent literature [26, 41, 150].

Training Configuration. To guarantee a fair comparison, all models (including RS-NN) utilized a ResNet50 backbone executed on NVIDIA A100 80GB GPUs. We employed a consistent training regimen: 200 epochs from scratch (using pre-trained weights only for ImageNet efficiency), a batch size of 128, and a learning rate scheduler initialized at 10^{-3} with decays at epochs 80, 120, 160, and 180. Standard data augmentation techniques [80], including random shifts and horizontal flips, were applied uniformly.

For the RS-NN specifically, the ResNet50 architecture was adapted to accommodate the random-set output. The final classification layer was replaced with a dense layer of size $|O|$ (the budget size), utilizing a sigmoid activation. This design choice reflects the multi-label nature of belief encoding (see Section 5.2.2). In contrast, all baseline models employed a standard softmax output.

Budgeting and Optimization. The RS-NN training process begins with the budgeting phase. Feature vectors were extracted using a pre-trained ResNet50. We employed 50 CPU cores for t-SNE dimensionality reduction and 150 CPU cores for computing class overlaps via GMM clustering. We set the budget of focal sets to 20 for MNIST/CIFAR-10/Intel Image, 200 for CIFAR-100, and 3000 for ImageNet. The model was trained using the proposed loss function $\mathcal{L}_{RSK}$(Eq. 5.1) with regularization hyperparameters set to $\alpha = \beta = 10^{-3}$.

5.3.2 *Standard Predictive Performance*

We first establish the baseline performance of the RS-NN in terms of generalization accuracy and computational efficiency. As detailed in Table 5.1, the RS-NN demonstrates competitive or superior test accuracy compared to computationally expensive ensembles (DE) and complex Bayesian methods (FSVI, LB-BNN). Crucially, the inference time of the RS-NN (1.91 ± 0.02 ms) matches that of a standard CNN, whereas Deep Ensembles require significantly higher latency (13163.50 ± 3.37 ms), highlighting the efficiency of our single-forward-pass approach.

Table 5.1: Test accuracies (%) and inference time (ms) for uncertainty estimation over 5 consecutive runs across methods and datasets. Average and standard deviation are shown for each experiment.

Datasets	MNIST	CIFAR-10	Intel Image	CIFAR-100	ImageNet (Top-1)	ImageNet (Top-5)	Inference time (ms)
RS-NN (ours)	**99.71 ± 0.03**	**93.53 ± 0.09**	**94.22 ± 0.03**	**71.61 ± 0.07**	**79.92**	**94.47**	**1.91 ± 0.02**
LB-BNN [65]	99.58 ± 0.04	89.95 ± 0.81	90.49 ± 0.42	59.89 ± 1.96	72.48	90.85	7.11 ± 0.89
FSVI [123]	99.18 ± 0.03	80.29 ± 0.05	88.92 ± 0.13	53.34 ± 0.09	62.56	84.69	340.25 ± 0.76
DE [82]	99.25 ± 0.01	92.73 ± 0.04	91.98 ± 0.11	70.53 ± 0.07	78.77	94.37	13163.50 ± 3.37
ENN [113]	99.07 ± 0.11	91.55 ± 0.60	91.49 ± 0.19	68.02 ± 0.26	71.82	89.48	3.10 ± 0.03
CNN	99.12 ± 0.04	92.08 ± 0.42	90.89 ± 0.10	65.50 ± 0.08	78.56	94.34	**1.91 ± 0.03**

5.4 Out-of-Distribution (OoD) Detection

We evaluated OoD detection capabilities using AUROC, AUPRC, and Expected Calibration Error (ECE). Table 5.2 demonstrates that RS-NN consistently outperforms baselines, particularly on the challenging ImageNet vs. ImageNet-O benchmark (AUROC 60.38 vs. 55.37 for DE), indicating a superior ability to distinguish between known and unknown data distributions.

Table 5.2: OoD detection performance and uncertainty estimation for models trained on ResNet50 on CIFAR-10 vs SVHN/Intel Image, MNIST vs F-MNIST/K-MNIST and ImageNet vs ImageNet-O. Evaluation metrics include AUROC/AUPRC (OoD); Entropy of predictions (uncertainty) and Expected Calibration Error (ECE).

In-distribution (iD)						Out-of-distribution (OoD)					
Dataset	Model	Test accuracy (%) (↑)	Uncertainty measure	In-distribution Entropy (↓)	ECE (↓)	SVHN			Intel Image		
						AUROC (↑)	AUPRC (↑)	Entropy (↑)	AUROC (↑)	AUPRC (↑)	Entropy (↑)
CIFAR-10	RS-NN	**93.53**	Pignistic entropy	**0.088 ± 0.308**	**0.0484**	**94.91**	**93.72**	**1.132 ± 0.855**	**97.39**	**90.27**	**1.517 ± 0.740**
	LB-BNN	89.95	Predictive Entropy	0.191 ± 0.412	0.0585	88.14	81.96	0.828 ± 0.243	82.21	55.17	0.763 ± 0.722
	FSVI	80.29	Predictive Entropy	0.118 ± 0.563	0.0521	80.59	80.84	0.413 ± 0.461	74.27	72.51	0.289 ± 0.670
	DE	92.73	Mean Entropy	0.154 ± 0.367	**0.0482**	93.84	91.88	0.939 ± 0.554	94.25	79.36	1.166 ± 0.552
	ENN	91.55	Mean Entropy	0.126 ± 0.323	0.0556	92.76	89.05	0.887 ± 0.514	85.67	58.09	0.600 ± 0.578
	CNN	92.08	Softmax Entropy	0.114 ± 0.304	0.0669	93.11	91.0	0.930 ± 0.610	87.75	65.54	0.719 ± 0.673
						F-MNIST			K-MNIST		
MNIST	RS-NN	**99.71**	Pignistic entropy	0.010 ± 0.111	**0.0029**	**93.89**	**93.98**	0.530 ± 0.770	**96.75**	**96.58**	**0.740 ± 0.917**
	LB-BNN	99.58	Predictive Entropy	**0.001 ± 0.085**	0.0032	89.65	90.36	0.287 ± 0.442	95.61	95.65	0.540 ± 0.621
	FSVI	99.18	Predictive Entropy	0.006 ± 0.265	0.0047	92.79	91.17	0.264 ± 0.289	91.65	95.75	0.313 ± 0.381
	DE	99.25	Mean Entropy	0.031 ± 0.155	0.0031	92.30	92.05	**0.584 ± 0.587**	95.81	94.71	0.564 ± 0.715
	ENN	99.07	Mean Entropy	0.022 ± 0.127	0.0039	81.79	82.92	0.313 ± 0.464	95.94	95.45	0.503 ± 0.672
	CNN	98.90	Softmax Entropy	0.023 ± 0.135	0.0052	83.77	84.14	0.278 ± 0.426	94.46	93.94	0.616 ± 0.688
						ImageNet-O					
						AUROC		AUPRC		Entropy	
ImageNet	RS-NN	**79.92**	Pignistic entropy	2.972 ± 2.108	**0.1416**	**60.38**		**55.16**		3.659 ± 3.771	
	LB-BNN	72.48	Predictive Entropy	2.471 ± 2.972	0.5812	41.08		30.99		1.383 ± 0.028	
	FSVI	62.56	Predictive Entropy	**1.328 ± 1.966**	0.3890	50.55		49.88		1.637 ± 1.328	
	DE	78.77	Mean Entropy	1.532 ± 1.325	0.1940	55.37		53.20		1.775 ± 1.343	
	ENN	71.82	Mean Entropy	1.395 ± 1.510	0.5961	54.67		43.73		1.617 ± 1.597	
	CNN	78.56	Softmax Entropy	6.386 ± 1.388	0.4004	54.28		48.73		**6.575 ± 1.512**	

Figure 5.4 illustrates the distribution of credal set widths (Eq. 7.6) across the CIFAR-10 test set. The density plot reveals a clear distinction based on classification accuracy: incorrect predictions (shown in red) are associated with larger credal widths and greater variance, signalling high epistemic uncertainty. In contrast, correct predictions (shown in blue) are tightly clustered around smaller values, reflecting the model's higher confidence.

Furthermore, Table 5.3 compares these credal intervals across In-Distribution (iD) and Out-of-Distribution (OoD) datasets. As expected, the model assigns significantly larger intervals to OoD samples, effectively flagging them as uncertain. An exception to this trend is observed in the comparison between ImageNet and

ImageNet-O, where the distinction is less pronounced due to the specific characteristics of the adversarial dataset. It is important to note that the credal set width is a distinct measure of epistemic ambiguity and is not semantically equivalent to metrics such as entropy, variance, or mutual information used in standard probabilistic models.

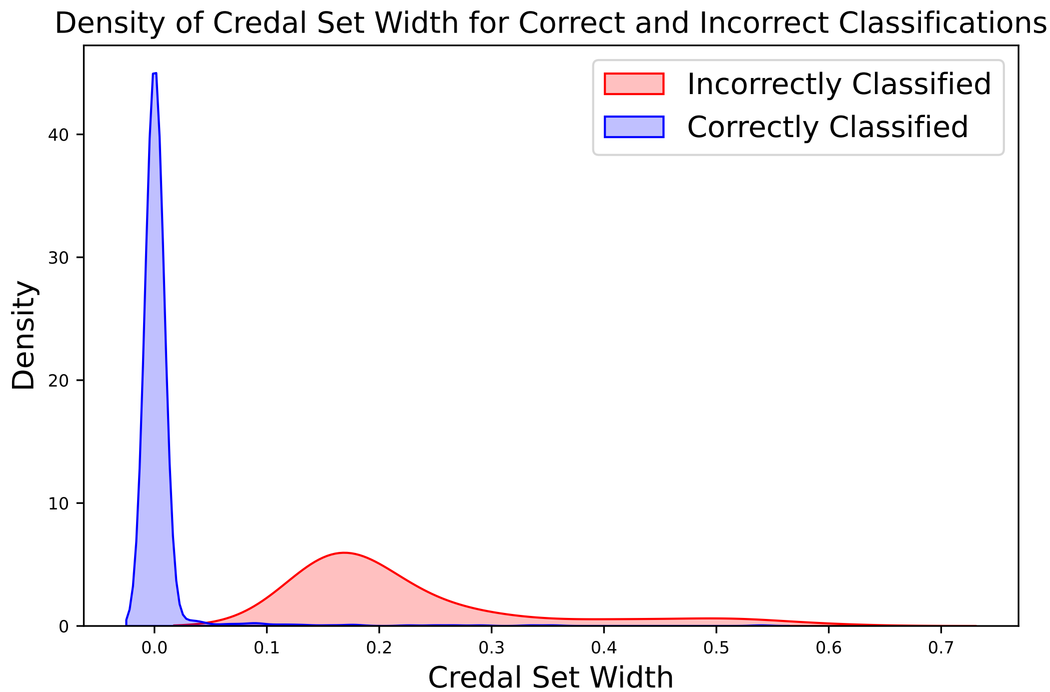


Figure 5.4: Width of credal predictions for CIFAR-10 test data (correctly classified, blue; incorrectly classified, red).

Table 5.3: Credal set width for RS-NN on iD vs OoD datasets: CIFAR10 vs SVHN/Intel Image, MNIST vs F-MNIST/K-MNIST and ImageNet vs ImageNet-O.

In-distribution (iD)		Out-of-distribution (OoD)			
CIFAR10	0.007 ± 0.044	SVHN	0.260 ± 0.322	Intel Image	0.587 ± 0.367
MNIST	0.001 ± 0.013	F-MNIST	0.070 ± 0.167	K-MNIST	0.103 ± 0.200
ImageNet	0.238 ± 0.266	ImageNet-O		0.272 ± 0.275	

5.4.1 Robustness to Distributional Shifts

Beyond standard benchmarks, a robust uncertainty estimator must maintain reliability under significant distributional shifts. We extended our evaluation to include synthetic perturbations, specifically investigating performance on Noisy and Rotated MNIST datasets.

We partitioned the MNIST dataset and subjected the test set to varying degrees of corruption. For the noisy condition, we introduced 50% random noise. For the rotation condition, images were subjected to random rotations within specific angular intervals (e.g., $-60°$ to $0°$).

Quantitative Analysis. Table 5.4 presents the accuracy of RS-NN versus a standard CNN under rotation shifts. The RS-NN consistently outperforms the standard CNN across all rotation angles. Notably, in the fully random rotation setting ($0°/360°$), the RS-NN achieves an accuracy of 47.71%, compared to 45.86% for the standard CNN, demonstrating superior resilience to geometric transformations.

Table 5.4: Test accuracies (%) for RS-NN and standard CNN on Rotated MNIST out-of-distribution (OOD) samples. Rotation angle is randomized within the specified intervals.

Rotation (angle)	-180/-120	-120/-60	-60/0	0/60	60/120	120/180	0/360
Standard CNN	36.41%	21.86%	74.41%	80.80%	23.89%	37.53%	45.86%
RS-CNN	**37.84**%	**23.54**%	**78.44**%	**81.46**%	**26.31**%	**38.48**%	**47.71**%

Qualitative Analysis: Mitigating Overconfidence. A key failure mode of standard deep learning models is high-confidence misclassification on OoD data. Table 5.6 provides a granular look at specific predictions on noisy and rotated samples.

The results illuminate the limitations of standard CNNs in ambiguous scenarios. For example, considering the noisy sample with True Label '3', the standard CNN confidently but incorrectly predicts class '8' with a probability of 0.969. In sharp contrast, the RS-NN assigns significant mass to the singleton {'3'} and composite sets containing '3' (e.g., {'3', '5'}), resulting in a corrected pignistic probability where the true class '3' is the most likely outcome (0.427), albeit with appropriately lowered confidence reflecting the ambiguity. Similarly, for the rotated label '9', where the standard CNN fails with 0.988 confidence in class '5', the RS-NN correctly identifies '9' as the most likely class.

5.4.2 *Hyperparameter Sensitivity Analysis*

The training stability and convergence of the RS-NN are influenced by the regularization hyperparameters α and β in the loss function $\mathcal{L}_{RSK}$(Eq. 5.1), which enforce the non-negativity and sum-to-one axioms of mass functions.

We conducted a sensitivity analysis on the CIFAR-10 dataset by varying α and βK across the set $\{0.5, 0.6, 0.9, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0\}$. Figure 5.5 illustrates the impact of these parameters on test accuracy.

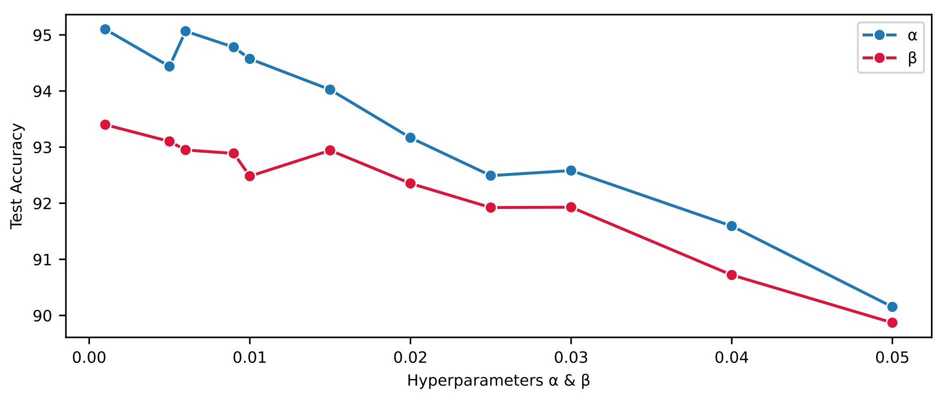


Figure 5.5: Ablation study of hyperparameters α and β on CIFAR-10 test accuracy. The model exhibits optimal performance when αK and βK are minimized, suggesting that while regularization is necessary for validity, excessive constraints can hamper optimization.

Our results indicate that test accuracy is maximized when αK and βK are small, striking a balance between enforcing theoretical constraints and allowing gradient flow for the primary classification objective.

5.4.3 Scalability to Large-Scale Architectures

Finally, to validate the architectural agnosticism of our approach, we applied the RS-NN wrapper to a variety of modern deep learning backbones. Table 5.5 demonstrates that the RS-NN maintains or improves upon the accuracy of standard deterministic models across WideResNet-28-10, VGG16, InceptionV3, EfficientNetB2, and Vision Transformers (ViT-Base-16). This confirms that the benefits of random-set uncertainty modeling are not confined to specific architectures but are generally transferable.

Table 5.5: Adaptability to large-scale model architectures with test accuracy (%) and parameters (in million) reported on CIFAR10.

	Model	Pre-trained R50	WRN-28-10	VGG16	IncepV3	ENetB2	ViT-Base
Test acc. (%)	RS-NN	**94.42**	**93.58**	**87.87**	**78.24**	**92.10**	86.75
	CNN	94.38	92.79	84.14	76.89	90.02	**87.21**
Params (M)	RS-NN	2.69	37.0	15.12	31.22	7.72	9.53
	CNN	2.62	36.7	15.11	31.21	7.71	9.52

5.5 Conclusion

In this chapter, we introduced the Random-Set Neural Network (RS-NN), a novel architecture that generalizes standard classification by predicting belief functions instead of probability distributions. By leveraging the budgeting strategy developed in Chapter 4, we successfully scaled the random-set framework to high-dimensional datasets like ImageNet, overcoming the computational intractability that has historically hindered this approach.

Our extensive experiments demonstrate that RS-NNs achieve a superior trade-off between predictive performance and uncertainty quantification compared to existing methods. Specifically:

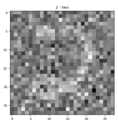
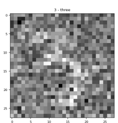
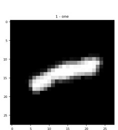
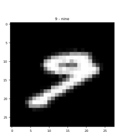
1. Accuracy and Efficiency: RS-NNs match or exceed the accuracy of Deep Ensembles and Bayesian Neural Networks while maintaining the inference speed of a standard deterministic CNN (1.91 ms vs. 13163 ms for Ensembles).

2. Robustness: The model exhibits remarkable resilience to adversarial attacks (FGSM) and distributional shifts (rotation, noise), correctly lowering its confidence in ambiguous scenarios where standard models remain overconfident.

3. Uncertainty Quantification: Through metrics like credal set width and pignistic entropy, RS-NNs provide a reliable signal for detecting Out-of-Distribution samples, significantly outperforming baselines on challenging benchmarks like ImageNet-O.

These results position the RS-NN as a robust and efficient solution for safety-critical applications, where the ability to ‘‘know when you don’t know’’ is just as important as being right.

Table 5.6: Comparison of predictions on Out-of-Distribution (OoD) noisy and rotated MNIST samples. Standard CNNs frequently predict incorrect classes with high confidence, whereas the RS-NN distributes belief mass to correctly encompass the true label or exhibit lower pignistic confidence.

	Standard CNN Predictions	Belief RS-CNN Predictions		
True Label = 2 	**Class 0 0.6287534** Class 2 0.2320247 Class 3 0.1172896 Class 8 0.0158272	**Belief values** {’2’} 0.98514729 {’2’, ’0’, ’1’} 0.9816562 {’2’, ’1’} 0.9806699 {’2’, ’4’} 0.97984832	**Mass values** {’2’} 0.9821171 {’0’, ’8’} 0.0090966 {’7’, ’0’, ’1’} 0.0029499 {’7’, ’8’} 0.00214586	**Pignistic Probability** **2 0.9821171** 8 0.007800460 0 0.100259813 3 0.00477754
True Label = 3 	**Class 8 0.96970748** Class 5 0.01849810 Class 9 0.00648078 Class 2 0.00339692	**Belief values** {’3’, ’5’} 0.77293962 {’6’, ’3’} 0.63796818 {’6’, ’3’, ’5’} 0.6200498 {’3’} 0.54052752	**Mass values** {’3’} 0.3612153 {’7’, ’8’} 0.1341305 {’0’, ’8’} 0.1040580 {’8’} 0.08483259	**Pignistic Probability** **3 0.42731909** 8 0.20502196 5 0.11517434 7 0.0801848
True Label = 1 	**Class 2 1.0** Class 1 3.3913153e-08 Class 6 1.0346271e-10 Class 3 4.9296049e-11	**Belief values** {’2’, ’1’} 0.99997830 {’1’, ’9’} 0.95868313 {’1’} 0.924648463 {’1’, ’5’} 0.82559686	**Mass values** {’1’} 0.55619530 {’2’} 0.423250733 {’1’, ’9’} 0.02047233 {’6’} 4.316419e-05	**Pignistic Probability** **1 0.56643147** 2 0.423250733 9 0.010245241 6 4.3164194e-05
True Label = 9 	**Class 5 0.9889484** Class 2 0.01012067 Class 3 0.00046459 Class 7 0.0002942	**Belief values** {’7’, ’5’, ’9’} 0.83161890 {’7’, ’9’} 0.70636504 {’5’, ’9’} 0.57625019 {’3’, ’9’} 0.5581744	**Mass values** {’7’, ’9’} 0.17115316 {’3’} 0.1345874 {’9’} 0.13252735 {’7’, ’8’} 0.1028051	**Pignistic Probability** **9 0.69936884** 7 0.02370170 5 0.004222436 3 0.001317382

Chapter 6

EVALUATION UNDER UNCERTAINTY

6.1 Introduction

In the preceding chapters, we introduced the Random-Set Neural Network (RS-NN) and demonstrated its capacity to model epistemic uncertainty via belief functions. However, a fundamental challenge in the field of uncertainty quantification is the lack of a standardized, objective framework for evaluating set-valued or epistemic predictions.

Standard evaluation metrics in machine learning, such as Accuracy, Negative Log-Likelihood (NLL), or Brier Score, are predicated on point-estimate probabilities. They penalize any deviation from a precise ground truth (typically a one-hot vector). This creates a ”precision trap” for epistemic classifiers: a model that correctly identifies ambiguity (e.g., predicting the set $\{Dog, Cat\}$ for a blurred image) is penalized for being imprecise, while an overconfident model that guesses wrong is penalized for error, but often less severely than it should be.

To address this, we propose a **Unified Evaluation Framework**. This framework introduces a novel metric, $\mathcal{E}$, which explicitly mathematically formalizes the trade-off between predictive accuracy (distance to ground truth) and epistemic uncertainty (non-specificity or credal set size). This allows us to evaluate models not just on their correctness, but on the quality of their uncertainty representation.

6.2 Classes of Epistemic Predictions

To evaluate different uncertainty models fairly, we must first understand the geometric nature of their predictions. The predictions of a classifier can be visualized within a simplex $\mathcal{P}$—the convex hull of one-hot probability vectors, where each vertex represents certainty in a specific class.

6.2.1 Point-Estimate Models

Standard Neural Networks (SNNs) predict a single probability vector $\hat{p}_s(y \mid \mathbf{x}, \mathbb{D})$ using the softmax function. Geometrically, this is a single point within the simplex. While accurate for in-distribution data, SNNs are often overconfident, collapsing to a vertex even for unknown data.

Deep Deterministic Uncertainty (DDU) models [106] also output softmax probabilities $\hat{p}_{ddu}$. However, they estimate uncertainty in the feature space by computing feature density, distinguishing in-distribution (iD) from out-of-distribution (OoD) samples without altering the prediction space geometry significantly.

6.2.2 Probabilistic and Ensemble Models

Bayesian Neural Networks (BNNs) [65] compute a predictive distribution by integrating over a posterior of model parameters. In practice, this integration is approximated via Bayesian Model Averaging (BMA), which averages predictions from multiple parameter samples. While BMA yields a single point prediction, it often smooths out the diverse, conflicting information present in the samples [63]. To capture the full epistemic uncertainty, we consider the *set* of prediction samples before averaging.

Deep Ensembles (DEs) [82] operate similarly but obtain diverse predictions by training independent models. The averaged prediction is $\hat{p}_{deK} = \frac{1}{} \sum \hat{p}_k$. Like BNNs, the spread of the individual ensemble members $\hat{p}_{kK}$contains valuable information about ambiguity that averaging obscures.

Evidential Deep Learning (EDL) [127] predicts the parameters of a Dirichlet distribution. While this defines a density over the simplex, for decision-making it is often collapsed to a mean point estimate. We effectively treat the Dirichlet distribution as a generator of probability samples.

6.2.3 Set-Valued Models

Credal Models represent uncertainty using a Credal Set [84, 32]—a convex set of probability distributions. This set can be defined, for instance, by probability intervals $[\underline{\hat{p}}(y), \hat{\overline{p}}(y)]$ for each class:

$$\hat{\mathbb{C}}r(y|\mathbf{x},\mathbb{D}) = \{p \in \mathcal{P} \mid \underline{\hat{p}}(y) \leq p(y) \leq \hat{\overline{p}}(y), \forall y \in \mathbf{Y}\}. \tag{6.1}$$

Belief Function Models, such as our Random-Set Neural Network (RS-NN), predict a mass function m over subsets of classes. As discussed in Chapter 5, a belief function $\hat{Bel}$ is mathematically equivalent to a credal set $\mathbb{C}r_{\hat{Bel}}$ containing all distributions consistent with the belief evidence:

$$\mathbb{C}r_{\hat{Bel}}(y|\mathbf{x},\mathbb{D}) = \{p \in \mathcal{P} \mid p(A) \geq \hat{Bel}(A)\}. \tag{6.2}$$

The center of mass of this set is the Pignistic Probability ($BetP$), used for making point predictions.

6.3 Mapping Predictions to Credal Sets

To compare these diverse models, we map all their outputs to a common representation: the Credal Set. For Credal and Belief Function models, this representation is native. For Bayesian, Ensemble, and Evidential models, we construct a credal set from the spread of their prediction samples using the theory of Coherent Lower Probabilities [103].

6.3.1 Construction Procedure

Given a sample of predicted probability vectors (e.g., from an ensemble or BNN posterior), we follow this three-step process:

1. Compute Lower Probabilities: We define a lower probability $\underline{P}(A)$ for any subset of classes $A \subseteq \mathbf{Y}$ as the minimum probability assigned to A across all samples

in the collection:

$$\underline{P}(A) = \min_{kK} P_k(A).K \tag{6.3}$$

To maintain computational feasibility for large class counts N, we calculate this only for a "budget" of relevant subsets obtained via clustering (as in RS-NN).

2. Compute Mass Function: We convert this lower probability into a mass function $m_{\underline{P}K}$using the Möbius inversion formula:

$$m_{\underline{P}}(A) = \sum_{B \subseteq AK} (-1)^{|A \setminus B|} \underline{P}(B).K \tag{6.4}$$

Negative masses are clipped to zero to ensure validity.

3. Compute Credal Vertices: Finally, we identify the vertices (extreme points) of the resulting credal set using the mass function. These vertices $p^{\pi K}$correspond to permutations of the class labels:

$$p^{\pi}(x_{\pi(i)}) = \sum_{A \ni x_{\pi(i)};\, A \not\ni x_{\pi(j)}\ \forall j<iK} m_{\underline{P}}(A).K \tag{6.5}$$

This procedure allows us to visualize standard ensembles as "clouds" or polytopes in the simplex, making them directly comparable to belief functions (See Figure 6.1 for a conceptual visualization).

6.4 Evaluation of Epistemic Predictions

We propose a unified evaluation metric $\mathcal{E}$ that assesses the quality of a credal set based on two conflicting objectives: Accuracy (closeness to the truth) and Non-Specificity (informativeness).

6.4.1 The Unified Metric

For a single data point with ground truth y, the metric is defined as:

$$\mathcal{E} = d(y, \hat{y}) + \lambda \cdot NS[m],K \tag{6.6}$$

where:

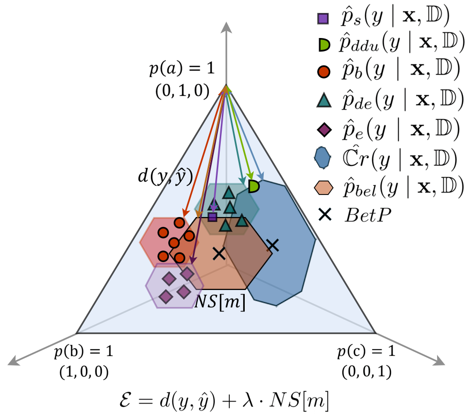


Figure 6.1: Different types of uncertainty-aware model predictions, shown in a unit simplex of probability distributions defined on the list of classes $\mathbf{Y} = \{a, b, c\}$. Our proposed evaluation framework uses a metric which combines, for each input $\mathbf{x}$, a distance (arrows) between the corresponding ground truth (e.g., $(0, 1, 0)$) and the *epistemic predictions* generated by the various models (in the form of credal sets), and a measure of the extent of the credal prediction (*non-specificity*).

- $d(y, \hat{y})$ is the Distance between the ground truth yK(a vertex of the simplex) and the predicted credal set $\hat{y}$. We typically use the Kullback-Leibler (KL) Divergence to the closest point (vertex) of the credal set. This rewards models whose set of plausible probabilities includes the truth.

- $NS[m]$ is the Non-Specificity, measuring the imprecision of the prediction. We use the definition by [45]:

$$NS[m] = \sum_{A \subseteq \mathbf{Y}} m(A) \log |A| .K \tag{6.7}$$

  High non-specificity means the model is vague (predicting large sets).

- λ is a trade-off parameter that reflects the user's risk tolerance.

6.4.2 Rationale and Scenarios

The metric penalizes two failure modes:

1. Overconfidence: High distance dK(the truth is outside the credal set) and low NS (the set is small).

2. Useless Vagueness: Low distance d (truth is inside) but very high NS (the set is huge, e.g., ”it’s one of the 1000 classes”).

The choice of λ depends on the application:

- Crop Disease Classification (Low λ): Abstention is allowed. We prioritize containing the truth (low KL) and tolerate some vagueness (high NS).
- Autonomous Driving (High λ): Decisions must be made instantly. We penalize vagueness heavily to force the model to be precise, ensuring decisiveness.

6.5 Experiments

We conducted extensive experiments to benchmark state-of-the-art uncertainty models using our unified metric.

6.5.1 Experimental Setup

Baselines: We evaluated (1) Standard Neural Network (SNN), (2) Laplace Bridge BNN (LB-BNN) [65], (3) Deep Ensemble (DE) [82], (4) Evidential Deep Learning (EDL) [127], (5) Deep Deterministic Uncertainty (DDU) [106], (6) Credal-Set Interval NN (CreINN) [146], (7) Evidential CNN (E-CNN) [138], and our (8) Random-Set NN (RS-NN).

Datasets: We used MNIST, CIFAR-10, and CIFAR-100. RS-NN generates belief functions for a budgeted set of outcomes (30 focal sets for CIFAR-10, 300 for CIFAR-100) to ensure efficiency.

6.5.2 Analysis of the Evaluation Metric

Table 6.1 details the performance of all models.

Results Discussion:

- CIFAR-10: Deep Ensembles (DE) achieve the highest accuracy (93.77%) and a very low Evaluation Metric ($\mathcal{E}$) when averaged. However, RS-NN is ex-

Table 6.1: Comparison of Kullback-Leibler divergence (KL), Non-Specificity (NS) and Evaluation Metric ($\mathcal{E}$) for uncertainty-aware classifiers (trade-off $\lambda = 1$). Mean and standard deviation are shown for CIFAR-10, MNIST and CIFAR-100 datasets.

Dataset	Model	Test accuracy (%)(↑)	ECE (↓)	KL divergence (KL)	Non-Specificity (NS)	Evaluation metric ($\mathcal{E}$)(↓)
CIFAR-10	LB-BNN	89.24	0.0565	0.243 ± 1.315	0.166 ± 0.398	0.409 ± 1.381
	DE	**93.77**	**0.0075**	0.031 ± 0.367	0.385 ± 0.715	0.415 ± 0.805
	EDL	59.13	0.0491	0.002 ± 0.011	2.267 ± 0.067	2.270 ± 0.066
	CreINN	88.36	0.0108	0.058 ± 0.374	0.596 ± 0.812	0.654 ± 0.892
	E-CNN	83.5	0.6497	0.193 ± 0.215	1.609 ± 0.003	1.802 ± 0.215
	RS-NN	92.99	0.0509	0.398 ± 1.895	0.009 ± 0.052	**0.407 ± 0.500**
	SNN	90.25	0.0668	0.481 ± 1.797	0.000 ± 0.000	0.481 ± 1.797
	LB-BNN Avg	89.24	0.0565	0.420 ± 1.520	0.000 ± 0.000	0.420 ± 1.520
	DE Avg	**93.77**	**0.0075**	0.195 ± 0.763	0.000 ± 0.000	**0.195 ± 0.763**
	DDU	91.34	0.0439	0.309 ± 1.115	0.000 ± 0.000	0.309 ± 1.115
MNIST	LB-BNN	99.55	0.0018	0.002 ± 0.126	0.091 ± 0.380	0.093 ± 0.401
	DE	99.32	**0.0012**	0.002 ± 0.072	0.067 ± 0.320	0.070 ± 0.331
	EDL	94.42	0.2418	0.00007 ± 0.002	2.260 ± 0.054	2.260 ± 0.054
	CreINN	98.23	0.0105	0.071 ± 0.609	0.005 ± 0.043	0.077 ± 0.612
	E-CNN	99.27	0.7878	0.037 ± 0.065	1.608 ± 0.004	1.645 ± 0.064
	RS-NN	**99.71**	0.0059	0.053 ± 0.740	0.001 ± 0.016	**0.054 ± 0.741**
	SNN	98.90	0.0057	0.043 ± 0.497	0.000 ± 0.000	0.043 ± 0.497
	LB-BNN Avg	99.55	0.0018	0.016 ± 0.251	0.000 ± 0.000	**0.016 ± 0.251**
	DE Avg	99.32	**0.0012**	0.020 ± 0.198	0.000 ± 0.000	0.020 ± 0.198
	DDU	99.28	0.0028	0.028 ± 0.336	0.000 ± 0.000	0.028 ± 0.336
CIFAR-100	LB-BNN	71.34	0.1332	0.146 ± 0.504	2.348 ± 1.771	2.494 ± 1.781
	DE	**74.08**	**0.0377**	0.019 ± 0.245	3.182 ± 1.909	3.201 ± 1.906
	EDL	45.76	0.3558	0.010 ± 0.192	3.434 ± 1.843	3.445 ± 1.840
	CreINN	44.30	0.1831	0.723 ± 0.646	2.050 ± 1.188	2.774 ± 0.945
	RS-NN	71.17	0.1336	1.518 ± 3.966	0.569 ± 1.164	**2.088 ± 4.025**
	SNN	65.51	0.2357	2.293 ± 4.199	0.000 ± 0.000	2.293 ± 4.199
	LB-BNN Avg	71.34	0.1332	1.617 ± 2.886	0.000 ± 0.000	1.617 ± 2.886
	DE Avg	**74.08**	**0.0377**	1.062 ± 1.924	0.000 ± 0.000	**1.062 ± 1.924**
	DDU	73.44	0.1142	1.180 ± 2.260	0.000 ± 0.000	1.180 ± 2.260

tremely competitive, with an accuracy of 92.99% and the lowest $\mathcal{E}$ (0.407) among the epistemic models (excluding averages).

- CIFAR-100: This more complex dataset highlights the strength of RS-NN. It achieves the lowest $\mathcal{E}$ (2.088) compared to DE (3.201) and LB-BNN (2.494), suggesting that RS-NN manages the trade-off between accuracy and precision better when the number of classes is large.

- Model Behavior: Figure 6.2 reveals that EDL and E-CNN generally exhibit consistently low KL but high Non-Specificity regardless of correctness, indicating they are ”safely vague.” In contrast, DE and RS-NN show a clear distinction, being precise for correct predictions and appropriately uncertain for incorrect ones.

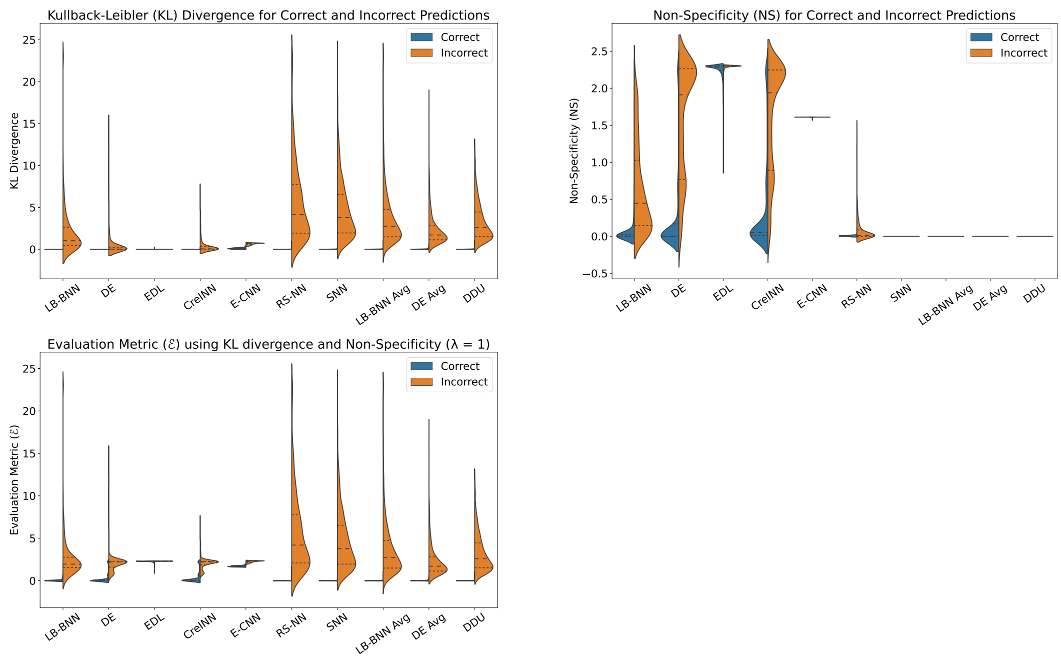


Figure 6.2: Measures of KL divergence (top left), Non-specificity (top right), Evaluation Metric (bottom left) for both Correctly (CC) and Incorrectly Classified (ICC) samples from CIFAR-10.

6.5.3 Evaluation of the Trade-off Parameter

The parameter λ controls the penalty for non-specificity. Figure 6.3 illustrates how $\mathcal{E}$ changes with λ.

- For LB-BNN, DE, and CreINN, $\mathcal{E}$ increases steadily with λ, reflecting that these models have moderate non-specificity.
- EDL and E-CNN show relatively flat or high metric values, confirming they are inherently imprecise (high NS) regardless of the penalty.

6.5.4 Model Selection

Table 6.2 provides a ranking of models based on $\mathcal{E}$ for various λ values on CIFAR-10.

This ranking reveals a crucial insight: Deep Ensembles are superior when precision is paramount (low λ), likely due to their high accuracy. However, as the penalty

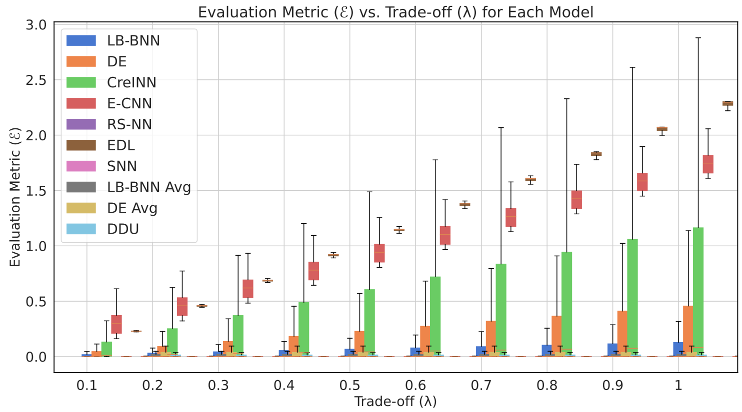


Figure 6.3: Evaluation metric vs trade-off parameter

for vagueness increases ($\lambda \rightarrow 1.0$), RS-NN moves to the top rank. This indicates that RS-NN is the most robust choice for high-stakes scenarios where both accuracy and tight uncertainty bounds are required.

6.6 Conclusion

In this chapter, we addressed the critical lack of a standardized framework for benchmarking epistemic classifiers. By proposing a Unified Evaluation Framework, we have provided a rigorous methodology to compare diverse uncertainty models—ranging from Deep Ensembles to Belief Networks—on a level playing field.

Central to this contribution is the introduction of the metric $\mathcal{E}$, which formalizes the intrinsic trade-off between *correctness* (distance to ground truth) and *informativeness* (non-specificity). This metric moves beyond simplistic accuracy scores, penalizing models that are either confidently wrong or uselessly vague.

Our extensive experiments reveal that no single model dominates across all regimes. While Deep Ensembles excel in low-penalty scenarios where raw accuracy is prioritized, our proposed Random-Set Neural Network (RS-NN) emerges as the superior choice in high-stakes environments (high λ). This validates the hypothesis that explicitly modelling epistemic uncertainty via belief functions provides a safer, more robust foundation for reliable AI than traditional probabilistic approaches.

Table 6.2: Model Rankings Based on KL and NS on the CIFAR-10 dataset for different values of trade-off λ. Model selection is based on the mean of Evaluation Metric ($\mathcal{E}$) with models with the lowest $\mathcal{E}$ ranking first.

Trade-off (λ)	**Model Ranking/Evaluation ($\mathcal{E}$) Mean**
0.1	DE, CreINN, EDL, LB-BNN, E-CNN, RS-NN [0.069, 0.117, 0.229, 0.259, 0.354, 0.399]
0.2	DE, CreINN, LB-BNN, RS-NN, EDL, E-CNN [0.108, 0.177, 0.276, 0.309, 0.456, 0.515]
0.3	DE, CreINN, LB-BNN, RS-NN, E-CNN, EDL [0.146, 0.237, 0.293, 0.309, 0.676, 0.682]
0.4	DE, CreINN, LB-BNN, RS-NN, E-CNN, EDL [0.184, 0.296, 0.309, 0.402, 0.837, 0.909]
0.5	DE, LB-BNN, CreINN, RS-NN, E-CNN, EDL [0.223, 0.326, 0.356, 0.403, 0.998, 1.136]
0.6	DE, LB-BNN, RS-NN, CreINN, E-CNN, EDL [0.261, 0.342, 0.404, 0.415, 1.159, 1.363]
0.7	DE, LB-BNN, RS-NN, CreINN, E-CNN, EDL [0.300, 0.359, 0.405, 0.475, 1.319, 1.589]
0.8	DE, LB-BNN, RS-NN, CreINN, E-CNN, EDL [0.338, 0.376, 0.405, 0.535, 1.480, 1.816]
0.9	DE, LB-BNN, RS-NN, CreINN, E-CNN, EDL [0.377, 0.392, 0.406, 0.594, 1.641, 2.043]
1.0	RS-NN, LB-BNN, DE, CreINN, E-CNN, EDL [0.407, 0.409, 0.415, 0.654, 1.802, 2.270]

Chapter 7

RANDOM-SET LARGE LANGUAGE MODELS

7.1 Introduction

In the previous chapters, we established the Random-Set Neural Network (RS-NN) for classification tasks. We now turn our attention to one of the most transformative technologies in modern AI: Large Language Models (LLMs). While LLMs produce high-quality text, their tendency to "hallucinate" and their lack of reliable uncertainty quantification remain significant barriers to trust [99].

In this chapter, I propose the Random-Set Large Language Model (RS-LLM). This approach fundamentally alters the generation process of an LLM. Instead of predicting a probability distribution over single tokens, RS-LLM predicts a *belief function* over sets of tokens. By leveraging the budgeting techniques proposed earlier and adapting them to the semantic space of language, we create a model that can explicitly quantify second-order (epistemic) uncertainty, providing a robust mechanism for hallucination detection and improved factual consistency.

7.2 Introduction and Motivation

Next-token prediction in language models is traditionally treated as a classification problem over a fixed vocabulary $\mathcal{V}$. Standard LLMs output a probability distribution P over $\mathcal{V}$ using a Softmax layer. However, this representation has a critical limitation: it cannot distinguish between *aleatoric uncertainty* (inherent ambiguity in the language) and *epistemic uncertainty* (lack of knowledge due to data limitations) [75].

Consider a simple example: an LLM is tasked to complete the sentence ''*Joe likes to play* ___''.

- **Scenario A**: The model has seen training data where Joe plays both baseball and basketball equally often.

- **Scenario B**: The model has never seen Joe before and has no information about his hobbies.

A standard probabilistic model might output $P(\text{baseball}) = 0.5, P(\text{basketball}) = 0.5$ in *both* cases. The probability distribution is identical, yet the source of uncertainty is fundamentally different.

The RS-LLM addresses this by predicting a **Belief Function** (or Random Set) rather than a probability vector.

- In Scenario A (Ambiguity), it might predict: $m(\{\text{baseball}\}) = 0.5, m(\{\text{basketball}\}) = 0.5$.

- In Scenario B (Ignorance), it might predict: $m(\{\text{baseball}, \text{basketball}\}) = 1.0$.

While both scenarios yield the same pignistic (decision) probability of 0.5 for each token, the belief function representation explicitly encodes the lack of knowledge in Scenario B via the mass assigned to the set.

In this chapter, we detail the architecture, the hierarchical budgeting strategy required to scale this to LLM vocabularies (32K tokens), and the training regimen. We demonstrate that RS-LLMs not only outperform standard LLMs in accuracy but provide superior uncertainty estimates for detecting hallucinations.

We present a scalable methodology involving:

1. A hierarchical clustering-based Budgeting strategy to handle the exponential output space of LLMs.

2. A modified Architecture and Training regimen using a specialized epistemic loss function.

3. A robust Uncertainty Estimation framework leveraging Credal Set Width to detect hallucinations.

7.3 Methodology

7.3.1 Budgeting: Constructing the Frame of Discernment

The core challenge in applying Random Set theory to LLMs is the cardinality of the vocabulary $\mathcal{V}$. For a standard tokenizer (e.g., Llama-2), $|\mathcal{V}| \approx 32,000$. The power set 2^{32000} is computationally intractable. Unlike the data-driven budgeting used in RS-NNs (Chapter 5), we propose a Semantic Budgeting strategy based on the linguistic properties of the tokens.

Hierarchical Clustering of Token Embeddings

We hypothesize that epistemic uncertainty in language modeling manifests primarily among semantically or syntactically similar tokens. Therefore, our focal sets should group these related tokens.

We employ an Agglomerative Hierarchical Clustering approach:

1. Embedding Extraction: We extract the input embedding weights $E \in \mathbb{R}^{|\mathcal{V}|\times d}$ from the pre-trained base model (e.g., Llama-2-7b).

2. Distance Computation: We compute a distance matrix using Cosine Dissimilarity (though Euclidean and Manhattan distances were also explored, see Section 7.5.5).

3. Clustering: We build a dendrogram of tokens. We cut this tree to obtain a fixed number of clusters .

The final Budget $\mathcal{O}$ comprises:

- All singleton tokens in $\mathcal{V}$ (to ensure precise generation is possible).
- The clusters derived from the hierarchy.

The total output size becomes $|\mathcal{V}| +$. For our experiments, we typically set $=$ 8000, resulting in an output dimension of $\approx 40,000$.

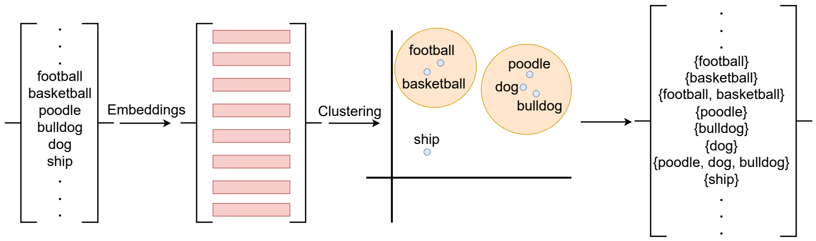


Figure 7.1: The RS-LLM Budgeting Pipeline. Token embeddings are extracted from the base model and processed via hierarchical clustering to form semantically meaningful focal sets (e.g., synonyms, morphological variants).

7.3.2 Architecture and Training

Any standard Causal Language Model (CLM) can be converted into an RS-LLM. We replace the final linear layer (the "LM Head") which maps hidden states $h \in \mathbb{R}^{dK}$ to logits $z \in \mathbb{R}^{|\mathcal{V}|}$, with a new head mapping $h \rightarrow z' \in \mathbb{R}^{|O|}$.

Crucially, we replace the Softmax activation with a Sigmoid activation. In our framework, the output logits represent the belief mass assigned to a set, not mutually exclusive probabilities. The problem becomes mathematically analogous to multi-label classification, where multiple sets (e.g., {cat} and {cat, dog}) can be "true" simultaneously.

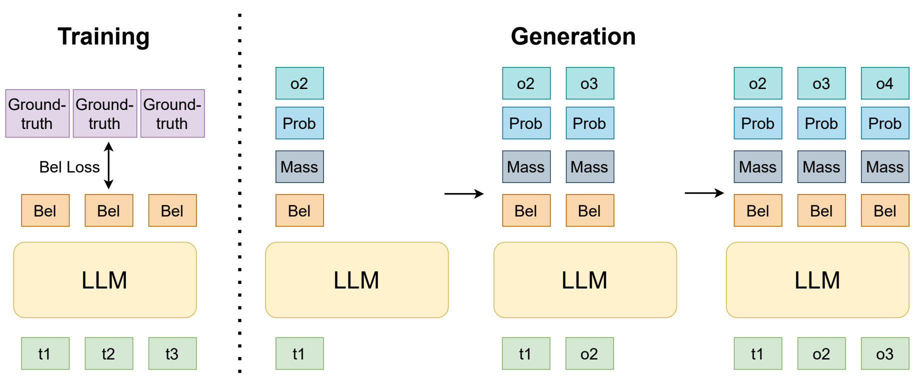


Figure 7.2: Training and Generation flow of RS-LLM. During training (teacher forcing), the model predicts belief functions for the next token. During generation, the predicted belief function is converted to a pignistic probability distribution for sampling.

Loss Function

We train the model using a compound loss function $\mathcal{L}_{RS}$. First, we encode the ground truth token y_t into a belief vector $\mathbf{bel}_t \in \{0,1\}^{|O|}$, where $\mathbf{bel}_t(A) = 1$ if $y_t \in A$, and 0 otherwise.

The primary loss is Binary Cross-Entropy (BCE):

$$\mathcal{L}_{BCE} = -\frac{1}{L}\sum_{j=1}^{L}\frac{1}{|O|}\sum_{A\in O}\left[Bel_j(A)\log(\hat{Bel}_j(A)) + (1-Bel_j(A))\log(1-\hat{Bel}_j(A))\right] \tag{7.1}$$

The Moebius inverse cannot be used here, as it requires the belief values for all the subsets of a given set to compute the corresponding mass function. This distincts it from the loss function used in [95], as this makes it more generalizable and usable with all kinds of budgetted focal sets.

$$m(A) = \hat{Bel}(A) - \sum_{B\subset A, B\neq A} m(B) \quad \text{(Recursive definition)} \tag{7.2}$$

To encourage the axioms of Dempster-Shafer theory (masses must be non-negative and sum to 1), we introduce regularization terms derived from the implicit mass assignment:

$$M_s = \max\left(0, \sum_{A\in O}\hat{m}(A) - 1\right) \quad \text{(Sum-to-one constraint)} \tag{7.3}$$

$$M_r = \sum_{A\in O}\max(0, -\hat{m}(A)) \quad \text{(Non-negativity constraint)} \tag{7.4}$$

The total loss is $\mathcal{L}_{RS} = \mathcal{L}_{BCE} + \alpha M_r + \beta M_s$. We empirically set $\alpha = \beta = 0.01$ (see Ablation Studies).

To ensure adherence to valid belief functions, regularisation terms are introduced to discourage deviations, akin to how training-time regularisation in neurosymbolic learning fosters predictions that align with commonsense reasoning [54]. Nevertheless, when α and β are too small, regularisation alone may not suffice to maintain the validity of belief-function predictions. In such instances, a corrective post-processing step is implemented: negative mass values are reset to zero, and an additional 'universal' set — comprising all classes — is incorporated into the final budget. This ensures that any residual mass is assigned to this set, preserving the requirement that the total mass across all focal sets in O sums to 1. This adjustment follows well-established approximation methods [33, 35], as detailed in [38], Part III or [36].

7.3.3 Uncertainty Estimation

RS-LLM provides two distinct measures of uncertainty for every generated token:

1. Pignistic Entropy (H_{RS}): The entropy of the pignistic probability distribution $BetP$ (the center of mass of the credal set). This captures total uncertainty (aleatoric + epistemic).

$$H_{RS} = mean\left\{-\sum_{t\in T}\left(BetP_j(t)\log BetP_j(t)\right)\right\}.K \tag{7.5}$$

2. Credal Set Width: This is a specific measure of *epistemic* uncertainty. It is defined as the difference between the upper probability and lower probability bounds for the predicted token t:

$$\overline{P}(t) = \max_{P\in \hat{Cre}} P(t), \quad \underline{P}(t) = \min_{P\in \hat{Cre}} P(t),K \tag{7.6}$$

A large width implies that the model has assigned significant mass to sets containing tK(e.g., $\{t, \text{other}\}$), indicating it cannot distinguish between them (ignorance).

7.4 Experiments

We evaluated RS-LLM against standard LLMs and other uncertainty quantification methods.

7.4.1 Experimental Setup

Models and Baselines: We utilized three base architectures: Llama2-7b [139], Mistral-7b [72], and Phi-2 [71]. We compared RS-LLM against:

- Standard LLM: The base model fine-tuned with standard Cross Entropy.
- BLoB [148]: A Bayesian Low-Rank Adaptation method.
- LoRA Ensembles [8]: An ensemble of Low-Rank Adapters.
- MC-Dropout [51]: Monte Carlo Dropout applied to LLMs.

Datasets: We employed four datasets covering different reasoning modalities:

- CoQA [121]: Conversational Question Answering (Free text).
- OBQA [101]: Open Book Question Answering (Multiple choice, STEM).
- BoolQ [28]: Boolean (Yes/No) reasoning.
- ARC-E [29]: Elementary science questions.

Training & Implementation: By Utilising the budgeting techinique highlighted in Section 7.3.1, we extract $= 8,000$ focal sets from 32,000 tokens in Llama2 and Mistral, and 51,200 tokens in Phi-2, so that the output size of the last layer for RS-Llama2 and RS-Mistral is $32,000 + 8,000 = 40,000$, and $51,200 + 8,000 = 59,200$ for Phi-2. The models are trained using the Supervised Fine-tuning (SFT) method for LLMs.

All models are trained on NVIDIA A100 80GB GPUs using Huggingface's trl framework [59] with default training parameters for 5 epochs, with a batch size of 8. To further boost the training, we add LoRA adapters [154] of rank 64 to all blocks of the model. Furthermore, we load and train the model in 4-bit mode for enhanced efficiency. Similary setting is used for BLoB and LoRA Ensembles with number of samples and number of ensemble set to 2 for them respectively. For RS-Llama2, we set $\alpha = \beta = 1e-2$ as hyperparameter values in the loss function.

Figure 7.3 highlights the training prompt template. The blue text represents the instructions to the LLM while black represents the actual question. The model is trained to predict the text in green. At generation time, the model is given input in the same template but without the answer. The model then continues the statement and produces the answer.

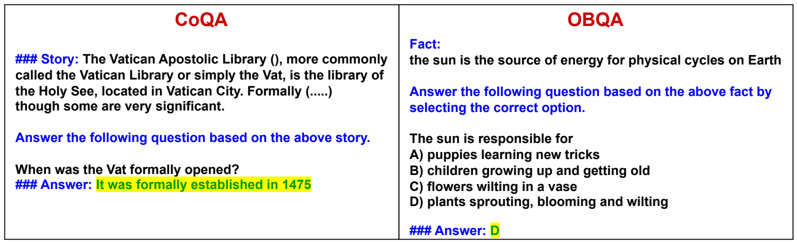


Figure 7.3: Training examples from CoQA and OBQA datasets. The text in black highlights the actual question, while the blue text represents prompt instructions. The model is trained to predict the text in green.

7.4.2 Generative Performance

We evaluate all model architecures on CoQA and OBQA datasets to higlight the performance improvments over standard LLMs. For CoQA, we produce free text. To evaluate the closeness of generated and ground truth text, we employ the cosine similarity metric between the two [9]. Cosine Similarity is a metric used to determine the cosine of the angle between two non-zero vectors in a multi-dimensional space. It is used in CoQA to measure semantic closeness between question and context (or candidate answers) text embeddings. Whereas, for OBQA, the option label is expected and generated which allows us to conveniently measure the accuracy w.r.t. the ground truth. Table 7.1 reports the cosine similarity and accuracy on the CoQA and OBQA datasets, respectively, for both models. RS-LLMs clearly outperform the standard LLMs model on both datasets across all models, even though all models are trained using the exact same regimen. This clearly shows the representative prowess of random sets in the output space.

7.4.3 Comparison with Uncertainty Baselines

We also compare our approach with other uncertainty methods; namely BLoB [148] and MC-Dropouts,bayesian based methods and LoRA Ensembles [8], an esemble

Table 7.1: Performance comparison on CoQA (Cosine Similarity) and OBQA (Accuracy). RS-LLMs consistently outperform standard baselines, indicating that the belief-function representation captures richer semantic information.

Model	*CoQA (Cosine Sim.)*	*OBQA (Accuracy)*
Llama2	0.69	83.20
RS-Llama2	**0.71**	**89.60**
Mistral	0.67	91.60
RS-Mistral	**0.72**	**93.00**
Phi2	0.72	87.60
RS-Phi2	**0.73**	**91.80**

based method on ARC-E [29], BoolQ [28] and OBQA [101] datasets. All methods are evaluated using Llama-2 models. Table 7.2 shows the quantitative results. RS-Llama2 significantly outperforms all the other methods on all datasets in terms of accuracy, clearly showcasing its superiority.

Table 7.2: Accuracy comparison on ARC-E, BoolQ, and OBQA. RS-Llama2 surpasses both Bayesian and Ensemble approaches.

	ARC-E	BoolQ	OBQA
Llama2	83.89	87.51	83.20
BLoB	84.78	87.23	85.12
LoRA Ensembles	84.31	87.09	81.38
MC-Dropout	85.02	87.32	82.51
RS-Llama2	**86.12**	**88.56**	**89.60**

7.4.4 Uncertainty quantification

. In Figure 7.4, instead, we show how the uncertainty measures produced by the two models behave with respect to the correctness of the answer. It illustrates the entropy distributions of correct vs. incorrect predictions for all models on the OBQA dataset. Ideally, these distributions should be clearly separable. Both BLoB and RS-Llama2 perform well in this regard, with RS-Llama2 further distinguishing itself by exhibiting a sharp spike at zero in the correct prediction distribution.

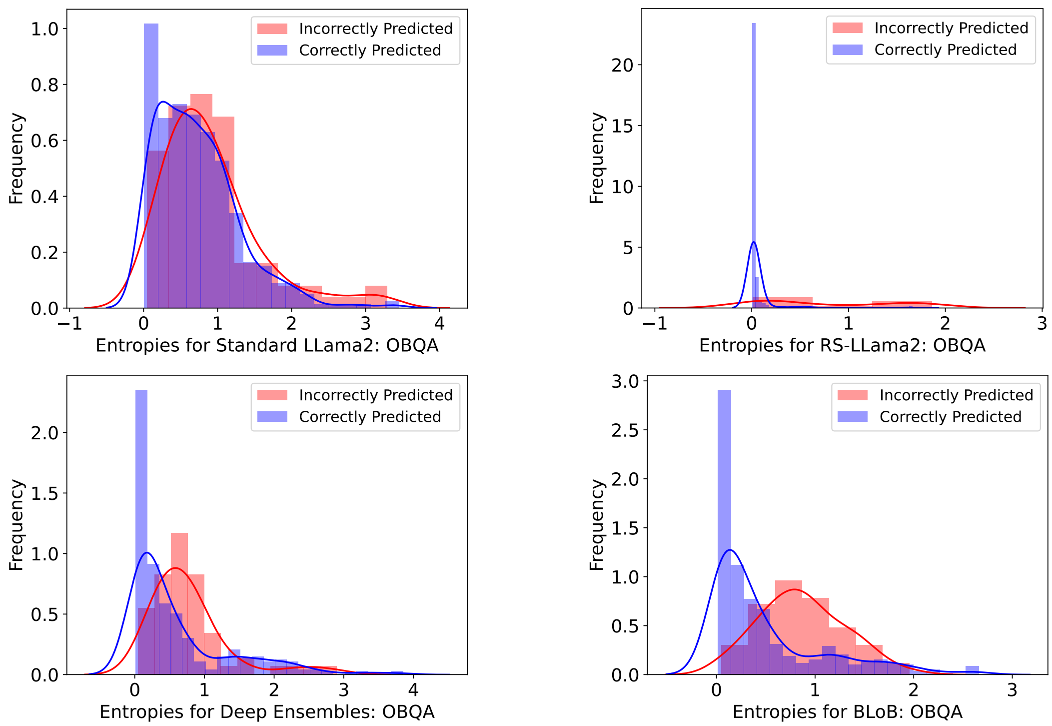


Figure 7.4: Behavior of entropy measures for all models with respect to the correctness on OBQA dataset.

7.4.5 *Hallucination Detection*

A primary goal of these models is detecting hallucinations. We simulated hallucinations by providing the model with incorrect context (e.g., irrelevant paragraphs for a reading comprehension task) and measuring the resulting uncertainty. A good model should exhibit high uncertainty when the context is wrong. Tabs. 7.3 & 7.4 presents the uncertainty evaluation and AUROC of all uncertainty methods under correct and incorrect context on BoolQ and OBQA datasets respectively. As evident, RS-Llama2 exhibits significantly superior performance on BoolQ while also showing comparable performance on the OBQA dataset.

We further evaluate the hallucination detection on CoQA dataset using entropy measures. Table 7.5 presents the uncertainty evaluation and AUROC for Llama2 and RS-Llama2 under both correct and incorrect context conditions. Note that the results for BLoB and LoRA Ensembles are not available for CoQA, as their imple-

Table 7.3: Hallucination detection performance (AUROC) on BoolQ using Entropy.

Model	**CC (↓)**	**IC (↑)**	**AUROC**
LLama2	0.50 ± 0.52	0.97 ± 0.64	78.43
BLoB	0.33 ± 0.26	0.45 ± 0.30	56.34
LoRA Ensembles	0.42 ± 0.36	0.60 ± 0.44	61.62
RS-LLama2	0.13 ± 0.22	0.62 ± 0.36	**86.34**

Table 7.4: Hallucination detection performance (AUROC) on OBQA using Entropy.

Model	**CC (↓)**	**IC (↑)**	**AUROC**
LLama2	0.75 ± 0.59	1.10 ± 0.93	66.45
BLoB	0.48 ± 0.53	1.46 ± 0.47	**92.45**
LoRA Ensembles	0.62 ± 0.72	1.12 ± 0.91	63.21
RS-LLama2	0.14 ± 0.38	0.61 ± 0.68	90.23

mentations do not currently support free-text generation tasks, limiting their applicability to this particular dataset. As shown in the results, RS-Llama2 demonstrates superior performance in detecting hallucinations on the CoQA dataset, outperforming Llama2 and further validating its effectiveness in handling hallucinations across different types of tasks.

Table 7.5: Uncertainty evaluation using entropy for Llama2 and RS-Llama2 for CoQA on correct and incorrect context.

Model	CoQA		
	CC (↓)	IC (↑)	AUROC
LLama2	0.39 ± 0.45	0.90 ± 0.80	70.56
RS-LLama2	2.39 ± 1.42	4.70 ± 3.25	**72.14**

Lastly, we evaluate the hallucination detection for RS-Llama2 using credal set wdith. The experiment setting is same as for previously. Only the metric for uncertainty measurement is changed. Table 7.6 presents the uncertainty evaluation and AUROC for RS-Llama uncertainty methods under both correct and incorrect context conditions. As shown in the results, RS-Llama2 effectively detects hallucinations under credal set width too. This reinforces the model's robustness in uncertainty estimation, making it a reliable choice for real-world applications where context errors may occur.

Table 7.6: Uncertainty evaluation using credal set width for RS-Llama2 on correct and incorrect context.

Dataset	CC (↓)	IC (↑)	AUROC
CoQA	0.13 ± 0.13	0.28 ± 0.21	69.24
OBQA	0.01 ± 0.05	0.10 ± 0.09	81.90
BoolQ	0.00 ± 0.01	0.01 ± 0.02	67.23

7.4.6 Align Score

In addition to using cosine similarity for the evaluation on the CoQA dataset, we also introduce another evaluation metric known as Align Score [155]. The Align Score is a measure of factual consistency between the generated and predicted text. While traditionally used for measuring alignment in machine translation tasks, we adapt it for the Question Answering (QA) setting. Specifically, in our case, we compute the Align Score between the ground truth and the generated answer, providing a more meaningful and relevant assessment for QA models. This adaptation allows us to assess how consistent the model's generated answers are with the true answers, focusing on factual accuracy rather than just semantic similarity.

Table 7.7 presents the Align Score between the ground truth and the generated answers for different models on the CoQA dataset. From the results, it is evident that RS-LLMs (Random-Set LLMs) consistently outperform standard LLMs across all tested models. This demonstrates the superiority of RS-LLMs in generating more factually consistent and reliable answers, highlighting their effectiveness in tasks where factual accuracy is critical.

Table 7.7: AlignScore on CoQA Dataset for Llama2 and RS-Llama2

Model	Llama2	RS-Llama2
Llama2	0.44	**0.57**
Mistral	0.46	**0.52**
Phi2	0.51	**0.56**

7.4.7 Ablation on Hyperparameters α & βK

The terms $M_{sK}$and M_r, as defined in Equation 7.3 are included in the loss function to enforce the validity of the belief functions generated by the model. Specifically, these terms encourage two key properties: the sum of the masses must be equal to 1, and all masses must remain non-negative. These constraints ensure that the model outputs consistent and valid belief functions. However, it is crucial to strike an appropriate balance in determining the weight of these terms within the overall loss function. Excessively penalizing deviations from these constraints can lead to undesirable consequences, as it may impair the model's ability to make accurate predictions.

This issue is similar to the problem faced in Variational Autoencoders (VAEs), where an overemphasis on the KL divergence term in the loss function can cause the model to focus too heavily on matching the latent distribution, potentially at the expense of reconstructing the input data faithfully. As a result, while the latent space may be well-regularized, the quality of the data reconstruction may suffer, leading to subpar performance.

To investigate the optimal balance between constraint enforcement and predictive performance, we conducted experiments with different values for the hyperparameters $\alpha K$and β. These hyperparameters control the relative importance of the mass-related terms in the loss function. Table 7.8 presents the cosine similarities for RS-Llama2 with varying values of $\alpha K$and β. Our findings indicate that values of $\alpha = \beta = 0.01$ and $\alpha = \beta = 0.0001$ yield similar performance in terms of cosine similarity. However, to further encourage the generation of valid belief functions, we opt for $\alpha = \beta = 0.01$ in our experiments, as this choice strikes a better balance between model accuracy and constraint satisfaction.

Table 7.8: Cosine Similarity values for different $\alpha = \beta$ settings

$\alpha = \beta K$	0.1	0.01	0.001	0.0001
Cosine Similarity	0.66	**0.71**	0.69	**0.71**

7.4.8 Ablation on Number of Focal Sets

The number of non-singleton focal sets, denoted as , to be budgeted is a critical hyperparameter that can influence model performance. It is important to explore the impact of different values of , as it can affect the balance between model expressiveness and computational complexity. Specifically, a lower value of may lead to results that closely resemble those of traditional large language models (LLMs), as it restricts the number of focal sets the model generates, thereby simplifying its task. On the other hand, a higher value of introduces greater complexity, as the model is tasked with managing a larger number of focal sets, potentially increasing the diversity and richness of its predictions, but at the cost of computational efficiency and potential overfitting.

To better understand the effect of on model performance, we conducted an ablation study on the CoQA dataset, as shown in Table 7.9. In this study, we evaluated the model's performance across various values of , ranging from smaller budgets (2,000 focal sets) to larger ones (16,000 focal sets). We also included a "Combined" budget, which represents the union of the budgets for 2,000, 4,000, 8,000, and 16,000 focal sets. This approach allows us to observe how combining different budgets impacts the model's ability to generalize and maintain performance.

Our findings indicate that using a medium value of —neither too small nor too large—yields the most optimal results. This balance ensures that the model has enough focal sets to generate meaningful predictions while avoiding excessive complexity that could hinder its ability to generalize effectively. These experiments were conducted using Llama2 with hyperparameters $\alpha = \beta = 0.01$, and the results suggest that tuning is essential for achieving the best performance.

Table 7.9: Cosine Similarity values across different budget sizes

Budget Size	2000	4000	8000	16000	Combined (24,538)
Cosine Similarity	0.67	0.68	**0.71**	0.66	0.69

7.5 Analysis of Budgeted Focal Sets

In this section, I present a comprehensive analysis of the budgeted focal sets generated using our novel budgeting approach. To conduct this evaluation, I employ the Llama2-7b-hf model, which supports a substantial token context size of 32,000. For the budgeting parameter, we set = 8000, meaning that from the enormous theoretical space of 2^{32000} possible focal sets, we constrain and select a manageable subset of 8,000 focal sets to analyze.

7.5.1 Qualitative Assessment of Focal Sets

The resulting focal sets exhibit strong semantic coherence, demonstrating that the budgeting technique effectively clusters related tokens. For instance, several focal sets group together semantically or contextually related words, such as:

- {`cattle`, `sheep`} — grouping related livestock terms.
- {`shame`, `pity`} — expressing related emotional states.
- {`quiet`, `calm`, `quietly`} — capturing synonyms and related adverbs describing tranquility.
- {`maintain`, `retain`, `maintained`, `retained`} — grouping verb forms and their conjugations.
- {`delight`, `pleasure`, `pleased`, `proud`, `pride`} — combining expressions of positive emotions and states.

These examples highlight the model's ability to capture both lexical similarity and subtle semantic relationships within the focal sets.

7.5.2 Quantitative Semantic Analysis: Centroid Distance

To quantify the semantic coherence of these focal sets, we perform a detailed semantic analysis by calculating the centroid distance for each set. Specifically, we use sentence-transformers [67] to obtain vector embeddings for each token within a focal set. The centroid distance is then computed as the mean Euclidean distance between each token's embedding and the centroid (mean vector) of the entire set.

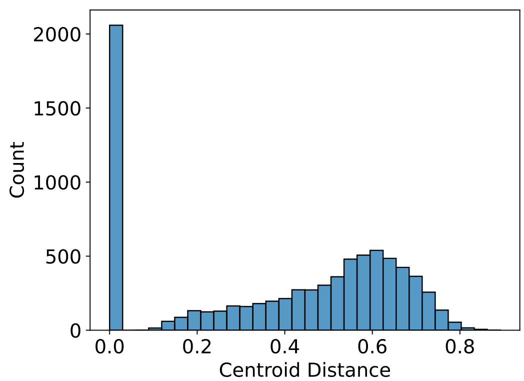
Figure 7.5a illustrates the distribution of these centroid distances across all 8,000 budgeted focal sets. The majority of the focal sets exhibit very small centroid distances, clustering tightly around zero, indicating high semantic similarity among the tokens in each set. Only a small fraction of sets show larger distances, which may correspond to more diverse or loosely connected token groups. This empirical evidence strongly supports the effectiveness of our budgeting method in generating semantically meaningful and coherent focal sets.

7.5.3 Distribution of Focal Set Sizes

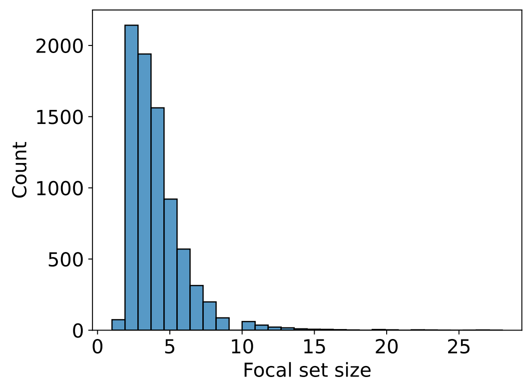
Further insights are gained by examining the distribution of focal set sizes, as depicted in Figure 7.5b. The data shows that most focal sets are relatively small, with sizes of 2, 3, and 4 being the most prevalent. Specifically, we observe 2,142 sets of size 2, 1,940 sets of size 3, and 1,562 sets of size 4, making these the dominant group sizes.

The largest focal set identified contains 162 tokens. Closer inspection reveals that this set predominantly consists of punctuation marks and symbols such as periods (.), exclamation marks (!), hashes (#), dollar signs ($), and percent signs (%). The formation of such a large set comprising non-lexical tokens is intuitive, as these symbols often co-occur or share syntactic roles, making them a natural cluster in the budgeting process.

Figure 7.5: (a) Frequency distribution of centroid distances of the obtained budgeted focal sets. (b) Frequency distribution of sizes of obtained budgeted focal sets. Note: There are 8 focal sets with sizes > 30. They are excluded here for better visualisation.



(a)



(b)

7.5.4 Morphology and synonymity analysis

Furthermore, we conduct a detailed semantic analysis of the obtained focal sets to understand the nature of the token groupings. Specifically, we investigate whether the sets primarily consist of tokens sharing the same root word—indicating morphological similarity—or if they instead contain different words that are semantically related, such as synonyms.

To quantitatively evaluate morphological similarity, we measure the length of the longest common substring among tokens within each set. Our analysis reveals that out of the 8,000 budgeted focal sets, 4,508 sets have a longest common substring length of two characters or fewer, suggesting that these sets mostly consist of tokens with little or no shared root structure. In contrast, only 721 sets have a longest common substring length of six characters or more, indicating a smaller subset of sets with strong morphological ties.

In addition to morphological analysis, we assess synonymy within the focal sets. Using a lexical resource to identify synonyms, we find that 2,703 sets contain at least one pair of synonymous tokens. This indicates that a substantial portion of the focal sets cluster semantically similar but lexically distinct words.

Taken together, these findings demonstrate that our budgeting technique produces a diverse collection of focal sets: some clusters are formed by morphologically related variants of the same root word, while others group distinct but semantically similar words. This diversity enhances the generality and flexibility of our budgeted sets, allowing them to capture a broader range of linguistic relationships rather than being restricted to one type of semantic similarity.

7.5.5 *Impact of Distance Metrics on Performance*

Finally, we investigate the performance of various distance metrics in hierarchical clustering. Specifically, we evaluate the effectiveness of different distance metrics for RS-Llama2 embeddings on the CoQA dataset. Table 7.10 presents the cosine similarities obtained using several distance measures for hierarchical clustering. Our analysis reveals that all the metrics exhibit comparable performance, with only minor variations observed across the different methods. Notably, Euclidean distance and Cosine similarity perform slightly better than Manhattan distance, likely due to their higher expressiveness in capturing the geometric relationships within the embedding space.

In addition to evaluating individual metrics, we also explore the performance when combining the distance metrics into a richer, multi-metric budget. The results show that it remains competitive, performing similarly to when as compared to stand-alone metric budgets. However, a key advantage of using a combined budget approach is that it yields better uncertainty estimates. This enhanced uncertainty estimation can allow for more accurate detection of hallucinations (see Table 7.11).

Table 7.10: Cosine Similarity values for different distance metrics in hierarchical clustering

Distance Metric	Cosine Similarity
Euclidean Distance	**0.71**
Manhattan Distance	0.70
Cosine Similarity	**0.71**
Combined Budget (16,354)	0.70

Table 7.11: Uncertainty evaluation using pignistic entropy of budgeting with a stand-alone metric and with combined metrics.

	CoQA		
	CC (↓)	IC (↑)	AUROC
Stand-alone Metric (Euclidean Distance)	2.39 ± 1.42	4.70 ± 3.25	72.14
Combined Metrics	2.62 ± 1.74	6.34 ± 3.87	**80.09**

7.6 Conclusion

This chapter introduced the Random-Set Large Language Model (RS-LLM), a transformative approach to uncertainty quantification in generative AI. By redefining the output space of an LLM as a belief function over sets of tokens, we moved beyond the limitations of standard point-estimate probabilities.

Our key innovation was the successful adaptation of the budgeting strategy to the semantic space of language. By clustering token embeddings, we reduced the exponential output space to a manageable set of semantically meaningful focal sets (e.g., synonyms, related concepts), making the approach computationally feasible for large vocabularies.

The experimental results validate this methodology across multiple dimensions:

1. Generative Performance: RS-LLMs consistently outperformed standard baselines in generation quality (Cosine Similarity) and factual consistency (AlignScore) across diverse datasets like CoQA and OBQA.

2. Hallucination Detection: The unique metrics derived from belief functions (Pignistic Entropy and Credal Set Width) proved highly effective at distinguishing correct from incorrect contexts, significantly outperforming Bayesian and Ensemble methods on benchmarks like BoolQ.

3. Interpretability: Our analysis of the focal sets confirmed that the model learns to group tokens based on both morphological and semantic similarity, providing a transparent view into the model's confusion.

These findings demonstrate that RS-LLMs offer a robust, scalable, and theoretically grounded path toward building trustworthy language models that can reliably signal their own ignorance.

Chapter 8

EPISTEMIC DIFFUSION MODELS

8.1 Introduction

While the previous chapters explored uncertainty in discriminative models and Large Language Models, this chapter turns to the frontier of visual generative modelling: Diffusion Models.

Diffusion Models [133] have recently established themselves as the state-of-the-art in image generation, capable of synthesizing samples of unprecedented quality [122]. Despite their success, standard diffusion models operate under a deterministic or rigidly scheduled variance assumption. This limits their ability to model the true underlying variability of the data generation process, often leading to a lack of diversity in the generated samples.

In this chapter, we propose Epistemic Diffusion Models, a ”two-level” diffusion process. Instead of predicting a single Gaussian noise distribution at each step, our model learns a second-order distribution (specifically, a Normal-Inverse-Gamma distribution) over the parameters of the Gaussian noise. This allows the model to capture epistemic uncertainty regarding the diffusion process itself, aiming to improve sample diversity and generation fidelity.

8.2 Background: Denoising Diffusion Probabilistic Models

Although originally proposed in 2015 [133], it was not until 2020 [64] that the research community began to fully exploit the potential of diffusion models. Broadly, these models can be categorized into three types:

1. Denoising Diffusion Probabilistic Models (DDPMs) [133, 64], inspired by non-equilibrium thermodynamics.

2. Noise-Conditioned Score Networks (NCSNs) [135], which learn a score function (gradient of the log-density) at various noise levels.

3. Stochastic Differential Equations (SDEs) [136], which generalize the process to continuous time.

This thesis focuses primarily on the first category: DDPMs. These models consist of two opposing processes: a fixed *Forward Diffusion Process* that destroys data, and a learned *Reverse Denoising Process* that creates data.

8.2.1 The Forward Diffusion Process

The forward process incrementally corrupts a real data sample $\mathbf{x}_0$ by adding Gaussian noise over TKsteps, resulting in a sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The magnitude of the noise added at each step is controlled by a pre-defined variance schedule $\{\beta_{tK} \in (0,1)\}_{t=1}^{T}$:

$$q(\mathbf{x}_{tK} | \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \quad q(\mathbf{x}_{1:TK} | \mathbf{x}_0) = \prod_{t=1}^{TK} q(\mathbf{x}_{tK} | \mathbf{x}_{t-1}) .K \tag{8.1}$$

As tKincreases, the original features of $\mathbf{x}_0$ are gradually obliterated. As $TK \to \infty$, the distribution of $\mathbf{x}_{TK}$converges to an isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

A critical property of this Gaussian Markov chain is that we can sample $\mathbf{x}_{tK}$at any arbitrary timestep t directly from $\mathbf{x}_0$ without iterating through intermediate steps. This is achieved via the ‘‘reparameterization trick’’. Let $\alpha_{tK} = 1 - \beta_{tK}$and $\bar{\alpha}_{tK} = \prod_{i=1}^{t} \alpha_i$. The derivation is as follows:

$$\begin{aligned}
\mathbf{x}_{tK} &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} && ;\text{ where } \boldsymbol{\epsilon}_{t-1}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
&= \sqrt{\alpha_t \alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} && ;\text{ merging two Gaussians} \\
&= \cdots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \\
q(\mathbf{x}_{tK} | \mathbf{x}_0) &= \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)
\end{aligned} \tag{8.2}$$

Typically, the schedule is defined such that $\beta_1 < \beta_2 < \cdots < \beta_T$, meaning we take larger corruption steps as the signal becomes noisier.

8.2.2 The Reverse Denoising Process

To generate data, we must reverse this process by sampling from the posterior $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. If the noise steps β_t are sufficiently small, this reverse distribution is also Gaussian. However, computing it exactly requires knowledge of the entire dataset. Therefore, we approximate it using a learnable model p_θ:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t); \quad p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\right).$$

To train this model, we utilize the fact that the reverse posterior *is* tractable when conditioned on the ground truth $\mathbf{x}_0$:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}\right).$$

By applying Bayes' rule, we can derive the closed-form expressions for the mean $\tilde{\boldsymbol{\mu}}$ and variance $\tilde{\beta}_t$:

$$\begin{aligned} q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) &= q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_t \mid \mathbf{x}_0)} \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1-\bar{\alpha}_t}\right)\right) \end{aligned}$$

Through algebraic manipulation (completing the square for $\mathbf{x}_{t-1}$), we identify the parameters of this distribution:

$$\begin{aligned} \tilde{\beta}_t &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \\ \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0. \end{aligned}$$

Using the reparameterization from Eq. (8.2) to express $\mathbf{x}_0$ in terms of $\mathbf{x}_t$ and ϵ_t, the mean becomes:

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right).$$

Optimization Objective

We optimize the model by minimizing the Variational Lower Bound (VLB) on the negative log-likelihood:

$$L_{\text{VLB}} = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log \frac{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right].$$

This objective can be decomposed into a sum of KL-divergence terms at each timestep:

$$L_{\text{VLB}} = \mathbb{E}_q[\underbrace{D_{\text{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^{T} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}_{L_0}]$$

This formulation allows us to train the neural network to predict the noise $\boldsymbol{\epsilon}_t$ (or equivalently the mean $\boldsymbol{\mu}_\theta$) at each step.

8.3 Problem Statement and Motivation

Diffusion models represent a significant leap forward from Generative Adversarial Networks (GANs), solving key issues such as training instability and mode collapse [69]. However, they introduce new challenges. The most prominent is the slow inference time [134], necessitating iterative sampling over thousands of steps.

While the community has aggressively optimized sampling speed—via distillation [125], flow matching [88], and latent space modelling [122, 141]—a fundamental issue remains under-addressed: Diversity.

Diffusion models are often praised for having better coverage than GANs [14]. However, quantitative analyses using the Vendi Score (VS) indicate that diffusion samples are actually *less* diverse than the training datasets (e.g., ImageNet, CIFAR-10) they mimic [49]. As illustrated in Figure 8.1, popular models like Stable Diffusion [122] often produce outputs that are startlingly similar in composition, pose, and style, failing to capture the full richness of the semantic prompt.

Furthermore, standard DDPMs [64] typically fix the variance β_t rather than learning it. Even "improved" versions [111] or implicit models [134] constrain the

Figure 8.1: Diversity analysis on images generated by the Stable Diffusion Model [122]. Top: ”an oil painting of a horse sitting in an airplane” yields samples with nearly identical composition. Middle: ”a picture of a girl eating dinner” produces subjects with similar demographics. Bottom: replacing ”girl” with ”female” results in identical poses, highlighting a lack of generative diversity.

variance to a deterministic schedule or a narrow learned range. This rigid treatment of stochasticity limits the model's ”creativity.” We argue that the randomness should be a controllable property of the model, not a fixed property of the process.

8.4 Epistemic Diffusion Models

To solve the diversity problem and introduce true uncertainty awareness, we propose Epistemic Diffusion Models.

In a standard diffusion model, at each reverse step, the model predicts the mean $\boldsymbol{\mu}_{\theta K}$of a single Gaussian distribution (assuming fixed variance). We propose a two-level approach. Instead of predicting the parameters of the Gaussian directly, the model learns a second-order probability distribution over the space of possible Gaussian noise parameters.

At each generation step, we first sample a specific mean and variance from this second-order distribution, and then sample the image transition from the resulting

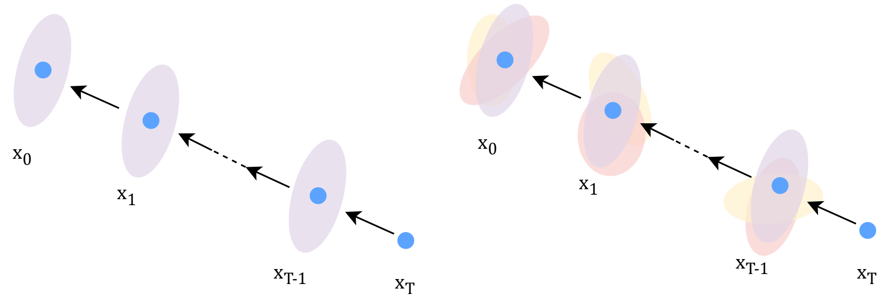


Figure 8.2: Left: Standard diffusion generation using a fixed Gaussian. Right: Epistemic diffusion generation. At each step, we sample a Gaussian from a learned second-order distribution (depicted by the shaded region representing the variance of the cloud of Gaussians), enhancing diversity.

Gaussian.

This approach offers several key benefits:

1. Enhanced Diversity: By sampling from a distribution of distributions, the model can access a broader range of generation pathways (see Figure 8.2).

2. Better Data Modelling: A single Gaussian is unimodal. A distribution over Gaussians can model complex, multi-modal data distributions more flexibly (see Figure 8.3).

3. Uncertainty Awareness: The variance of the predicted second-order distribution provides a measure of epistemic uncertainty—how unsure the model is about the denoising step.

8.4.1 Methodology: The Normal-Inverse-Gamma Prior

To make this two-level inference tractable, we need a suitable distribution to model the uncertainty over the Gaussian parameters (mean μKand variance σ^2). We select the Normal-Inverse-Gamma (NIG) distribution [2]. The NIG distribution is the conjugate prior for a normal distribution with unknown mean and variance, which allows for closed-form Bayesian updates without expensive Monte Carlo simulations.

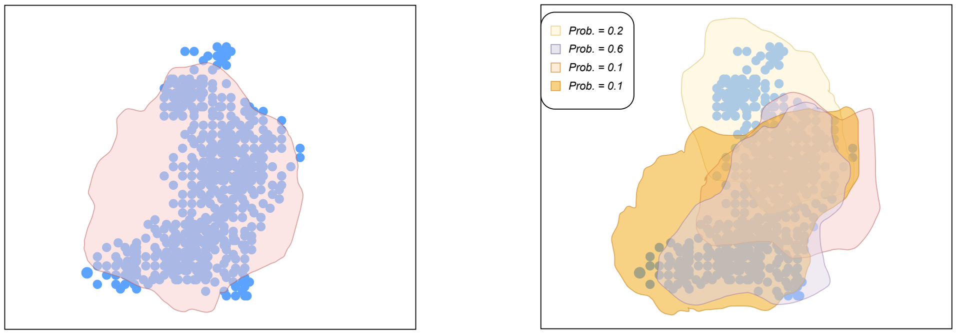


Figure 8.3: Left: Standard diffusion models capture data variability using a single Gaussian. Right: Epistemic diffusion models model this variability as a second-order distribution over possible Gaussian PDFs, allowing for more flexible data modelling.

Under this framework, the noise parameters (μ, σ^2) are distributed as follows:

$$(\mu, \sigma^2) \sim \text{N-}\Gamma^{-1}(\mu', \lambda, \alpha, \beta).K \tag{8.3}$$

Specifically, the variance σ^2 follows an Inverse-Gamma distribution:

$$\sigma^2 \mid \alpha, \beta \sim \Gamma^{-1}(\alpha, \beta)$$

And conditioned on σ^2, the mean μ follows a Normal distribution:

$$\mu \mid \sigma^2, \mu', \lambda \sim \mathcal{N}(\mu', \sigma^2/\lambda) \tag{8.4}$$

Here, the model must predict four parameters: μ' and λ (controlling the mean), and α and β (controlling the variance).

8.4.2 Loss Function

To train the epistemic diffusion model, we design a composite loss function comprising three terms:

1. Variance MSE: The error between the true variance of the noise schedule and the predicted variance sampled from the Inverse-Gamma distribution.
2. Mean MSE: The error between the true mean and the predicted mean. The predicted mean is obtained via sampling: first sample σ_{pred}, then compute $\mu_{predK} = \mu' + \sigma_{predK} \lambda K^{-1} \cdot \epsilon$.

3. KL Divergence: We enforce consistency by minimizing the KL divergence between the Normal distribution defined by the true parameters $(\mu_{true}, \sigma_{true})$ and the distribution defined by the sampled parameters $(\mu_{pred}, \sigma_{pred})$.

The KL divergence between two Gaussians p and q is:

$$L(p,q) = \log \frac{\sigma_{predK}}{\sigma_{trueK}} + \frac{\sigma_{true}^2 + (\mu_{true} - \mu_{pred})^2}{2\sigma_{predK}^2} - 0.5$$

The total loss is a weighted sum:

$$\mathcal{L} = w_1 \cdot \text{MSE}(\sigma_{true}, \sigma_{pred}) + w_2 \cdot \text{MSE}(\mu_{true}, \mu_{pred}) + w_3 \cdot \text{KLD} \tag{8.5}$$

8.5 Bayesian Formulation and Challenges

Implementing this fully Bayesian diffusion process is non-trivial. We propose a hybrid approach where the mean parameters $(\mu', \lambda K^1)$ are predicted per-pixel by the U-Net, while the variance parameters $(\alpha_{ig}, \beta_{ig})$ are learned as global properties of the model (or per timestep) independent of the specific input image.

A major challenge is establishing the priors. Since $\alpha_{igK}$and β_{igK}must be initialized, an inappropriate prior can prevent convergence. We explored two strategies: 1. Uninformative Priors: Minimizing the impact on the posterior. 2. Learned Priors: Estimating the prior parameters from a set of randomly sampled, monotonically increasing variance schedules using Maximum Likelihood Estimation [90].

Updating these parameters involves calculating a posterior, for which we explored Variational Inference [91] and Monte-Carlo sampling.

8.6 Evaluation metrics

To rigorously evaluate both the fidelity and the diversity of our outputs, we employ two complementary metrics: the Fréchet Inception Distance (FID) [62] for image quality and the Vendi Score [50] for sample diversity.

Fréchet Inception Distance (FID): The Fréchet Inception Distance has become a standard metric for assessing the visual realism of images synthesized by generative models. FID operates by embedding both real and generated images into the feature space of a pretrained Inception network and then modeling each set of activations as a multivariate Gaussian distribution. Lower FID values indicate that the generated distribution more closely matches the real data distribution in feature space, reflecting higher sample fidelity and fewer visual artifacts.

Vendi Score: While FID effectively captures sample quality, it does not directly quantify how well a model covers the full support of the target distribution. To address this gap, we utilize the Vendi Score, a metric explicitly designed to measure diversity among generated samples. It is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix. Higher Vendi Scores thus reflect richer diversity, complementing FID's assessment of image fidelity.

8.7 Experimental Results

We conducted initial experiments on the CIFAR-10 dataset using a U-Net architecture with $TK= 1000$ timesteps.

8.7.1 Fixed Variance Baseline

In our first set of experiments, we fixed the variance schedule to the standard DDPM trajectory and only learned the distribution over the mean (i.e., predicting μ' and λ). As shown in Figure 8.4, the model successfully generated high-quality images resembling CIFAR-10. However, the model learned a very narrow distribution (high certainty) for the means, behaving almost identically to a standard deterministic diffusion model. Consequently, the diversity of the output was not significantly improved.

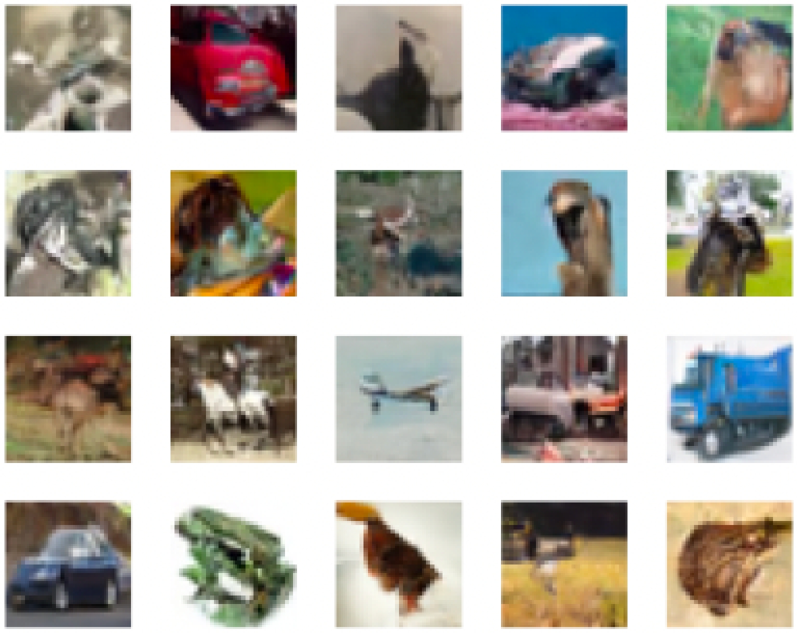


Figure 8.4: Images generated by an Epistemic Diffusion Model trained on CIFAR-10. In this experiment, the variance was fixed, resulting in high-quality but standard diversity outputs.

8.7.2 Full Bayesian Learning

We then attempted the full formulation: learning the Inverse-Gamma distribution over the variance alongside the mean. We tested both uninformative priors and priors learned from variance schedules. Unfortunately, these experiments resulted in a degradation of image quality. Our analysis suggests that the unconstrained learning of the variance gave the model ”too much freedom.” Without sufficient constraints, the model struggled to converge to the delicate noise schedule required for stable diffusion. Another set of experiments was to parameterize the variance schedule and learn the distribution over the schedule. This led to better results but it still lacked in overall qaulity.

	CelebA		Cifar-10	
	FID ↓	Vendi Score ↑	FID↓	Vendi Score ↑
Reference	-	6.95	-	8.48
Standard Diffusion	16.2	5.73	22.1	4.37
Fixed Variance	15.6	5.48	21.8	4.11
Full Bayseian	27.8	6.75	37.1	5.87
Parameterized Schedular	23.3	6.21	34.9	5.01
Image Manifold	17.2	5.97	22.9	4.58

Table 8.1: Performance of Standard GAN and different modes of epistemic diffusion model on Celeb-A and Cifar-10 dataset. Reference represents the Vendi score (diversity) of the training data.

8.7.3 Manifold-Aware Diffusion

The failure of unconstrained variance learning points the way forward: we must constrain the training process to the geometry of the data. Images reside on a low-dimensional manifold within the high-dimensional pixel space. Standard diffusion models implicitly learn to point towards this manifold.

To stabilize Epistemic Diffusion, we propose to explicitly incorporate Manifold Learning. By forcing the diffusion process to respect the approximate image manifold, we can constrain the variance estimation to ”tangential” and ”normal” components relative to the data surface [149].

In this direction, projection-based Diffusion was primarily investigated. This entails projecting the noisy samples back onto the estimated manifold after each sampling step to prevent drifting into low-probability regions. This was achived using an autoencoder’s decode–encode cycle. As shown in Figure 8.5, this displayed considerable promise and much better results.

By constraining the epistemic uncertainty to the geometry of the data, we aim to achieve the dual goals of high-fidelity generation and enhanced sample diversity. Table 8.1 presents the qualitative results accross different experimental settings on cifar-10 [78] and CelebA [89] datasets.

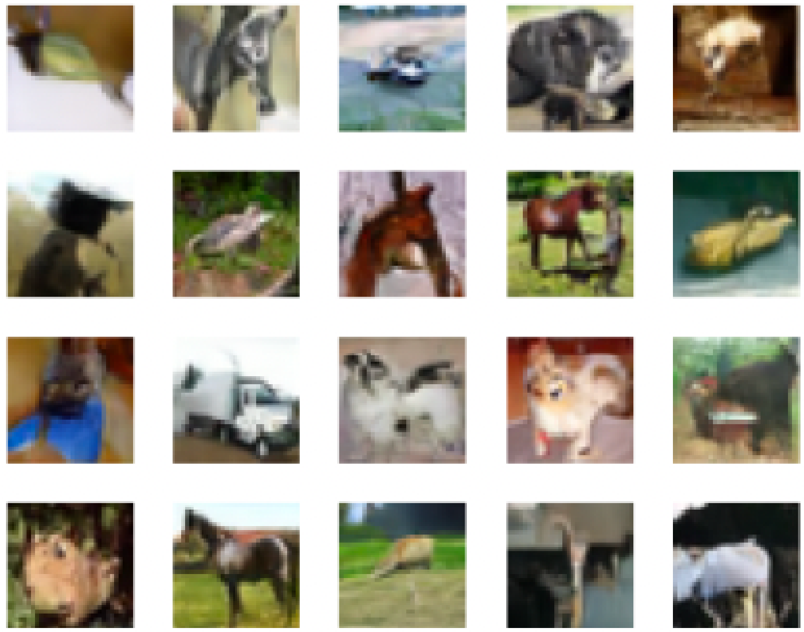


Figure 8.5: Images generated by an Epistemic Diffusion Model with image manifold trained on CIFAR-10.

8.8 Conclusion

This chapter introduced Epistemic Diffusion Models, a novel framework designed to inject second-order uncertainty into the generative process. By learning a distribution over the noise parameters—specifically using a Normal-Inverse-Gamma prior—we moved beyond the deterministic or rigid variance schedules of standard diffusion models.

Our initial experiments highlighted a critical trade-off: fixed-variance models maintain high fidelity but fail to improve diversity, while unconstrained fully Bayesian models suffer from training instability and quality degradation. However, by introducing Manifold-Aware Diffusion, specifically through projection-based constraints, we successfully stabilized the training.

The results demonstrate that constraining the epistemic uncertainty to the intrinsic geometry of the data manifold yields a sweet spot. As evidenced by the Vendi

Scores on CIFAR-10 and CelebA, our manifold-aware approach enhances sample diversity (e.g., Vendi Score 5.97 vs. 5.73 for baseline on CelebA) while maintaining competitive image fidelity. This confirms that modelling the uncertainty of the generation process is a viable path toward more creative and diverse generative AI, provided the exploration is grounded in the data's underlying structure.

Chapter 9

EPISTEMIC GENERATIVE ADVERSARIAL NETWORKS

9.1 Introduction

Generative models have achieved remarkable progress in synthesizing high-fidelity images, audio, and text, with Generative Adversarial Networks (GANs) standing out for their sample quality and training efficiency [18, 117, 92]. These models have revolutionized various domains, from computer vision to natural language processing, demonstrating unprecedented capabilities in generating realistic synthetic data [152]. Yet, despite numerous advances in objectives, architectures, and regularization techniques, lack of output diversity persists as a major challenge plaguing GANs and related generative approaches [117, 92, 124].

9.1.1 The Problem of Mode Collapse

GANs often exhibit mode dropping and mode collapse, producing samples that look plausible but concentrate on a narrow subset of the target distribution [120, 117]. This phenomenon occurs when the generator fails to capture the complete range of diversity present in the training data, instead focusing on generating limited and repetitive variations of samples [92, 13]. Mode collapse arises from imbalances in the training dynamics between the generator and the discriminator [120]. For example, if the discriminator is weak or learns too slowly, the generator can exploit this by converging to a narrow set of outputs that consistently fool the discriminator [92]. The discriminator, seeing only limited patterns, gives the generator no incentive to explore other modes [120].

This imbalance erodes coverage, biases downstream analyses, and obscures failure modes that are crucial for trustworthy deployment [120, 30]. In applications such as medical imaging or autonomous driving, where diverse and varied outputs are essential, mode collapse poses critical challenges [30, 124].

Addressing the diversity problem requires not only better training dynamics but also principled approaches to quantify and leverage uncertainty in the generative process [112, 60]. While traditional approaches mitigate mode collapse through heuristics or architectural tweaks, they often lack a principled theoretical framework for modeling the uncertainty inherent in the generative process itself.

9.1.2 Our Contribution: Epistemic GANs

In this chapter, we introduce a novel approach that leverages the Dempster-Shafer Theory of Evidence to address the diversity problem in GANs. By extending the traditional GAN framework with evidence-theoretic principles, we create generative models that capture greater diversity and provide meaningful measures of confidence.

This is achieved through three key innovations:

1. Evidential Discriminator: Modifying the discriminator to predict a *belief function* [34] rather than a probability distribution. This allows the discriminator to express ”ignorance” regarding ambiguous samples, rather than being forced into a binary decision.

2. Region-wise Uncertainty in Generator: Introducing architectural enhancements to the generator that enable region-wise uncertainty estimation via belief function prediction over perceptual states.

3. Generalized Evidential Loss: Developing a GAN loss formulation that operates within the belief function framework, explicitly penalizing conflicting beliefs and encouraging the generator to explore diverse modes.

9.2 Theoretical Framework

9.2.1 Dempster-Shafer Theory in GANs

The Dempster-Shafer theory provides a robust mathematical framework for reasoning under uncertainty [42, 39]. Unlike traditional Bayesian approaches, it handles incomplete evidence by assigning belief masses to *sets* of propositions.

In the context of a GAN, the propositions are typically {Real} and {Fake}. Standard GAN discriminators output a probability $P(\text{Real})$, implying $P(\text{Fake}) = 1 - P(\text{Real})$. This forces a decision even when evidence is scarce (e.g., pure noise inputs). Our Epistemic GAN allows the discriminator to assign mass to the composite set $\Theta = \{\text{Real}, \text{Fake}\}$, representing ignorance. This leads to a belief assignment:

$$b_{\text{real}} = bel(\{\text{real}\}), \quad b_{\text{fake}} = bel(\{\text{fake}\})$$

subject to the constraint $b_{\text{real}} + b_{\text{fake}} \leq 1$. The remaining mass $u = 1 - (b_{\text{real}} + b_{\text{fake}})$ represents the epistemic uncertainty.

9.3 Architecture

To embed evidential reasoning within the GAN architecture, we introduce targeted modifications to both the discriminator and the generator.

9.3.1 Discriminator with Belief Outputs

We replace the standard single-neuron output of the discriminator with a dual-output layer predicting belief masses (see Figure 9.1).

$$D(\mathbf{x}) = (b_{\text{real}}, b_{\text{fake}})$$

These values are produced via sigmoid activations. The constraint $b_{\text{real}} + b_{\text{fake}} \leq 1$ is enforced via a regularization term in the loss function. Aside from this change, the feature extraction backbone (e.g., DCGAN) remains identical.

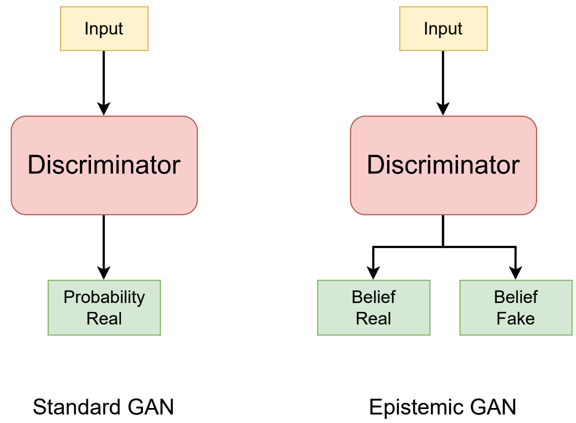


Figure 9.1: Discriminator architecture comparison. The Epistemic Discriminator outputs two belief values, allowing for a residual ”ignorance” mass.

By encoding both the strength of support for “real” and for “fake,” the discriminator conveys richer uncertainty information. If the discriminator is ”ignorant” ($b_{\text{real}} + b_{\text{fake}} < 1$), the gradients driving the generator are dampened compared to a confident decision, preventing over-optimization against a confused discriminator.

9.3.2 *Generator with Pixel-Wise Mass Prediction*

Our generator is restructured into two sequential modules to incorporate uncertainty into the synthesis pipeline (see Figure 9.2).

1. **Mass Function Prediction Stage:** The first module takes a latent vector $z \sim p(z)$ and outputs parameters $\alpha_{ijK}$for a Dirichlet distribution at each spatial region (i, j). This defines a discrete mass function over intensity intervals.

2. **Interval Sampling:** We sample intervals from these Dirichlet distributions. This acts as a structured, learned noise injection.

3. **Image Construction Stage:** The second module decodes this map of intervals into the final image $\hat{\mathbf{x}}$ via upsampling and convolutions.

This architecture allows the generator to ”operationalize” intervals. The variance in the sampled intervals forces the second module to be robust to a wider range of

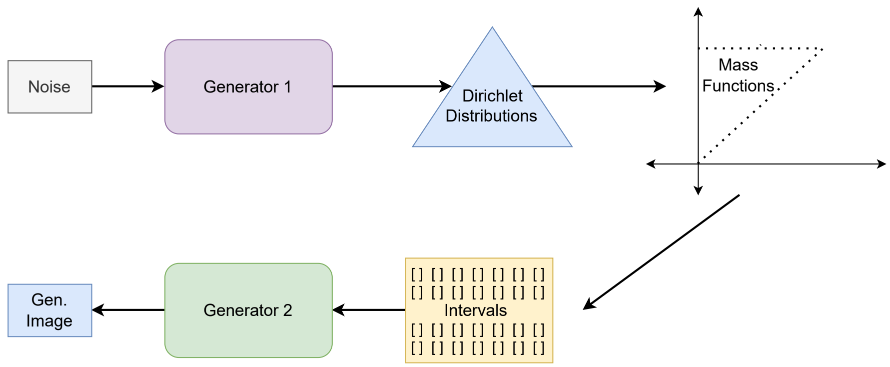


Figure 9.2: Generator Architecture and flow for Epistemic GANs. The generator predicts mass functions (Dirichlet parameters) for regions, samples intervals, and then constructs the image.

inputs, preventing it from collapsing to a single deterministic output and thereby enhancing diversity.

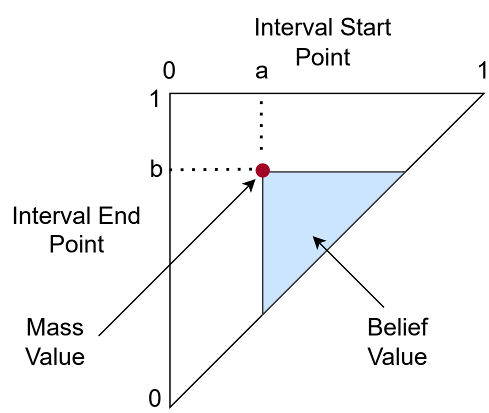


Figure 9.3: Representation of a belief function with Borel intervals used in the generator's intermediate representation.

9.4 Loss Function

We redesign the adversarial objectives to incorporate evidential reasoning.

9.4.1 Discriminator Loss

The discriminator D is trained to maximize belief in the correct label while minimizing the "ignorance" gap for clear samples.

$$
\begin{aligned}
L_{D} = & -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\log b_{\text{real}}(\mathbf{x}) + \log\left(1 - b_{\text{fake}}(\mathbf{x})\right)\right] \\
& -\mathbb{E}_{z\sim p(z)}\left[\log b_{\text{fake}}\left(G(z)\right) + \log\left(1 - b_{\text{real}}\left(G(z)\right)\right)\right] \\
& +\lambda\,\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\max\left(0,\, b_{\text{real}}(\mathbf{x}) + b_{\text{fake}}(\mathbf{x}) - 1\right)\right] \\
& +\lambda\,\mathbb{E}_{z\sim p(z)}\left[\max\left(0,\, b_{\text{real}}\left(G(z)\right) + b_{\text{fake}}\left(G(z)\right) - 1\right)\right].
\end{aligned}
$$

The regularization terms (weighted by λ) strictly penalize violations of the belief summation constraint $b_{\text{real}} + b_{\text{fake}} \leq 1$.

9.4.2 Generator Loss

The generator G aims to fool the discriminator but is also regularized to maintain diversity.

$$
\begin{aligned}
L_{G} = & -\mathbb{E}_{z\sim p(z)}\left[\log b_{\text{real}}\left(G(z)\right) + \log\left(1 - b_{\text{fake}}\left(G(z)\right)\right)\right] \\
& +\beta\,\mathbb{E}_{i,j}\left[\text{Var}\left[\text{Dir}\left(\alpha_{ij}\right)\right]\right] \\
& +\gamma\,\mathbb{E}_{i,j}\left[w_{ij}\right].
\end{aligned}
$$

- The Adversarial Term drives G to generate samples classified as real.
- The Variance Term (weighted by β) encourages the Dirichlet distributions to have high variance, promoting exploration and diversity.
- The Interval Width Term (weighted by γ) penalizes overly wide intervals, encouraging precision and realism.

This dual-objective creates a minimax-style equilibrium: variance encourages diversity, while width enforces precision.

Table 9.1: Performance of Standard GAN and Epistemic GAN on CelebA, CIFAR-10, and Food-101. Reference represents the Vendi score (diversity) of the training data. Epistemic GAN achieves lower FID and higher Vendi Scores.

	CelebA		Cifar-10		Food101	
	FID ↓	Vendi ↑	FID↓	Vendi ↑	FID↓	Vendi ↑
Reference	-	6.95	-	8.48	-	16.29
Standard GAN	18.5	5.70	25.9	4.25	33.76	12.78
Epistemic GAN	**17.3**	**5.86**	**24.1**	**4.53**	**29.1**	**13.82**

9.5 Experiments

9.5.1 Experimental Setup

Datasets: We evaluate on CelebA [89] (faces), CIFAR-10 [78] (objects), and Food-101 [16] (fine-grained categories). These datasets provide a robust testbed for both fidelity and diversity.

Implementation: We use a DCGAN [119] backbone for both the Standard GAN baseline and our Epistemic GAN. We deliberately chose this simple backbone to isolate the contribution of the evidential framework from architectural engineering (like StyleGAN). Hyperparameters were set to $\lambda = 1, \beta = 1, \gamma = 1$. All models were trained for equal epochs with Adam optimizer.

Metrics:

- Fréchet Inception Distance (FID) [62]: Measures image quality (lower is better).
- Vendi Score [50]: Measures sample diversity (higher is better).

9.5.2 Results and Analysis

Table 9.1 presents the quantitative results. The Epistemic GAN consistently outperforms the Standard GAN across all datasets in both quality and diversity.

On Food-101, a challenging dataset with 101 classes, we see a significant reduction in FID (33.76 → 29.1) and a boost in Vendi Score (12.78 → 13.82). This confirms that explicitly modeling epistemic uncertainty helps the generator cover more modes of the distribution.

Qualitative Results

Figure 9.4 compares generated samples. While visual differences can be subtle, the Epistemic GAN avoids ”mode dropping,” generating a wider variety of facial features and accessories compared to the baseline.

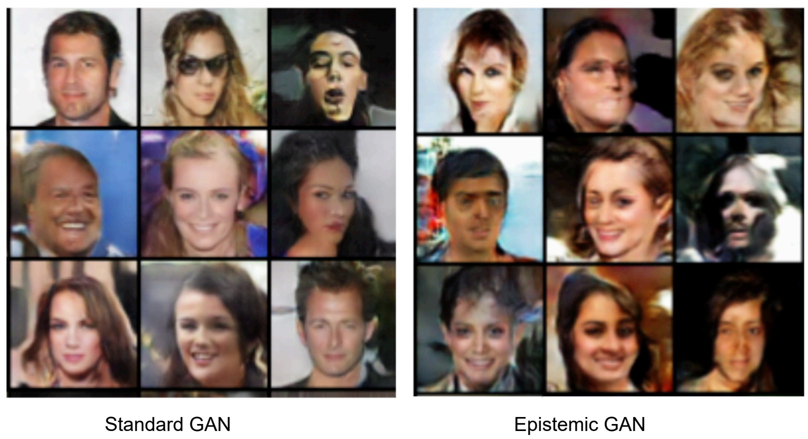


Figure 9.4: Generations for Standard GAN (left) and Epistemic GAN (right) on Celeb-A.

9.5.3 Ablation Studies

We conducted extensive ablations to validate our design choices.

Loss Components: Table 9.2 analyzes the impact of β (variance) and γ (width) terms. The results show that without β (variance term), diversity drops. Without γK

Table 9.2: Ablation for γK and β on CelebA. [FID (Vendi Score)]. The best balance is achieved when both terms are active (≈ 1).

γK \ βK	0	0.5	1	2
0	17.4 (5.35)	17.1 (5.11)	20.1(6.20)	23.7 (6.30)
0.5	18.9 (5.43)	17.5 (5.71)	18.4 (5.98)	20.2 (5.87)
1	19.1 (5.05)	19.71 (5.70)	**17.4 (5.82)**	22.6 (6.01)
2	21.6 (5.01)	23.1 (5.33)	19.6 (5.77)	21.7 (5.92)

(width term), quality (FID) degrades as the model produces overly vague intervals.

Furthermore, Table 9.3 presents the ablation study on the CelebA dataset for different values of the trade-off parameter λ, analyzing the impact of the normalization constraint.

Table 9.3: Ablation study for λ on the CelebA dataset. Results are reported in **FID (Vendi Score)** format.

λK	0	0.5	1	2
Score	19.4 (5.88)	17.4 (5.70)	**17.3 (5.90)**	20.1 (6.01)

Architecture Components: Table 9.4 shows that using *both* the Evidential Discriminator and Generator is necessary for optimal performance.

Table 9.4: Ablation for architecture components on CelebA.

Configuration	FID (Vendi Score)
Evidential Discriminator Only	15.1 (5.15)
Evidential Generator Only	19.2 (5.97)
Full Epistemic GAN	**17.3 (5.86)**

Training Stability: We measured the computational cost (Table 9.5). The Epistemic GAN incurs a negligible overhead ($\approx$ 1.5% increase in time per epoch), making it a practical drop-in replacement.

Table 9.5: Training time per epoch on CelebA (5 runs).

Model	Time (sec)
Standard GAN	75.91 ± 1.79
Epistemic GAN	77.10 ± 1.86

9.6 Application: Synthetic Data for Autonomous Driving

One of the biggest challenges in developing generalizable and robust deep learning models is the limited availability of diverse training data. This issue is particularly critical in the domain of autonomous driving, where real-world data collection is expensive, time-consuming, and often unable to capture the full range of rare or edge-case driving scenarios. A promising solution to this challenge is the use of synthetic data generation, which allows researchers to supplement real datasets with artificially created but realistic samples.

Synthetic data has already been shown to significantly enhance model performance, robustness, and generalization by exposing models to a broader distribution of inputs than real-world data alone can provide [128, 4]. Building on this idea, we propose Epistemic Generative Adversarial Networks for creating synthetic dataset for object detection and segmentation which can be utilized in autonomous driving pipeline. Unlike conventional GANs, Epistemic GANs are specifically designed to capture and model epistemic uncertainty, thereby enabling the generation of more diverse and representative training samples.

By leveraging Epistemic GANs, we aim to produce high-quality synthetic road scenarios—ranging from common traffic conditions to rare and unpredictable edge cases—that can be used to train next-generation autonomous driving models. This approach not only addresses the problem of limited data diversity but also paves the way for safer, more reliable, and more generalizable autonomous vehicle systems.

To this end, we train our proposed Epistemic GAN on the Cityscapes dataset [31], using 2,975 training images and a Pix2Pix-style [70] conditional adversarial setup. Given a semantic segmentation mask as input, the generator synthesizes a photorealistic street-scene image, while the discriminator judges realism conditioned on the same mask. The experiments look promising: the model preserves road topology and overall layout, and it avoids most mode-collapse. Figure 9.5 represnts qualitative results for Epistemic GANs on the Cityscapes dataset.

9.7 Conclusion

This chapter introduced Epistemic Generative Adversarial Networks, a novel generative framework grounded in the Dempster-Shafer theory of evidence. By equipping the discriminator with the capacity to express ”ignorance” and the generator

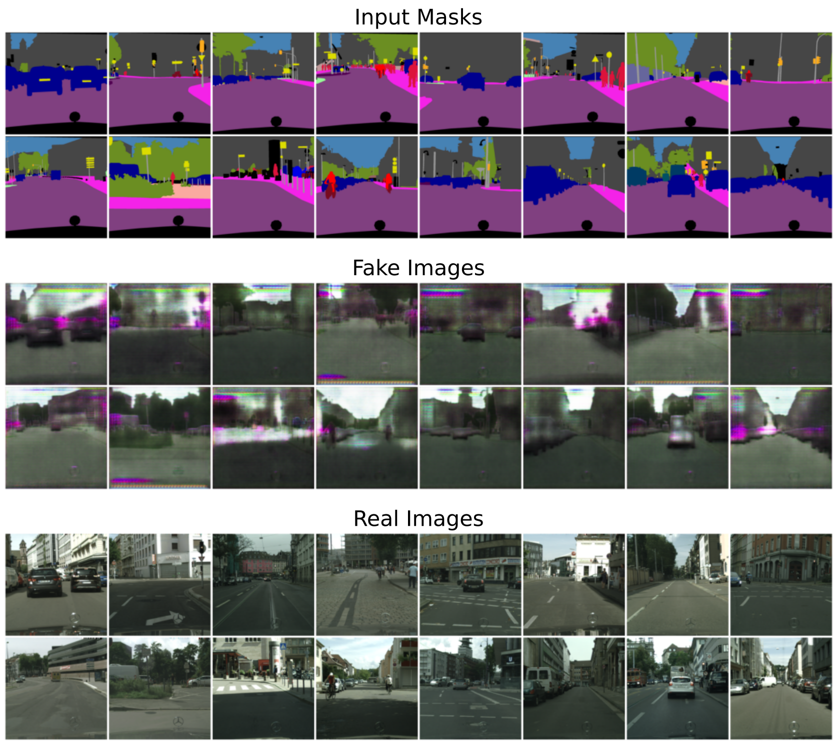


Figure 9.5: Epistemic GAN generations for road scenarios on Citiscape dataset.

with region-wise uncertainty awareness, we fundamentally altered the adversarial dynamics that typically lead to mode collapse.

The extensive experimental evaluation validates this theoretical shift. Across diverse datasets—from facial synthesis (CelebA) to fine-grained object generation (Food-101)—Epistemic GANs consistently outperformed standard baselines. Crucially, the improvement was not merely in visual fidelity (lower FID) but in the breadth of the learned distribution, as evidenced by significantly higher Vendi Scores. Our ablation studies further confirmed that the delicate balance between the variance-promoting term (β) and the precision-enforcing term (γ) is essential for achieving these gains.

Furthermore, the application to the Cityscapes dataset demonstrates the practical utility of this approach. By generating diverse, high-fidelity synthetic road scenes, Epistemic GANs offer a scalable solution for creating robust training data for autonomous driving systems, particularly for rare or edge-case scenarios that are difficult to capture in the wild.

Chapter 10

APPLICATION IN AUTONOMOUS DRIVING

10.1 Introduction

In the previous chapters, we developed a suite of methods for quantifying epistemic uncertainty in deep learning, ranging from classification (RS-NN) to generative modeling (RS-LLM, Epistemiv Diffusion Models, and Epistemic GANs). In this final chapter, we apply these theoretical advancements to one of the most demanding and safety-critical applications of modern AI: Autonomous Driving (AD).

A fundamental limitation of existing autonomous driving datasets is the absence of explicit, temporally persistent annotations of road user *intent* and long-term goals. Current datasets typically provide low-level information—trajectories, bounding boxes, or short-horizon manoeuvre labels—but rarely explain *why* an agent behaves in a certain way. Consequently, models are trained to predict *what* an agent is doing, but not the underlying intent that drives that behavior.

To address this, we introduce two major contributions:

1. **The ROAD-INTENT Dataset**: A novel dataset featuring choreographed scenarios with known ground-truth intent, specifically designed to capture ambiguous and unexpected events.

2. **Random-Set Vision Language Models (RS-VLMs)**: An extension of our RS-LLM framework to the multi-modal domain. We use RS-VLMs to automate the labeling of the ROAD-INTENT dataset, providing video descriptions enriched with epistemic uncertainty scores.

10.2 The ROAD-INTENT Dataset

The primary motivation behind ROAD-INTENT is to enable the rigorous training of advanced prediction algorithms, particularly those relying on epistemic inverse reinforcement learning [109], preference learning [25], and theory-of-mind reasoning [7].

10.2.1 Data Collection at the RACE Facility

Data collection will be conducted at the RACE facility at the Culham Centre for Fusion Energy. This facility offers a large, sandboxed campus that serves as a realistic proving ground. It features a closed-road network with junctions, pedestrian crossings, parking areas, and RADAR-realistic movable foam infrastructure (see Figure 10.1). This allows for the controlled repetition of scenarios while preserving the sensor noise characteristics of the real world.

Figure 10.1: RADAR-realistic roadside infrastructure at the RACE facility, allowing for reconfigurable urban layouts.

The data capture involves a multi-perspective sensor suite:

- Ego Vehicle (EV): Equipped with automotive-grade LiDAR, Cameras, GPS, IMU, and CANBus data (velocity, steering, throttle).
- External Sensors: Vehicle-mounted GoPros on non-ego vehicles.

- Wearable Sensors: Chest/head-mounted cameras on pedestrian and driver actors.
- Aerial View: Drone-mounted cameras providing a bird's-eye view for global trajectory alignment.

10.2.2 Scenario Design and Complexity

A distinguishing feature of ROAD-INTENT is the deliberate inclusion of unexpected events—sudden changes of intent, hesitation, or near-misses—that are dangerous to capture on public roads. We defined varying levels of Complexity for each scenario to systematically test epistemic uncertainty.

Example Scenario: Giving Way This scenario examines interactions where intent is revealed late.

- **Complexity 1 (Expected)**:

  Car 1 is parked on the same side of the road as the ego vehicle. Before the ego vehicle reaches Car 1, Car 1 signals and leaves the parking spot. The ego vehicle passes while Car 2 continues driving from the opposite direction.

  Intent: Car 1 leaves the parking spot; Car 2 drives on.

  Actions: Signalling, driving out, passing.

- **Complexity 2 (Ambiguous)**:

  The ego vehicle reaches Car 1 and stops to give way. Car 2 passes from the opposite direction. The ego vehicle then continues while Car 1 remains parked.

  Intent: Car 1 stays in the parking spot; Car 2 drives on.

  Actions: Waiting, passing.

- **Complexity 3 (Unexpected)**:

The ego vehicle stops to allow Car 2 to pass. After Car 2 has passed and the ego vehicle begins to move, Car 1 unexpectedly signals and pulls out, forcing the ego vehicle to stop abruptly.

Intent: Car 1 leaves the parking spot; Car 2 drives on.

Actions: Sudden signalling, unexpected merge, abrupt stopping.

This structure ensures that the dataset contains the necessary ”epistemic shocks” required to train robust uncertainty-aware models.

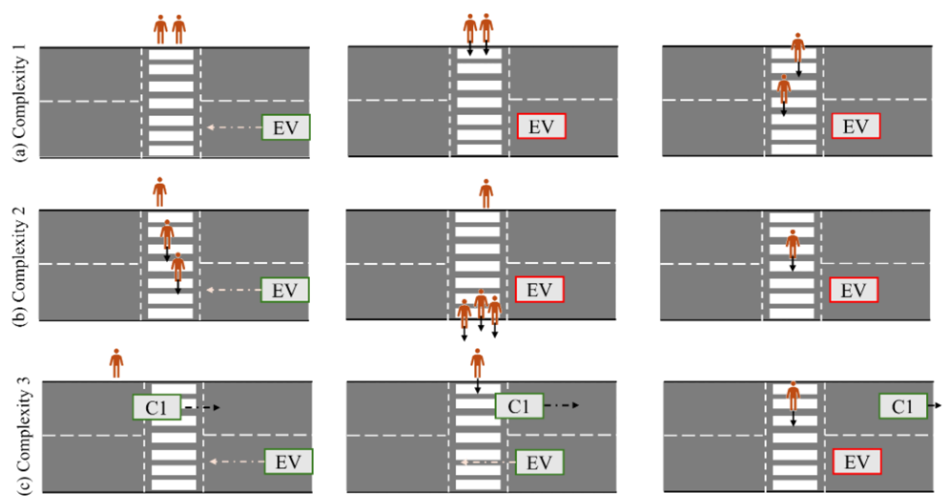


Figure 10.2: Illustration of the Giving Way scenario across three levels of complexity. (a) Expected behaviour. (b) Ambiguous behaviour. (c) Unexpected behaviour causing a near-miss.

10.2.3 Annotation Protocol: Intent vs. Action

The annotation strategy explicitly separates **Observable Action** from **Latent Intent**.

- **Action Labels**: Low-level behaviors aligned with sensor data (e.g., *walking, waiting, signaling*). These change dynamically.
- **Intent Labels**: Medium-to-long term goals (e.g., *intention to cross*, *intention to load vehicle*). These are derived from the a priori choreography and remain fixed even if the agent hesitates.

This separation is crucial for Inverse Reinforcement Learning (IRL) for autonomous driving, as it allows algorithms to learn that an agent stopping at a curb (action) still possesses the desire to cross (intent), but is inhibited by oncoming traffic.

10.3 Random-Set Vision Language Models (RS-VLMs)

To scale the annotation of such a complex dataset, manual labeling is insufficient. We require an automated system that can describe video scenes while quantifying its own reliability. To this end, we extend the RS-LLM framework (Chapter 7) to the multi-modal domain, introducing the Random-Set Vision Language Model (RS-VLM).

10.3.1 Methodology and Architecture

The RS-VLM aligns closely with the RS-LLM but incorporates a visual encoder to process video frames. We utilize the InternVideo2.5-Chat-8B architecture [147], which combines:

1. Visual Encoder: InternViT [24] to encode frames into embeddings.
2. Language Model: InternLM2.5-7B [20] to generate text.
3. Projector: An MLP to map visual tokens into the LLM's input space.

Epistemic Modification: Standard VLMs predict a probability distribution over the next token using Softmax. This forces a choice even when visual evidence is ambiguous (e.g., a blurry road sign). RS-VLM replaces the final layer to predict a Belief Function over sets of tokens. The output space is defined by a budget O of focal sets obtained via hierarchical clustering of the vocabulary (as described in Section 7.3.1).

The model is trained using the epistemic loss function $\mathcal{L}_{RS}$:

$$\mathcal{L}_{RSK} = \mathcal{L}_{BCE}(\mathbf{bel}, \hat{\mathbf{bel}}) + \alpha M_{rK} + \beta M_{sK} \tag{10.1}$$

where $M_{rK}$and M_{sK}enforce the non-negativity and sum-to-one constraints of the Dempster-Shafer theory.

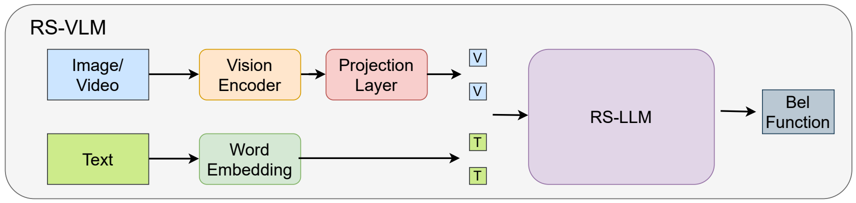


Figure 10.3: Architecture flow of RS-VLM. Visual tokens from the video encoder are concatenated with text prompts. The model predicts a belief function over a budget of token sets, which is converted to a mass function for uncertainty estimation.

10.3.2 Experiments

We evaluated RS-VLM on standard Video QA benchmarks to ensure that the epistemic modification does not degrade performance.

Datasets

- MSVD-QA [153]: A dataset of short YouTube clips with descriptions.
- TGIF-QA [85]: A large-scale dataset of animated GIFs focusing on spatio-temporal reasoning (repetition counting, state transitions).

Performance Comparison

Table 10.1 shows the accuracy of RS-VLM compared to the standard InternVideo2.5 baseline. RS-VLM matches or exceeds the performance of the standard model, confirming that belief-function training preserves the representational power of the VLM.

Hallucination Detection

A critical requirement for automated labeling is detecting when the model is ”hallucinating” (generating text not supported by the video). We tested this by randomly

Table 10.1: Performance (Accuracy) of Standard VLM and RS-VLM on MSVD-QA and TGIF-QA.

Model	**MSVD-QA**	**TGIF-QA**
Standard VLM	**0.72**	0.71
RS-VLM	0.71	**0.73**

replacing the input video with an irrelevant one while keeping the question constant. The model's uncertainty was measured using the entropy of the Pignistic Probability.

As shown in Table 10.2, RS-VLM demonstrates superior separability between Correct Context (CC) and Incorrect Context (IC), achieving significantly higher AUROC scores (89.62 vs 75.31 on MSVD-QA). This proves that RS-VLM can reliably flag its own hallucinations.

Table 10.2: Hallucination Detection: Uncertainty evaluation (Entropy) on Correct (CC) vs Incorrect (IC) video contexts.

Model	**MSVD-QA**			**TGIF-QA**		
	CC (↓)	IC (↑)	AUROC	CC (↓)	IC (↑)	**AUROC**
Standard VLM	0.58 ± 0.36	0.73 ± 0.55	75.31	0.77 ± 0.46	1.24 ± 0.71	73.68
RS-VLM	0.19 ± 0.15	0.45 ± 0.29	**89.62**	0.26 ± 0.25	0.79 ± 0.81	**88.13**

10.4 Annotating ROAD-INTENT with RS-VLM

We applied the validated RS-VLM to the task of annotating the ROAD-INTENT dataset. The goal is to generate high-level textual descriptions of actor intent to facilitate Inverse Reinforcement Learning (IRL).

10.4.1 Qualitative Results

We prompted the RS-VLM with: *"This video is the dashcam view of a car. This is the road view and there are different actors in this video. Describe in detail the actions and intent of all actors (cars and pedestrians). How did their actions affect the driving?"*

Using a sample video from the dataset (Figure 10.4), the model generated the following description:

> *"The video shows a road with a building on the left and a fence on the right. A black car is approaching from the distance. As the car gets closer, a person wearing a red jacket and carrying a backpack walks towards the car from the right side of the road. The car stops near the person, and another person wearing a black jacket and carrying a yellow backpack walks towards the car from the left side of the road. The car then starts moving forward and drives away from the camera."*

The model correctly identified the key actors, their clothing (red/black jackets), and the interaction dynamics (stopping/moving). Crucially, these generations come with an uncertainty score derived from the credal set width.

10.4.2 Impact on Inverse Reinforcement Learning

The integration of RS-VLM annotations enhances the IRL pipeline in three ways:

1. Semantic Abstraction: Converting raw pixels to text helps the IRL algorithm learn rewards based on high-level concepts (e.g., "pedestrian waiting") rather than visual noise.

2. Intent Modeling: By explicitly prompting for "intent", the model provides the latent variables necessary for Theory of Mind reasoning.

3. Uncertainty Integration: The RS-VLM's uncertainty score acts as a filter. IRL algorithms can be designed to weigh high-confidence annotations heavily while disregarding ambiguous ones, preventing the learning of dangerous behaviors from noisy labels.

Figure 10.4: Sample frames from the ROAD-INTENT dataset used for automated annotation.

10.5 Conclusion

This chapter demonstrated the practical translation of our theoretical frameworks into the high-stakes domain of autonomous driving. By establishing the ROAD-INTENT dataset, we addressed a critical gap in existing resources: the absence of ground-truth intent labels for ambiguous road scenarios. The deliberate inclusion of ”epistemic shocks”—situations where intent is hidden or changes suddenly—provides a necessary proving ground for the next generation of robust prediction algorithms.

To make this complex dataset usable at scale, we developed the Random-Set Vision Language Model (RS-VLM). This architecture successfully extends our RS-LLM methodology to multi-modal inputs, allowing for the automated generation of intent descriptions that are both semantically rich and statistically reliable. Crucially, the RS-VLM does not just describe a scene; it quantifies its own uncertainty about that description.

The experimental validation confirms that RS-VLMs match the accuracy of standard models while offering superior hallucination detection capabilities (AUROC of 89.62 on MSVD-QA). By integrating these uncertainty-aware annotations into the Inverse Reinforcement Learning pipeline, we pave the way for autonomous systems that can learn not just to mimic human actions, but to understand the intent behind them—and, vitally, to recognize when that intent is unclear.

Chapter 11

CONCLUSION AND FUTURE WORK

11.1 Summary of Contributions

This thesis began with a simple but troubling observation: modern Deep Learning models are incredibly powerful, yet dangerously overconfident. From a classifier mistaking a blurred image for a definite class, to a Large Language Model confidently hallucinating a historical fact, the inability of standard AI to say ‘‘I don’t know’’ poses a massive risk to their deployment in the real world.

The central hypothesis of this work was that Epistemic Uncertainty—the uncertainty arising from a lack of knowledge—can be rigorously modelled using the mathematical framework of Random Sets (Belief Functions), and that this can be done without sacrificing the scalability required for modern neural networks.

Over the course of this thesis, we have substantiated this hypothesis through several key contributions:

1. **Solving the Scalability Bottleneck:** We addressed the exponential complexity of random sets (2) by introducing Budgeting. By using unsupervised clustering (GMMs for RS-NN, Hierarchical Clustering for RS-LLM), we showed that we can restrict the output space to a manageable set of semantically meaningful focal sets. This was the foundation that made everything else possible.

2. **Random-Set Neural Networks (RS-NN):** We moved beyond point-estimate classification. The RS-NN predicts belief functions, allowing it to achieve state-of-the-art results in Out-of-Distribution (OoD) detection and adversarial robustness, while remaining as fast as a standard CNN at inference time.

3. **A Unified Evaluation Framework:** We recognised that existing metrics like

Accuracy are insufficient for set-valued predictions. We proposed a new metric ($\mathcal{E}$) that balances the trade-off between being correct (Distance) and being informative (Non-specificity), providing a fair way to benchmark epistemic models.

4. **Epistemic Generative AI:** We extended these ideas to the generative frontier.
   - With RS-LLMs, we showed that predicting belief functions over token sets helps detect hallucinations and improves factual consistency in text generation.
   - With Epistemic GANs and Diffusion Models, we demonstrated that modelling the uncertainty of the generation process itself significantly improves the diversity of generated images, mitigating mode collapse.
5. **Real-World Application:** Finally, we grounded these theories in the high-stakes domain of Autonomous Driving. We created the ROAD-INTENT dataset and developed RS-VLMs to automatically annotate actor intent with uncertainty scores, paving the way for safer self-driving systems.

11.2 Limitations

Despite these successes, there are limitations to the current approach that must be acknowledged.

Dependence on Budgeting: The performance of our models relies heavily on the quality of the pre-computed budget. If the initial clustering fails to capture the true semantic ambiguities of the dataset, the model's ability to express uncertainty is compromised. Currently, this budget is fixed before training. Ideally, the budget should be dynamic and learned end-to-end along with the model weights.

Hyperparameter Sensitivity: Our loss functions introduce regularization terms (e.g., α, β for mass validity) and trade-off parameters. While we found stable values for our experiments, these add an extra layer of complexity to the training process compared to standard Cross-Entropy.

Computational Overhead in Generation: While RS-NN classification is fast, sampling from Epistemic Diffusion models or RS-LLMs involves slightly more complex operations (sampling from Dirichlet or Inverse-Gamma distributions) than standard sampling. While still small for large models, it is a non-zero cost.

11.3 Future Work

The work presented here opens several exciting avenues for future research.

11.3.1 Dynamic and End-to-End Budgeting

A natural next step is to remove the fixed budget constraint. We aim to explore methods where the focal sets are learned dynamically during training, perhaps using attention mechanisms to identify which subsets of classes are currently causing confusion and adding them to the budget on the fly.

11.3.2 Manifold-Aware Diffusion

As discussed in Chapter 8, unconstrained variance learning in diffusion models can be unstable. Future work will focus on constraining the epistemic diffusion process to other intrinsic geometry (manifold) of the image data, potentially using Riemannian flow matching techniques.

11.3.3 Towards Random-Set Concept Models

One of the most promising extensions of RS-LLMs work lies in moving beyond tokens. Despite the amazing progress of large language models, there is still a gap between what they do and how people actually think. Humans don't just predict the next word; we plan, revise, and reason across different levels of abstraction—from a rough idea, to a structured outline, down to the exact phrasing. LLMs, even the strongest ones, often struggle with that kind of explicit, multi-level planning. They can do chain-of-thought or tool use, sure, but it is not the same as holding and manipulating concepts across layers like a person does. Large Concept Models (LCMs) [12] take a step toward that goal by moving away from pure token-level processing and working in an abstract embedding space where reasoning can be more hierarchical.

These LCMs can extended with Random-Set inference, creating Random-Set Large Concept Models (RS-LCMs). Instead of committing to a single point in concept space, an RS-LCM will predict a mass function over regions of the embedding manifold, placing belief on *sets* of candidate concepts rather than just one. Practically, this can be achived by belief using a low-dimensional Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha}_{ij})$ with three concentration parameters $\alpha_1, \alpha_2, \alpha_3$.

11.4 Final Words

This thesis has argued that AI systems must be self-aware of their limitations to be truly trustworthy. By marrying the rigour of Random Set theory with the power of Deep Learning, we have provided a set of tools to make this possible. We hope this work serves as a foundation for a future where AI not only answers our questions but also tells us, with honesty and precision, when it simply does not know.

BIBLIOGRAPHY

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[5] Alessandro Antonucci and Fabio Cuzzolin. Credal sets approximation by lower probabilities: application to credal networks. In *Computational Intelligence for Knowledge-Based Systems Design: 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010, Dortmund, Germany, June 28-July 2, 2010. Proceedings 13*, pages 716–725. Springer, 2010.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[7] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[8] Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned llms using lora ensembles. *arXiv preprint arXiv:2402.12264*, 2024.

[9] Debarag Banerjee, Pooja Singh, Arjun Avadhanam, and Saksham Srivastava. Benchmarking llm powered chatbots: methods and metrics. *arXiv preprint arXiv:2308.04624*, 2023.

[10] Puneet Bansal. Intel image classification. *Available on https://www. kaggle. com/puneet6060/intel-image-classification, Online*, 2019.

[11] Jonathan Baron. Second-order probabilities and belief functions. *Theory and Decision*, 23:25–36, 1987.

[12] Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024.

[13] Farhat Lamia Barsha and William Eberle. Mode collapse detection strategies in generative adversarial networks for credit card fraud detection. *Proceedings of the Florida Artificial Intelligence Research Society Conference*, 2024.

[14] Reza Bayat. A study on sample diversity in generative models: Gans vs. diffusion models. 2023.

[15] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

[16] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[17] Remco Ronaldus Bouckaert. *Bayesian belief networks: from construction to inference*. PhD thesis, 1995.

[18] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[19] Wray L. Buntine and Andreas S. Weigend. Bayesian back-propagation. *Complex Syst.*, 5, 1991.

[20] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[21] Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Imprecise Bayesian neural networks. *arXiv preprint arXiv:2302.09656*, 2023.

[22] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

[23] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[24] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding per-

formance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[25] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[26] Tristan Cinquin and Robert Bamler. Regularized kl-divergence for well-defined function-space variational inference in bayesian neural networks. *arXiv preprint arXiv:2406.04317*, 2024.

[27] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[28] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

[29] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[30] Matthew Cobbinah, Henry Nunoo-Mensah, Prince Ebenezer Adjei, Francisca Adoma Acheampong, Isaac Acquah, Eric Tutu Tchao, Andrew Selasi Agbemenu, Jerry John Kponyo, and Emmanuel Abaidoo. Diversity in stable gans: A systematic review of mode collapse mitigation strategies. *Engineering Reports*, 2025.

[31] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele.

The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[32] Fabio Cuzzolin. On the credal structure of consistent probabilities. In *European Workshop on Logics in Artificial Intelligence*, pages 126–139. Springer, 2008.

[33] Fabio Cuzzolin. Complexes of outer consonant approximations. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 10th European Conference, ECSQARU 2009, Verona, Italy, July 1-3, 2009. Proceedings 10*, pages 275–286. Springer, 2009.

[34] Fabio Cuzzolin. Three alternative combinatorial formulations of the theory of evidence. *Intelligent Data Analysis*, 14(4):439–464, 2010.

[35] Fabio Cuzzolin. On consistent approximations of belief functions in the mass space. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 287–298. Springer, 2011.

[36] Fabio Cuzzolin. Lp consonant approximations of belief functions. *IEEE Transactions on Fuzzy Systems*, 22(2):420–436, 2013.

[37] Fabio Cuzzolin. Visions of a generalized probability theory. *arXiv preprint arXiv:1810.10341*, 2018.

[38] Fabio Cuzzolin. *The Geometry of Uncertainty: The Geometry of Imprecise Probabilities*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, 2020.

[39] Fabio Cuzzolin. Reasoning with random sets: An agenda for the future. *arXiv preprint arXiv:2401.09435*, 2023.

[40] Fabio Cuzzolin. Uncertainty measures: A critical survey. *Information Fusion*, page 102609, 2024.

[41] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

[42] Arthur P. Dempster. *Upper and Lower Probabilities Induced by a Multivalued Mapping*, volume 38. Institute of Mathematical Statistics, 1967.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[44] Thierry Denoeux. Nn-evclus: Neural network-based evidential clustering. *Information Sciences*, 572:297–330, 2021.

[45] Didier Dubois and Henri Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy sets and systems*, 24(2):161–182, 1987.

[46] Didier Dubois and Henri Prade. Consonant approximations of belief functions. *International Journal of Approximate Reasoning*, 4:419–449, 1990.

[47] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.

[48] David Freedman. Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141, 1999.

[49] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[50] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[51] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[52] Junyu Gao, Mengyuan Chen, Liangyu Xiang, and Changsheng Xu. A comprehensive survey on evidential deep learning and its applications. *arXiv preprint arXiv:2409.04720*, 2024.

[53] Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8513–8521, 2018.

[54] Eleonora Giunchiglia, Mihaela Cătălina Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. Road-r: The autonomous driving dataset with logical requirements. *Machine Learning*, 112(9):3261–3291, 2023.

[55] Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. Uncertainty-aware machine translation evaluation. *arXiv preprint arXiv:2109.06352*, 2021.

[56] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[57] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[58] Joseph Y. Halpern. *Reasoning About Uncertainty*. MIT Press, 2017.

[59] Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. trlx: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8578–8595, 2023.

[60] Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. A survey on uncertainty quantification methods for deep neural networks. *arXiv preprint arXiv:2302.13425*, 2023.

[61] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[62] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[63] Max Hinne, Quentin F Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215, 2020.

[64] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[65] Marius Hobbhahn, Agustinus Kristiadi, and Philipp Hennig. Fast predictive uncertainty for classification with Bayesian deep networks. In *Uncertainty in Artificial Intelligence*, pages 822–832. PMLR, 2022.

[66] Carl-Johan Hoel, Krister Wolff, and Leo Laine. Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[67] https://sbert.net/. SentenceTransformers Documentation &x2014; Sentence Transformers documentation — sbert.net. https://sbert.net/. [Accessed 23-07-2025].

[68] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[69] Guillermo Iglesias, Edgar Talavera, and Alberto Díaz-Álvarez. A survey on gans for computer vision: Recent research, analysis and taxonomy. *Computer Science Review*, 48:100553, 2023.

[70] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[71] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

[72] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[73] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[74] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[75] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *arXiv:1703.04977*, 2017.

[76] Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[77] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2012.

[78] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (Canadian Institute For Advanced Research). Technical report, CIFAR, 2009.

[79] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

[80] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[81] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[82] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[83] Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2005.

[84] Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. The MIT Press, Cambridge, Massachusetts, 1980.

[85] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

[86] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

[87] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[88] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[89] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[90] A Llera and CF Beckmann. Estimating an inverse gamma distribution. *arXiv preprint arXiv:1605.01019*, 2016.

[91] A Llera, D Vidaurre, RHR Pruim, and CF Beckmann. Variational mixture models with gamma or inverse-gamma components. *arXiv preprint arXiv:1607.07573*, 2016.

[92] Yixin Luo and Zhouwang Yang. Dyngan: Solving mode collapse in gans with dynamic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5493–5503, 2024.

[93] David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 05 1992.

[94] Shireen Kudukkil Manchingal and Fabio Cuzzolin. Epistemic deep learning. *arXiv preprint arXiv:2206.07609*, 2022.

[95] Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[96] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[97] Marie-Hélène Masson and Thierry Denoeux. Ecm: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008.

[98] Georges Matheron. *Random sets and integral geometry*. Wiley Series in Probability and Mathematical Statistics, New York, 1975.

[99] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[100] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[101] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[102] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

[103] Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

[104] Ilya Molchanov. Random sets and random functions. *Theory of Random Sets*, pages 451–552, 2017.

[105] Ilya S Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

[106] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv preprint arXiv:2102.11582*, 2021.

[107] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[108] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, Spain, 2011.

[109] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[110] Hung T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

[111] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[112] Philipp Oberdiek, Gernot A. Fink, and Matthias Rottmann. Uqgan: A unified model for uncertainty quantification of deep classifiers trained via conditional gans. *arXiv preprint arXiv:2201.13279*, 2022.

[113] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[114] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[115] Ziqi Pan, Li Niu, and Liqing Zhang. Unigan: Reducing mode collapse in gans using a uniform generator. *Advances in neural information processing systems*, 35:37690–37703, 2022.

[116] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016.

[117] Sen Pei, Richard Yi Da Xu, Shiming Xiang, and Gaofeng Meng. Alleviating mode collapse in gan via diversity penalty module. *arXiv preprint arXiv:2108.02353*, 2021.

[118] Benjamin Plaut, Khanh Nguyen, and Tu Trinh. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *arXiv preprint arXiv:2402.13213*, 2024.

[119] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[120] Nikitha Rao, Piyush M. Surana, and Nypunya Devraj. Analyzing the mode collapse problem in gans. Available at: https://raonikitha.github.io/files/academic-posts/ModeCollapseProblem.pdf, 2022.

[121] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[122] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[123] Tim GJ Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractable function-space variational inference in Bayesian neural networks. *Advances in Neural Information Processing Systems*, 35:22686–22698, 2022.

[124] Muhammad Muneeb Saad, Mubashir Husain Rehmani, and Ruairi O'Reilly. Adaptive input-image normalization for solving the mode collapse problem in gan-based x-ray images. *arXiv preprint arXiv:2309.12245*, 2023.

[125] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[126] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[127] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3183–3193, Red Hook, NY, USA, 2018. Curran Associates Inc.

[128] Pratinav Seth, Akshat Bhandari, and Kumud Lakara. Analyzing effects of fake training data on the performance of deep learning systems. *arXiv preprint arXiv:2303.01268*, 2023.

[129] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[130] Glenn Shafer. A theory of statistical evidence. In W. L. Harper and C. A. Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 2, pages 365–436. Reidel, Dordrecht, 1976.

[131] Philippe Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38(2):133–147, 2005.

[132] Philippe Smets et al. Constructing the pignistic probability function in a context of uncertainty. In *UAI*, volume 89, pages 29–40, 1989.

[133] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Ma-*

chine Learning Research, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[134] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[135] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[136] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[137] Vincent Spruyt. How to draw a covariance error ellipse. 2014. *Computer Vision for Dummies.[Available Online-http://www. visiondummy. com/2014/04/drawerror-ellipse-representing-covariance-matrix/]*, 2013.

[138] Zheng Tong, Philippe Xu, and Thierry Denoeux. An evidential classifier based on Dempster-Shafer theory and deep learning. *Neurocomputing*, 450:275–293, 2021.

[139] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[140] Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All you need is a good functional prior for Bayesian deep learning. *arXiv preprint arXiv:2011.12829*, 2020.

[141] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

[142] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[143] Manuel A Vega and Michael D Todd. A variational Bayesian neural network for structural health monitoring and cost-informed decision-making in miter gates. *Structural Health Monitoring*, 21(1):4–18, 2022.

[144] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[145] Daniel J Walters, Gülden Ülkümen, David Tannenbaum, Carsten Erner, and Craig R Fox. Investor behavior under epistemic vs. aleatory uncertainty. *Management Science*, 69(5):2761–2777, 2023.

[146] Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *arXiv preprint arXiv:2401.05043*, 2024.

[147] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

[148] Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *arXiv preprint arXiv:2406.11675*, 2024.

[149] Li Kevin Wenliang and Ben Moran. Score-based generative models learn manifold-like structures with constrained mixing. *arXiv preprint arXiv:2311.09952*, 2023.

[150] Luhuan Wu and Sinead A Williamson. Posterior uncertainty quantification in neural networks using data augmentation. In *International Conference on Artificial Intelligence and Statistics*, pages 3376–3384. PMLR, 2024.

[151] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[152] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, Brussels, Belgium, 2018. Association for Computational Linguistics.

[153] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[154] Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[155] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.