

Classification From Synthetic Data

When Diffusion-Generated Synthetic Data Stops Improving and Starts Memorising

Oliver Greenhalgh

Student ID: 3023322

Supervised by Per Kristian Lehre

Abstract

This study investigates whether classifier accuracy obtained from synthetic data reaches a near-maximum threshold before a diffusion model begins memorising its training examples, and how the interval between these events changes with training-set size. Conditional diffusion models were trained on MNIST subsets containing 200 to 10,000 images and evaluated between 250 and 64,000 optimiser steps. At each checkpoint, 5,000 generated images were used to train a classifier evaluated using a Train on Synthetic, Test on Real protocol, while memorisation was measured against the generator's exact training subset using a nearest-neighbour ratio criterion. The operational accuracy plateau, defined as the first checkpoint reaching 95% of a comparable reference maximum, occurred between approximately 2,000 and 4,000 steps. Plateau accuracy reached 93.1% to 98.5% of the corresponding step-matched real-data baseline under the study's fixed-synthetic-sample design. Memorisation onset moved later as the training set grew, while the operational plateau remained nearly fixed. From $n = 700$ upwards, the plateau preceded memorisation with no large or consistent subsequent decline in accuracy. These results provide a method for identifying a potential early-stopping interval after classifier accuracy reaches a selected threshold but before memorisation begins.

Honour Code

I certify that this dissertation is my own work. I used generative AI, including Claude Sonnet 5 and Claude Opus 5 to: generate plotting scripts for figures, format experimental results into tables, proofread, restructure writing. AI has also provided limited assistance with code development, debugging and review. The research, analysis, interpretation and conclusions drawn are my own. I reviewed, verified and edited all AI-assisted material and take full responsibility for the accuracy and integrity of the final submission.

Acknowledgements

I would like to extend special thanks to my supervisor Per Kristian Lehre for his ongoing support throughout this project. His guidance, support and weekly meetings have been invaluable throughout this process, and I wish him the best for his future research. I would also like to extend special thanks to my parents for their continued support throughout my studies.

Contents

Abstract . . . 1
Honour Code . . . 2
Acknowledgements . . . 2

1 Introduction **6**
1.1 Background & Motivation . . . 6
1.2 Aims & Objectives . . . 6
1.3 Research Questions . . . 7
1.4 Contributions . . . 7
1.5 Report Structure . . . 8

2 Literature Review **9**
2.1 Conditional Diffusion Models . . . 9
2.2 Synthetic Data for Classification . . . 9
2.3 Fidelity, Diversity and Classifier Accuracy . . . 10
2.4 Training Dynamics and the Generalisation Window . . . 11
2.5 Checkpoint Selection . . . 11
2.6 Summary and Research Gap . . . 12

3 Methodology **13**
3.1 Research Design & Rationale . . . 13
3.2 Experimental Procedure . . . 14
3.3 Dataset . . . 14
3.4 Generator . . . 15
3.5 Sampling Protocol . . . 16
3.6 Measuring Memorisation . . . 16
3.7 Measuring Classifier Accuracy . . . 17
3.8 Baselines & Controls . . . 18
3.9 Event Definitions & Scaling Analysis . . . 19
3.10 Reproducibility & Computational Environment . . . 20
3.11 Limitations . . . 20
3.12 CIFAR-10 . . . 20

4 Results **21**
4.1 Scope of the Completed Runs . . . 21
4.2 Validation of Measurement Instruments . . . 21
4.2.1 Sensitivity and Specificity . . . 21
4.2.2 False-Positive Calibration . . . 22
4.2.3 Sampler Comparison . . . 22
4.2.4 Sampler Comparison . . . 22
4.2.5 Generator Validation Loss . . . 22
4.2.6 Real-Data Baselines . . . 22
4.3 Research Question 1 . . . 23
4.4 Research Question 2 . . . 24
4.4.1 The Relationship Between Utility and Memorisation . . . 24
4.4.2 Diversity . . . 25
4.5 Research Question 3 . . . 26
4.5.1 The Classifier Accuracy Plateau . . . 26
4.5.2 Memorisation Onset . . . 26
4.6 CIFAR-10 Extension . . . 27

4.7 Summary of Results . . . 27

5 Discussion **29**
5.1 Interpretation of Main Findings . . . 29
5.2 Classifier Accuracy from Synthetic Data (RQ1) . . . 30
5.2.1 How Does Accuracy Compare With a Real-Data Baseline? . . . 30
5.2.2 How Does Accuracy Vary With Generator Training Length? . . . 31
5.3 The Plateau, Memorisation and Diversity (RQ2) . . . 31
5.3.1 Does the Classifier-Accuracy Plateau Occur Before the Onset of Memorisation? 31
5.3.2 How Does Classifier-Accuracy Relate to Changes in Sample Diversity? . . . 33
5.4 Scaling of the Two Timings (RQ3) . . . 34
5.4.1 The Timing of the Classifier-Accuracy Plateau . . . 34
5.4.2 The Timing of Memorisation Onset . . . 34
5.4.3 The Interval Between the Two Timings . . . 35
5.5 CIFAR-10 and Transfer beyond MNIST . . . 35
5.6 Threats to Validity . . . 36
5.6.1 Internal Validity . . . 36
5.6.2 Construct Validity . . . 36
5.6.3 Conclusion Validity . . . 37
5.6.4 External Validity . . . 37
5.7 Future Work . . . 37

6 Conclusion **39**

Bibliography **41**

A Software Repository **43**
A.1 Repository Location and Version . . . 43
A.2 Repository Organisation . . . 43
A.3 Results and Figure Pipeline . . . 43
A.4 Weights & Biases Provenance . . . 44
A.5 Excluded Artefacts and Reproducibility Scope . . . 44

B Full Experimental Results **45**

C Supplementary Measurement Controls **48**
C.1 Null Calibration of the Memorisation Criterion . . . 48
C.2 Sampler Control . . . 48
C.3 Validation-Loss and Baseline Checks . . . 49

D CIFAR-10 Extension **50**
D.1 Experimental Scope . . . 50
D.2 Results . . . 50
D.3 Interpretive Limit . . . 51

E Project Timeline **52**

List of Figures

4.1 Downstream classifier accuracy against generator training step for five of the eight arms (n = 200, 500, 1,000, 2,000 and 10,000), the remaining three omitted for legibility. The step-matched real-data baseline for each is shown as a dashed line in the same colour, and the range across generator seeds is shaded where more than one seed was run. . . . 23

4.2 Plateau accuracy against the step-matched real-data baseline at the same n, with one point per training-set size and classifier and the line of equality drawn. 24
4.3 Classifier Accuracy (Upper Panel) and memorised fraction (Lower Panel) against generator training step for all eight training-set sizes. The shaded band marks the 2,000 to 4,000 step classifier-accuracy plateau range, and vertical dotted lines mark observed memorisation onsets. Solid lines and filled markers denote arms with at least two generator seeds; dashed lines and open markers denote single-seeded arms. 25
4.4 Coverage as a share of its arithmetic ceiling and median nearest-neighbour distance, both against generator training step. 26
4.5 Memorisation onset and the operational classifier-accuracy plateau against training-set size on logarithmic axes. Measured onsets are shown as filled markers, arms that did not cross within the budget are represented as open markers with arrows, and the conditional relation permitted by the included arms is shown. 26

5.1 Generated samples from the $n = 1,000$ generator at generator training steps, 2,000, 16,000 and 64,000. 31

List of Tables

B.1 Complete arm-level results for the MNIST experiments. 46
B.2 Memorisation onset and maximum memorised fraction at the primary copying threshold and the two sensitivity thresholds. 47
B.3 Sensitivity of the operational plateau step to the selected accuracy threshold for the primary MLP results. 47
B.4 Individual-seed and arm-level memorisation onsets for the replicated MNIST arms. 47

C.1 Null calibration of the MNIST memorisation criterion. 48
C.2 Sampler comparison at $n = 200$ and generator step 32,000. 49

D.1 Configuration of the CIFAR-10 feasibility extension. 50
D.2 Complete summary of the CIFAR-10 extension. 51

Chapter 1 Introduction

1.1 Background & Motivation

Labelled datasets are needed to train image classifiers, but real data is often scarce, sensitive, or difficult to share. One alternative for real data is synthetic data, synthesised using a generative model trained to produce new examples resembling presently available data. In this study, a generator is trained on real images and used to create a synthetic dataset. A separate classifier is then trained only on those generated images and evaluated on a held-out set of real images that it has not seen during training. The subsequent accuracy of this classifier, hereafter referred to as classifier accuracy, measures how much class-relevant information the synthetic dataset preserves.

The generators examined here are diffusion models, a class of generative models that learn to create new images by reversing a gradual noising process [1]. During training, Gaussian noise is progressively added to real images until the result is nearly indistinguishable from a sample of pure noise. A neural network is then trained to predict and remove this noise, enabling new images to be generated by iteratively de-noising a random input. The main alternative class of image generators is the Generative Adversarial Network (GAN), trained by pitting a generator against a discriminator. This adversarial objective can be unstable, and GANs are susceptible to mode collapse where the generator only covers part of the data distribution and produces a narrow range of outputs regardless of its input. Optimising a de-noising objective avoids this adversarial dynamic, and diffusion models can match or exceed GAN sample quality on image synthesis [2]. Only diffusion models are used for this study.

A primary concern for image generation models is memorisation, since continued training exposes the generator to the same images repeatedly and can eventually cause the reproduction of samples from the original dataset [3]. This matters when synthetic data is intended to reduce direct access to that dataset, particularly because generated images can still train an accurate classifier while remaining unusually close to the generator's training examples meaning classifier accuracy and memorisation must be measured separately. Recent work shows that high quality generation and memorisation emerge at different points during training, with memorisation generally beginning later as the training set grows [3, 4]. Those findings indicate an early-stopping interval in which the model generates plausible images without detectable copying, but the interaction between classification accuracy and memorisation has not been measured directly.

A practical question is therefore how long a generator should be trained. If classifier accuracy reaches a plateau before memorisation begins, later optimisation adds computational cost and copying risk without improving the measured downstream outcome [4]. If the two events coincide, or if classifier accuracy continues to improve after the onset, the stopping decision is less clear. This dissertation measures both events at the same generator checkpoints and examines their timing changes with the number of real images used to train the generator.

Training-set size is the second variable of interest where small training sets provide fewer distinct examples and expose each image to the generator more frequently [5]. Larger training by contrast delay memorisation, but it has not been established whether they also change when classifier accuracy plateaus.

1.2 Aims & Objectives

The aim of this dissertation is to determine how classifier accuracy and memorisation change throughout diffusion-model training, and whether their temporal relationship depends on the size of the generator's training set.

The objectives are:

1. Train conditional diffusion models on controlled subsets of a real image dataset.
2. Generate a fixed quantity of class-balanced synthetic data at multiple checkpoints during each training run.
3. Measure the accuracy of classifiers trained exclusively on each synthetic dataset and evaluated on held-out real images.
4. Measure copying against the exact subset used to train each generator.
5. Identify and compare the classifier-accuracy plateau and memorisation onset.
6. Determine how the timing of both events changes with generator training-set size.

These objectives lead directly to three research questions concerning the accuracy trajectory, its relationship with memorisation, and the scaling of both event timings with training-set size.

1.3 Research Questions

This dissertation addresses the following research questions:

RQ1: How does the accuracy of a classifier trained solely on synthetic data and evaluated on real data compare with a real-data baseline, and how does this vary with diffusion-model training length?

RQ2: Does classifier accuracy reach its operational plateau before the onset of memorisation, and how does it relate to changes in sample diversity?

RQ3: How do the timings of a classifier's accuracy plateau and memorisation onset change as the diffusion model's training-set size varies?

1.4 Contributions

The principal contribution is a methodology measuring classifier accuracy and the memorised fraction at the same generator checkpoints. The individual measurements build on established approaches including Train-on-Synthetic, Test-on-Real evaluation [6, 7] and nearest-neighbour measurement of diffusion model memorisation [5, 3, 8]. Their combination makes it possible to compare the timing of a classifier's accuracy plateau directly with the onset of copying.

The empirical contribution is an examination of this relationship across eight MNIST training-set sizes from 200 to 10,000 images. The real-data baselines use the same source training-set condition and an approximately matched classifier optimisation budget. Each baseline is therefore matched to the generator on the quantity of real data and to the synthetic classifier on optimiser budget, which prevents differences in classifier training from being mistaken for properties of the generated data. What is not matched is the number of images the classifier itself receives: the synthetic classifier always receives 5,000 generated images, whereas the baseline receives n real ones. This asymmetry is measured and considered explicitly in Chapters 4 and 5. Together these controls distinguish changes caused by additional generator optimisation from changes caused by the quantity of real data available to the generator.

Finally, the study tests whether the measurement pipeline can be extended from MNIST to CIFAR-10. This extension is inconclusive with respect to memorisation, but identifies the model-capacity, training-budget and criterion validation requirements needed before the same comparison can reliably be made on more complex images. All empirical relationships reported here remain conditional on the selected generator architecture and its approximately 1.7-million-parameter capacity.

1.5 Report Structure

Chapter 2 provides context by reviewing conditional diffusion models, synthetic data evaluation, fidelity and diversity, and memorisation during generator training. Chapter 3 describes the experimental design, model architectures, operational definitions, controls and analysis procedure, before Chapter 4 reports the completed MNIST experiments in research question order and presents the CIFAR-10 extension. Chapter 5 then answers the research questions in relation to the previous literature and states the overall qualified conclusion.

Chapter 2 Literature Review

This chapter reviews the literature needed to examine how classifier accuracy and memorisation change during diffusion-model training. It first introduces conditional diffusion models and the evaluation of synthetic data through classifier accuracy. It then reviews fidelity, diversity, memorisation measurement and checkpoint selection before identifying the gap addressed by the three research questions.

2.1 Conditional Diffusion Models

The original De-noising Diffusion Probabilistic Models (DDPM) formulation established that a neural network trained to predict noise at different stages of the diffusion process could generate high-quality images through iterative de-noising [1]. Sampling starts from pure Gaussian noise, and the trained network predicts and removes a small amount of noise at each of many timesteps, gradually turning that noise into an image. Subsequent work extended this approach to class-conditional image generation, in which a class-conditional model receives a requested category during generation, and can therefore produce, for example, a specified digit as opposed to an unrestricted sample. This is enabled through mechanisms such as classifier guidance. Introduced by Dhariwal and Nichol, this uses gradients from a separately trained image classifier to steer the generation process towards a selected class, producing images that represent that class more clearly [2]. Ho and Salimans later built on this with classifier-free guidance, which achieves a similar effect without a separate classifier. The diffusion model instead learns conditional and unconditional predictions, which are combined during generation [9]. Both approaches allow the strength of class conditioning to be adjusted, but create a trade-off between more convincing individual images and the range of examples generated.

This trade-off between image fidelity and dataset diversity, a central concept for this dissertation and explored in more depth in Section 2.3, is particularly relevant when synthetic images are used for training a downstream classifier rather than visual inspection [7]. Guidance may improve perceptual quality but reduces distributional coverage available to a classifier [9]. The literature therefore establishes diffusion models as capable generators, while motivating direct measurement of how their training dynamics affect synthetic datasets used by classifiers.

2.2 Synthetic Data for Classification

A central question in synthetic data research is whether visually realistic synthetic images are also useful for training classifiers [7]. This distinction has led to criticism of traditional generative-model metrics, particularly Fréchet Inception Distance (FID), which measures similarity between real and generated image distributions in the feature space of an Inception network. FID evaluates perceptual realism rather than the classifier accuracy obtained from the generated dataset, motivating a separate measure of downstream performance.

Ravuri and Vinyals thus introduced the Classification Accuracy Score (CAS), which evaluates generative models by training a classifier solely on synthetic images before testing it on real data [7]. In class-conditional ImageNet generation, they found that synthetic-trained classifiers underperformed against real-trained classifiers, and models with better FID or Inception Scores were not able to consistently achieve higher CAS. These results demonstrate that conventional sample-quality metrics are poor predictors of classifier accuracy.

CAS implements the Train on Synthetic, Test on Real (TSTR) protocol, first introduced by Esteban et al. for synthetic medical time-series data generated by recurrent conditional GANs [6]. Its central premise is that a generator should be assessed by whether the data it produces

can support learning that generalises to previously unseen data. Movahedi et al. reinforced this principle, showing that testing on synthetic data can produce misleadingly optimistic performance estimates [10]. Although their work focuses on privacy-preserving data generation rather than diffusion models, it demonstrates that evaluation on a held-out real test set is the appropriate operational measure for the present classifier-training question.

Despite establishing CAS and TSTR as meaningful evaluation protocols, this body of literature has important limitations. Ravuri and Vinyals evaluated generative models during the GAN era, meaning their quantitative findings cannot be assumed to transfer directly to diffusion models, as they differ substantially in both training dynamics and sample generation [7]. Also, although their work established that fidelity metrics are poor predictors of classifier performance, the extent to which this conclusion holds across newer diffusion-based generators is incompletely understood.

These studies support direct evaluation through classifier accuracy, but do not establish how that accuracy changes across diffusion-model checkpoints. The relationship between diffusion-model training and classifier accuracy obtained from its generated data remains largely unexplored.

2.3 Fidelity, Diversity and Classifier Accuracy

Fidelity and diversity are two properties of a generated dataset. Fidelity describes the realism of individual samples, whereas diversity reflects how completely a generator captures the underlying distribution [11]. Although both properties influence sample quality, they can conflict as a generator might produce more convincing samples from only a restricted part of the distribution. For example, GANs are well known for producing highly realistic images while remaining susceptible to mode collapse, where only a subset of the true data distribution is represented. Diffusion models expose this trade-off more explicitly through guidance mechanisms, where increasing guidance strength typically improves perceptual quality at the expense of sample diversity [2, 9].

Evidence from classifier evaluation confirms that fidelity alone is insufficient, with Ravuri and Vinyals finding that neither FID nor Inception Score reliably predicts CAS [7]. Distributional coverage may therefore be important for classifier accuracy even when individual samples are more realistic.

Azizi et al. showed that synthetic images can improve classifier training by fine-tuning Imagen, a pre-trained diffusion model, to generate class-conditional ImageNet samples. They augmented real training data with synthetic images, improving the performance of both ResNet and Vision Transformer classifiers [12]. Their pipeline balanced fidelity and diversity through the choice of guidance and sampling parameters. Sariyildiz et al. [13] generated synthetic ImageNet datasets using Stable Diffusion, with classifiers trained on these datasets approaching real-data performance. however, their prompt engineering strategy increased sample diversity. On a smaller scale, Rather and Kumar demonstrated that GAN-generated MNIST images enabled lightweight convolutional networks to achieve performance close to real-data baselines [14], although this used a single generator configuration.

These settings differ from complete replacement of real data because the strongest results often use generators pre-trained on much larger external corpora, or use synthetic images to augment rather than replace real examples. Geng et al. instead propose an information ceiling, whereby a generator cannot create task-relevant information absent from its training data [15], while Fan et al. show diminishing returns as synthetic datasets grow [16]. At the furthest extreme, Shumailov et al. demonstrate that recursively training generative models on their own outputs progressively erodes the tails of the data distribution until model collapse occurs [17]. Together, these studies indicate that generating more observations does not necessarily add the diversity that would be supplied by additional real examples, as the generated samples cannot be novel within the present data distribution.

These findings are therefore not contradictory as synthetic augmentation, external pre-training and complete replacement reflect different experimental regimes. These papers do not provide

direct insight into a strict replacement setting where classifiers are trained exclusively on generated images or on the corresponding real data, without task-specific fine-tuning.

2.4 Training Dynamics and the Generalisation Window

Research on diffusion models has shifted attention from model capabilities to how generalisation and memorisation develop throughout training. Generalisation refers to learning shared patterns that allow new examples to be generated, whereas memorisation refers to reproducing particular members of the training set.

Gu et al. show that diffusion model memorisation is more pronounced with smaller datasets and depends on architecture and training procedure [5]. Yoon et al. distinguish unusually close reproduction from ordinary class similarity using the two closest images in the generator's training set [8]. A generated image satisfies their memorisation criterion when its distance to the nearest training image is less than one third of its distance to the second-nearest. The second-nearest distance thereby supplies a local reference for the nearest one. A low ratio indicates that the generated image is unusually close to one particular training example, rather than being similarly close to several examples from the same class. This relative comparison is also less sensitive to the size of the reference set than a single absolute distance threshold, because larger reference sets tend to produce smaller nearest-neighbour distances even when generator behaviour has not changed [8].

Bonnaire et al. provide a temporal account of diffusion-model generalisation and memorisation [3], identifying an early timescale, τ_{gen} at which sample quality emerges and a later timescale when memorisation becomes apparent, τ_{mem}. The former remains approximately constant as training-set size increases, whereas memorisation begins later. Their critical size $n_{\mathrm{gm}}(p)$ describes when that window opens for a model of capacity p, where n is the number of training examples. Below this size, memorisation has already begun by the time high-quality generation is achieved. Above it, the two events are separate, creating a period in which early stopping can retain sample quality while avoiding later memorisation.

They distinguish this from a second threshold, above which the dataset is too large for the model to memorise even after continued training. The first boundary concerns avoiding memorisation by stopping at an appropriate time, while the second concerns avoiding it because the model lacks sufficient capacity to reproduce the complete training set.

Favero et al. independently report a similar separation and scaling of memorisation across image and language diffusion models [4]. The agreement between these studies supports the view that memorisation develops as a distinct phase of diffusion-model optimisation whose timing depends on the amount of training data.

However, the early timescale of these studies is defined through generative quality using FID, while memorisation is measured through similarity to the training set [8]. Neither study trains a downstream classifier on samples produced at each generator checkpoint. This distinction is particularly important given Ravuri and Vinyals' findings mentioned in Section 2.3 demonstrating that improvements in conventional image quality metrics do not reliably predict classifier performance [7], making it plausible that classifier accuracy and memorisation follow different trajectories during training.

2.5 Checkpoint Selection

A checkpoint is a saved model state recorded during training, and evaluating multiple checkpoints gives insight into whether a final generator state is more useful than an earlier version. Lomurno and Matteucci provide a precedent for evaluating downstream classifier performance at multiple GAN checkpoints [18]. However, their work does not evaluate diffusion models or compare checkpoint selection with memorisation onset. As a result, it does not establish how classifier-accuracy behaves around the points of generalisation and memorisation identified in earlier research.

2.6 Summary and Research Gap

The literature establishes TSTR classifier accuracy as a direct measure of whether generated data supports learning on real examples. It also establishes that fidelity alone is insufficient and complete replacement of real data exposes information and diversity limitations that augmentation studies do not necessarily encounter. Meanwhile, research investigating synthetic data for classification has produced mixed conclusions. Under favourable conditions, including large-scale pre-training, augmentation of real datasets and evaluation under distribution shift, where the data encountered during testing differs from the data used for training, synthetic data can approach or even improve upon classifiers trained solely on real data. Separately diffusion training studies show that generative quality can emerge before memorisation, and that memorisation generally occurs later as training-set size increases [5, 3, 4, 19].

Three connected questions therefore remain unresolved: how classifier accuracy changes during generator training, whether its plateau precedes memorisation onset, and whether the timings of those events scale differently. These gaps motivate RQ1, RQ2 and RQ3 respectively in Section 1.3. Chapter 3 converts these questions into an experimental design in which classifier accuracy and memorised fraction are measured at the same generator checkpoints.

Chapter 3 Methodology

This chapter explains how generator training, memorisation and downstream classifier accuracy were measured. It first defines the experimental protocol and datasets, then describes the generator, sampling protocol and two measurement instruments. It concludes by defining the plateau and memorisation-onset events used in the results, documenting the controls and computational environment, and stating the limitations of the design.

3.1 Research Design & Rationale

The generators used in this study are trained from scratch under controlled conditions, as opposed to using existing sources, because two of the required measurements need direct access to the training process. Classifying a generated image as a copy requires distances measured against the generator's training set and against that set alone, so its exact membership must be recoverable at evaluation time [8]. Similarly, measuring how classifier accuracy responds to the onset of memorisation requires the training trajectory to be sampled at multiple points rather than observing a final, trained model. Although a small number of published generators release their training data and therefore satisfy the first requirement, checkpoint series are rarely distributed, meaning the second cannot be satisfied without training models directly. Consequently, training from scratch constrains the remainder of the design, as each generator must be sufficiently small to be trained many times over on a single consumer GPU. This is why the study considers training sets ranging from 200 to 10,000 images, and why its conclusions are stated at that scale.

Within this range, the primary independent variable is generator training length, recorded at between nine and fifteen checkpoints spanning 250 and 64,000 optimiser steps, with training-set size as a second independent variable. Each checkpoint yields two dependent measures:

1. The proportion of generated samples classified as copies
2. The accuracy of a classifier trained on those samples and evaluated on real data.

Generator capacity, subset composition, sampler, sample count and classifier training budget are held constant throughout.

Generator capacity is the most consequential of these controls; with the model fixed at approximately 1.7 million parameters, dividing model parameters by training-set size gives approximately $8,500$ training parameters per image at the smallest size, $n = 200$, and around 170 at the largest, $n = 10,000$. This ratio only describes how a fixed-capacity model is exposed to progressively more training data and is not treated as a direct measure of the capacity required to store an image. An alternative design would vary model width at a fixed training-set size, but width also changes the model's inductive bias and optimisation behaviour. Holding capacity constant therefore isolates training-set size, aligning more closely with the experimental protocol adopted by Bonnaire et al. and Favero et al. [3, 4]. The resulting scaling relationship is conditional on this approximately 1.7-million-parameter model and cannot establish whether the same relationship holds at other capacities.

The second dependent measure could have been a distributional metric, such as FID [20]. As discussed in Section 2.2 however, FID assesses how closely generated samples resemble the underlying data distribution rather than directly measuring if samples are useful for downstream learning as discussed in Chapter 2. Since the present question concerns whether synthetic data can substitute for real data during classifier training, the downstream task itself provides a more appropriate measurement [7]. The resulting accuracy remains a joint property of the synthetic dataset and the classifier using it for training, therefore, only comparisons where the classifier is held constant are meaningful, a constraint addressed in Section 3.10.

3.2 Experimental Procedure

Let D_{train} denote the complete real training dataset, and let $D_n \subset D_{\text{train}}$ denote the nested, class-balanced subset containing n images. For generator seed r, let $G_{n,r}^{(s)}$ denote the generator after s optimiser steps. Sampling this checkpoint produces the class-balanced synthetic dataset

$$S_{n,r}^{(s)} = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{m}, \qquad m = 5{,}000, \tag{3.1}$$

where $\tilde{x}_j$ is a generated image and $\tilde{y}_j$ is its conditioning label. The fundamental experimental unit is therefore the tuple (n, r, s), identifying a training-set size, generator seed and generator checkpoint.

Algorithm 1 Evaluation of a generator training trajectory

Require: Nested training subset D_n, generator seed r, checkpoint set $\mathcal{S}$, synthetic sample count $m = 5{,}000$

Train $G_{n,r}$ on D_n, saving its parameters at every $s \in \mathcal{S}$
for each checkpoint $s \in \mathcal{S}$ **do**
  Generate the balanced synthetic dataset $S_{n,r}^{(s)}$
  Compute memorisation and coverage statistics against D_n
  for each classifier architecture a **do**
    Train $f_{n,r,s,a}$ from scratch on $S_{n,r}^{(s)}$
    Evaluate $f_{n,r,s,a}$ on the real test set
    Record test accuracy as the downstream classifier accuracy
  end for
end for

Generator training and checkpoint evaluation are deliberately separated. Checkpoints and generated datasets are written to disk and reused when already available, allowing the evaluation stage to be repeated or extended without retraining the generator. This separation is necessary because individual generator runs require several hours and may be interrupted.

3.3 Dataset

MNIST was selected as the primary dataset for three reasons. It is a closed and finite corpus in which every training image is addressable by index, meaning the reference set used for memorisation measurement can be exactly the set of images a given generator was trained on, without ambiguity in defining what counts as a copy. Its dimensionality is sufficiently low, and its digits sufficiently centred and size-normalised, for Euclidean distance in pixel space to provide a reasonable approximation of perceptual similarity. This is a precondition for the criterion adopted in Section 3.6. Finally, a class-conditional diffusion model can be trained to convergence on MNIST repeatedly using a single consumer GPU, which the experimental design requires given the number of independent training runs involved.

The native resolution of 28 by 28 pixels constrains both network architectures. Because $28 = 7 \times 2^2$, the generator admits three resolution levels of 28, 14 and 7; further halving would require a non-integer dimension. The convolutional classifier applies two pooling operations, producing a flattened representation of $32 \times 7 \times 7$, while the MLP accepts 784 inputs. These values follow from the dataset rather than a tuning decision and are stated here so that the architectures described in Sections 3.4 and 3.7 are understood as constrained rather than freely chosen.

Training subsets are constructed once before any experimental run and stored as index files, so that the exact composition of each is fixed and recoverable. Each subset contains precisely $\frac{n}{10}$ images per class, since random sampling would produce noticeable imbalance at the smallest sizes considered. The subsets are nested prefixes of a single seeded permutation within each class. Increasing n therefore adds images without altering those already included, reducing subset composition as a

competing explanation for changes across training-set size. The validation set was drawn from positions beyond the largest subset in the same permutation and is therefore disjoint from every subset used to train a generator. This disjointness holds only up to the maximum n considered and would need to be re-established if larger subsets were examined.

3.4 Generator

The generator is a class-conditional DDPM built on a U-Net backbone, with channel widths of 32, 64 and 64 across its three resolution levels. Two residual layers are used at each level, and self-attention is applied only at the two lower resolutions; its cost grows quadratically with the number of spatial positions, so it is omitted at full resolution. The complete model contains approximately 1.7 million parameters. Class conditioning uses a learned four-dimensional embedding per class, which is broadcast spatially and concatenated to the image as additional input channels. This provides enough capacity to distinguish ten classes at negligible parameter cost, while allowing generated datasets to be exactly class-balanced by construction. A reference implementation was adopted rather than a bespoke architecture [21], keeping the noise-prediction formulation standard and reducing the number of design choices.

Training follows the simplified noise-prediction objective introduced by Ho et al. [1]. The forward diffusion process comprises 1,000 timesteps under a cosine noise schedule, with images represented in the range $[-1, 1]$ as the forward process assumes zero-mean Gaussian noise. The cosine schedule is preferred to a linear alternative because the latter destroys information too rapidly at low resolution, leaving the final portion of the trajectory close to pure noise and contributing little training signal. At each step, a timestep is sampled uniformly for every image, noise is added according to the schedule, and the mean squared error between predicted and true noise is minimised. Optimisation uses Adam at a constant learning rate of 10^{-3} with a batch size of 128, without a learning rate schedule, gradient clipping or an exponential moving average of weights. The absence of weight averaging departs from common diffusion practice and thus must be acknowledged as a limitation, although it primarily affects absolute sample quality rather than the relative changes across training length that this study measures.

For each training image x_0 with class label y, a diffusion timestep t is sampled uniformly and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is drawn. The corresponding noised image is

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \tag{3.2}$$

where $\bar{\alpha}_t$ is determined by the cosine noise schedule. The class-conditional network $\epsilon_\theta(x_t, t, y)$ is trained to recover the noise added to x_0 [1]. Its objective is

$$\mathcal{L}_{\mathrm{DDPM}}(\theta) = \mathbf{E}_{x_0, y, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, y)\|_2^2\right]. \tag{3.3}$$

Equation 3.3 teaches the model to predict the noise present at each diffusion timestep. During generation, this prediction is applied iteratively to transform an initial Gaussian sample into an image conditioned on the requested class.

Training length is measured in optimiser steps, following Bonnaire et al. [3]. One optimiser step is one parameter update and, as batch size is fixed, represents approximately the same amount of computation across training-set sizes. This makes checkpoint positions directly comparable. Epochs, used by Gu et al. [5], describe the average number of times each training image has been encountered. Because the number of optimiser steps per epoch increases with training-set size, expressing onset only in epochs would remove part of the dependence on n that Research Question 3 is intended to measure.

Held-out validation loss was recorded at every checkpoint using the disjoint validation set described in Section 3.3. The validation timesteps and noise were drawn once before training and used at every subsequent evaluation. Fixing these values makes validation loss a paired comparison across

checkpoints, ensuring observed changes reflect the generator rather than variation in the evaluation draw.

3.5 Sampling Protocol

Synthetic images are generated using a deterministic implicit sampler with 100 denoising steps, rather than the 1,000 timesteps used during training [22]. This reduction fits the design within the available computational budget. Generating 5,000 images at every checkpoint across multiple training-set sizes and seeds with full-length sampling would have consumed a substantial proportion of the available compute and reduced the number of checkpoints available for scaling analysis. Because the accelerated sampler approximates the full reverse process, its effect on reported measurements is assessed separately as a control in Section 3.8.

3.6 Measuring Memorisation

Memorisation is quantified using the nearest-neighbour ratio criterion introduced by Yoon et al. [8]. For each generated image, Euclidean distances are calculated to every image in the generator's training subset, and the two smallest distances are retained. A generated image is classified as a copy when the ratio of the nearest to the second-nearest distance falls below one third. The memorised fraction at a checkpoint is the proportion of generated images satisfying this condition when evaluated against the exact subset used to train the generator. The ratio is preferred to an absolute distance threshold because absolute distances are not directly comparable across training-set sizes. As the training set grows, it becomes denser in image space, so nearest-neighbour distances can decrease even without copying. An absolute threshold could therefore report an apparent increase in memorisation caused by reference-set density, leaving Research Question 3 difficult to interpret. The ratio asks whether a generated sample is unusually close to one particular training image relative to the local scale of its neighbourhood.

Distances are computed in raw pixel space, which is defensible on MNIST for the reasons expanded on in Section 3.3 and is consistent with the work this study extends, although it represents the principal limitation on the portability of measurement. For natural images, pixel distance is dominated by colour and low-frequency layout, and the nearest and second nearest distances concentrate as dimensionality increases, driving the ratio towards unity for most samples and leaving the criterion little scope to fire. Extensions to natural images could instead use membership-inference attacks [23], although applying these across many checkpoints would be computationally expensive. A more efficient alternative would be to calculate the same nearest-neighbour ratio in a learned feature space, such as the embedding produced by a self-supervised encoder, rather than directly in pixel space.

Let $D_n = \{(x_i, y_i)\}_{i=1}^{n}$ denote the generator's exact training subset. For generated image $\tilde{x}_j$, its distance to training image x_i is

$$d_{ij} = \|\tilde{x}_j - x_i\|_2 . \tag{3.4}$$

Let $d_{j,(1)}$ and $d_{j,(2)}$ denote the smallest and second-smallest distances for $\tilde{x}_j$. Its nearest-neighbour ratio is

$$\rho_j = \frac{d_{j,(1)}}{d_{j,(2)}}. \tag{3.5}$$

At ratio threshold k, the copy indicator is

$$C_j(k) = \mathbf{I}\left[\rho_j < k\right]. \tag{3.6}$$

where $\mathbf{I}(\cdot)$ equals one when its argument is true and zero otherwise. The memorised fraction at checkpoint s is therefore

$$M_{n,r}^{(s)}(k) = \frac{1}{m}\sum_{j=1}^{m} C_j(k), \qquad m = 5{,}000. \tag{3.7}$$

The primary analysis uses $k = \frac{1}{3}$, with $k = \frac{1}{4}$ and $k = \frac{1}{2}$ used for sensitivity analysis. This sensitivity analysis tests whether the conclusions depend on the selected threshold [8]. The median nearest-neighbour distance and median ratio are also recorded as continuous measures because thresholded memorisation may obscure gradual changes.

Finally, nearest-neighbour coverage measures how many distinct training images are represented among the generated samples. Let

$$a_j = \arg\min_{1 \le i \le n} d_{ij} \tag{3.8}$$

denote the training image nearest to generated image $\tilde{x}_j$. Raw coverage is defined as

$$C_{\text{raw}} = \frac{|\{a_j : 1 \le j \le m\}|}{m}, \tag{3.9}$$

with the arithmetic ceiling

$$C_{\max} = \min\left(1, \frac{n}{m}\right). \tag{3.10}$$

Raw coverage must be interpreted relative to this ceiling. For example, when $n = 200$ and $m = 5,000$ no more than $200/5,000 = 4\%$ of generated samples can have distinct nearest training neighbours. A raw coverage approaching this ceiling indicates that generated samples are distributed across much of the available training subset, whereas substantially lower coverage suggests concentration around a smaller number of training images. Coverage is therefore reported alongside its ceiling rather than interpreted in isolation.

As much of the analysis rests on this criterion, an automated test compares three cases against a controlled reference set of synthetic images drawn uniformly at random. These are near-duplicates made by adding small Gaussian perturbations, further images drawn independently from the same distribution, and the reference set compared with itself. The criterion must flag over 90% of the near-duplicates, under 5% of the independent images and all of the self-comparison, together establishing sensitivity, specificity and correct behaviour in the degenerate case. This synthetic test checks that the criterion behaves correctly by construction, while its observed false-positive behaviour on real data is assessed through the null calibration described in Section 3.8.

3.7 Measuring Classifier Accuracy

Classifier accuracy is measured under the TSTR protocol described in Section 2.2. A classifier is trained exclusively on generated images and evaluated on a real test set of 10,000 images, which is not used to train the generator or to calculate memorisation. This operationalises the downstream usefulness of generated data, following Esteban et al. [6] and Ravuri and Vinyals [7].

For classifier architecture a, its real-test accuracy is

$$\text{Acc}_{n,r,s,a} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{(x,y)\in\mathcal{D}_{\text{test}}} \mathbf{I}\left[\arg\max_c f_{n,r,s,a}(x)_c = y\right]. \tag{3.11}$$

Memorisation does not necessarily imply a loss of downstream classifier accuracy because reproduced training images remain valid examples of their respective classes, although accuracy may

still fall if copying is accompanied by reduced coverage of the training distribution. The analysis therefore compares the classifier-accuracy plateau with memorisation onset and examines accuracy after onset using two downstream classifier architectures. The first is a multilayer perceptron (MLP) with hidden layers of 256 and 128 units and ReLU activations, while the second is a convolutional neural network (CNN) comprising two convolutional and pooling stages followed by a fully connected layer with dropout. Their different inductive biases provide a robustness check on whether the observed trajectory is architecture-specific. The dropout parameter in the shared configuration applies only to the convolutional network, as the MLP is trained without dropout or another explicit form of regularisation.

Each synthetic classifier is trained on a synthetic dataset of 5,000 images for 10 epochs, using Adam with a learning rate of 10^{-3}, a batch size of 128 and cross-entropy objective. This gives a fixed budget of 400 optimiser steps per classifier.

3.8 Baselines & Controls

Three controls support the interpretation of the primary measurements, the first of which estimates the observed false-positive behaviour of the memorisation criterion. Real test images, which no generator is trained on and which are therefore novel by construction, are passed through the criterion as if they were generated samples and scored against each training subset. Any image classified under these conditions is a false positive, so repeating the calibration across reference-set sizes tests whether increasing reference density alone is sufficient to cross the memorisation threshold.

The second control provides the reference against which downstream accuracy is interpreted by training classifiers across several seeds on the same real subsets used by the generators. The resulting learning curve is matched to the generator's source subset rather than to the size of the synthetic set, allowing the study to ask whether a given quantity of real data is better used directly or used to train a generator that is then sampled. Without this curve, accuracy obtained from synthetic data could only be compared with other synthetic results, while a classifier trained on the complete set provides the corresponding upper reference. The synthetic classifier's optimisation budget is:

$$K_{\text{syn}} = 10 \left\lceil \frac{5000}{128} \right\rceil = 400 \tag{3.12}$$

optimiser steps. For a real subset containing n images, the baseline epoch budget is

$$E_{\text{real}}(n) = \text{round}\left(\frac{K_{\text{syn}}}{\lceil n/128 \rceil} \right). \tag{3.13}$$

The realised baseline budget is therefore

$$K_{\text{real}}(n) = E_{\text{real}}(n) \left\lceil \frac{n}{128} \right\rceil. \tag{3.14}$$

Rounding to complete epochs results in between 395 and 402 optimiser steps across the examined values of n.

The third control assesses the accelerated sampler. Both samplers are applied to the same checkpoint, selected where memorisation is most pronounced, and their memorised fractions and classifier accuracy are compared. Because the control covers one checkpoint at one training-set size, it indicates the direction and approximate magnitude of sampler-related bias rather than equivalence between the two samplers.

3.9 Event Definitions & Scaling Analysis

The classifier-accuracy plateau is defined separately for each training-set size and classifier architecture. First, a reference maximum is calculated for each arm, taken over only those checkpoints represented by the largest number of generator seeds available for that arm. The plateau step is the first checkpoint whose mean accuracy reaches at least 95% of that reference maximum. Plateau accuracy is reported as the mean accuracy over the final three observed checkpoints. This distinction prevents a late checkpoint represented by fewer seeds from moving the event threshold while still allowing the final level to summarise the completed trajectory.

Using the primary ratio threshold $k = \frac{1}{3}$, memorisation onset is defined using an onset-fraction threshold $\gamma = 0.05$:

$$s_{\mathrm{mem}}(n, r) = \min \left\{ s \in \mathcal{S} : M_{n,r}^{(s)}(k) \geq \gamma \right\}. \tag{3.15}$$

Because the generator is evaluated only at discrete checkpoints, a crossing first observed at s_ℓ locates onset within the interval

$$s_{\mathrm{mem}}(n, r) \in (s_{\ell-1}, s_\ell]. \tag{3.16}$$

If no crossing is observed, onset is right-censored beyond the final checkpoint s_L:

$$s_{\mathrm{mem}}(n, r) > s_L. \tag{3.17}$$

The relationship between training-set size and onset is represented as

$$s_{\mathrm{mem}}(n) = an^\beta, \tag{3.18}$$

or equivalently,

$$\log s_{\mathrm{mem}}(n) = \log a + \beta \log n. \tag{3.19}$$

The scaling parameters are constrained using the observed onset intervals rather than treating each first-crossing checkpoint as exact onset time. An onset first detected at checkpoint s_ℓ imposes the constraint

$$s_{\ell-1} < an^\beta \leq s_\ell. \tag{3.20}$$

A run with no observed onset instead imposes the right-censoring constraint

$$an^\beta > s_L. \tag{3.21}$$

Candidate values of a and β are retained only when they satisfy the constraints from all included runs. The reported range of β is therefore the range permitted by the checkpoint interval, rather than a regression confidence interval.

3.10 Reproducibility & Computational Environment

Three seeds are maintained separately, each controlling a distinct source of variation. The subset seed determines which images comprise each training subset and is held constant throughout so that any generator at a given set size sees identical data. The generator seed controls weight initialisation and batch ordering during generator training and is varied to assess replication. The classifier seed performs the equivalent role for the downstream classifier. A fourth fixed seed governs the validation noise and timesteps described in Section 3.4. Determinism is enforced through a single function that seeds the Python, NumPy and PyTorch random number streams, selects deterministic convolutional algorithms and disables algorithm auto-tuning, which would otherwise vary between runs on identical inputs. Data loaders are given explicitly seeded generator objects rather than relying on global state, making batch ordering reproducible. Disabling auto-tuning carries a throughput cost, which is accepted given that the study's central claims concern the location of a transition, rather than absolute training speed.

The MNIST experiments were conducted on a consumer NVIDIA RTX 3070 8GB. An NVIDIA Tesla T4 accessed via vLab was used for the CIFAR-10 Extension. Metrics were logged with a metric tracking service Weights & Biases (W&B).

3.11 Limitations

Several limitations follow from the design described above. Replication is uneven with three generator seeds completed at $n = 200$ and $n = 1,000$, two at $n = 350$, and one at each remaining size. Results from single-seeded arms cannot be separated from generator optimisation stochasticity. The generator also omits the exponential moving average of weights commonly used in diffusion training and its learning rate was inherited rather than tuned. These choices limit comparison with state-of-the-art generators. The MLP applies no explicit regularisation, and may therefore be especially sensitive to reductions in generated-data diversity.

Two limitations concern the memorisation measurement specifically. Although raw pixel distances are defensible for MNIST, they are not expected to transfer directly to natural images, while memorisation onset depends on both the conventional 5% threshold and the available checkpoint spacing. Individual-seed onsets for the replicated arms are therefore reported in Appendix B alongside the arm-level results.

More broadly, the primary findings rest on one dataset, one generator architecture at a single resolution, at one model capacity. These design limitations establish the scope of the claims; Chapter 5 considers the corresponding threats to validity to the observed results. Many of these limitations follow from the short duration of the project, and are therefore natural extensions for future work.

3.12 CIFAR-10

A parallel extension adapts the same pipeline to CIFAR-10, a dataset of natural colour images at 32×32 resolution. Changes were confined to image dimensions, channel counts and output namespaces, while the training procedure and pixel-space memorisation criterion were retained to test whether the existing measurement transferred beyond MNIST.

The intended validation was to compare known copies with novel CIFAR-10 images before interpreting generator results. The retained extension did not provide a complete positive-control result, so it is treated as a feasibility study rather than evidence that the MNIST findings generalise. This distinction is carried into Chapters 4 and 5.

Together, these methods define the accuracy plateau and memorisation onset used throughout the dissertation. Chapter 4 reports when they occur, how their separation changes with training-set size, and which conclusions remain supported by the controls.

Chapter 4 Results

This chapter reports the completed experiments in order of the three research questions. It first establishes the scope of available runs and validates the measurement instruments. It then reports the classifier-accuracy trajectories (RQ1), their timing relative to memorisation and coverage (RQ2), and the relationship between memorisation onset and training-set size. The final section reports the limited CIFAR-10 extension. Interpretation and threats to validity are reserved for Chapter 5.

4.1 Scope of the Completed Runs

The experimental programme described in Chapter 3 included eight training-set sizes: $n = 200$, 350, 500, 700, $1,000$, $2,000$, $5,000$ and $10,000$. Across these sizes, generators were evaluated at between nine and fifteen checkpoints spanning 250 to 64,000 optimiser steps. Table B.1 summarises the checkpoints completed for each experimental arm, together with the available replication, classifier accuracies and memorisation measurements.

The amount of replication varied across the training-set sizes. Three generator seeds were completed at $n = 200$ and $n = 1,000$, and two were completed at $n = 350$, while each of the remaining sizes was represented by one seed. Five arms continued to 64,000 steps. At $n = 350$, one generator reached 6,000 steps and the other reached 8,000, while the runs at $n = 500$ and $n = 700$ ended at 6,000 and 12,000 steps respectively. These shorter budgets were chosen in advance to continue beyond the memorisation onset expected at those sizes, before the classifier-accuracy trajectories were inspected. There were no other experimental reasons for ending an arm.

This uneven replication limits the comparisons that can be made. Seed-to-seed variation can be estimated at $n = 200$, $n = 350$ and $n = 1,000$, but results at the other sizes cannot be separated from variation caused by generator optimisation. A similar limitation applies to the plateau-accuracy summaries reported in Section 4.3, which use the mean accuracy from the final three checkpoints of each arm. Averaging these checkpoints avoids basing the summary on a single measurement, but all three came from the same generator seed at every training-set size and therefore do not represent independent replications.

Where multiple generator seeds were available, the mean observed range in accuracy between seeds was 4.5 percentage points, changes smaller than this may reflect ordinary differences between training runs rather than an effect of generator training length. The experiment can therefore rule out a large decline in accuracy but cannot distinguish smaller changes with confidence.

The MLP was evaluated at all eight training-set sizes, while the convolutional classifier was evaluated at $n = 200$ and $n = 1,000$. Memorisation is measured from the generator and its samples, so both classifier architectures use the same memorisation measurements when they are trained on samples from the same generator checkpoint.

4.2 Validation of Measurement Instruments

Before the primary results were interpreted, the controls described in Section 3.8 were used to examine the behaviour of the memorisation criterion, the accelerated DDIM sampler, the generator validation loss and the real-data classifier baselines.

4.2.1 Sensitivity and Specificity

The automated test described in Section 3.6 used a controlled reference set of 1,000 synthetic images drawn independently from a uniform distribution. Near-duplicates were created by adding

Gaussian noise with a standard deviation of 0.01 to 50 members of this set, while a further 50 images were drawn independently from the same distribution to represent novel samples. More than 90% of the near-duplicates were identified as copies, fewer than 5% of the independently drawn images were incorrectly identified, and every image was detected when the reference set was compared with itself. All three pre-specified conditions therefore passed.

These results show that the criterion behaves as expected under controlled conditions. Real images may behave differently because naturally similar examples occur and nearest-neighbour distances are smaller, so the false-positive behaviour on MNIST was examined separately through the null calibration.

4.2.2 False-Positive Calibration

The null calibration treated unseen MNIST test images as though they were generated samples and evaluated them against real training reference sets. Since the test images were not used to train a generator, every detection under these conditions would constitute a false positive.

At the primary ratio threshold of $k = \frac{1}{3}$, no detections were observed among the 10,000 probe images at any of the examined reference-set sizes: $n = 200$, $n = 1,000$, $n = 2,000$, $n = 5,000$, $n = 10,000$ and $n = 60,000$. The memorised fraction was therefore 0.000 in every case. The value at $n = 60,000$ uses the complete MNIST training set as an additional check of reference-set density and does not represent a generator training arm. Within this finite calibration, the observed false-positive result remains separated from the memorisation-onset threshold of $\gamma = 0.05$.

The continuous statistics recorded during the same calibration demonstrate the effect of reference-set density. The proportion of test images assigned to distinct nearest-neighbours rose from 0.020 at $n = 200$ to 0.865 at $n = 60,000$, while the median nearest-neighbour distance fell from 13.09 to 8.67. These changes occurred without generator involvement and therefore reflect the increasing density of the reference set.

4.2.3 Sampler Comparison

The 100-step accelerated sampler was compared with the full 1,000-step reverse process at $n = 200$ and generator step 32,000, where memorisation was already pronounced. Full sampling produced a downstream classifier accuracy of 0.793 and a memorised fraction of 0.999. Accelerated sampling produced an accuracy of 0.774 and a memorised fraction of 0.954.

4.2.4 Sampler Comparison

The continuous measurements recorded during the same calibration show how the density of the reference set affects the nearest-neighbour statistics even when no generator is involved. As the reference-set size increased from $n = 200$ to $n = 60,000$, the proportion of test images assigned to distinct nearest neighbours rose from 0.020 to 0.865, while the median nearest-neighbour distance fell from 13.09 to 8.67. These changes therefore reflect the increasing density of the reference set rather than memorisation.

4.2.5 Generator Validation Loss

Held-out validation loss was recorded using fixed validation noise and diffusion timesteps, as described in Section 3.4. Validation loss was available across all eight training-set sizes and 137 checkpoints. It declined throughout training at every size and showed no visible inflection at either the classifier-accuracy plateau or the onset of memorisation. Neither transition was therefore apparent from the generator's own optimisation objective.

4.2.6 Real-Data Baselines

Real-data classifiers were trained on the same nested subsets used to train the generators. Three classifier seeds were evaluated against each of the eight training-set sizes for both architectures,

using the step-matched optimisation budget described in Section 3.8.

For the multilayer perceptron, real-test accuracy increased from 0.819 at $n = 200$ to 0.920 at $n = 10,000$. For the convolutional classifier, it increased from 0.902 to 0.971 across the same range. Variation between classifier seeds was comparatively small, at $n = 200$, for example, multilayer perceptron accuracy ranged from 0.816 to 0.824. These baselines provide the real-data reference used in Section 4.3.

4.3 Research Question 1

Accuracy obtained from synthetic data rises steeply during early training, with the MLP reaching the operational plateau between 2,000 and 4,000 steps at all eight training-set sizes, as shown in Figure 4.1. The CNN trajectories at $n = 200$ and $n = 1,000$ show the same broad pattern, and although later checkpoints continue to show small changes, no large or consistent sustained decline was observed during any arm's completed budget.

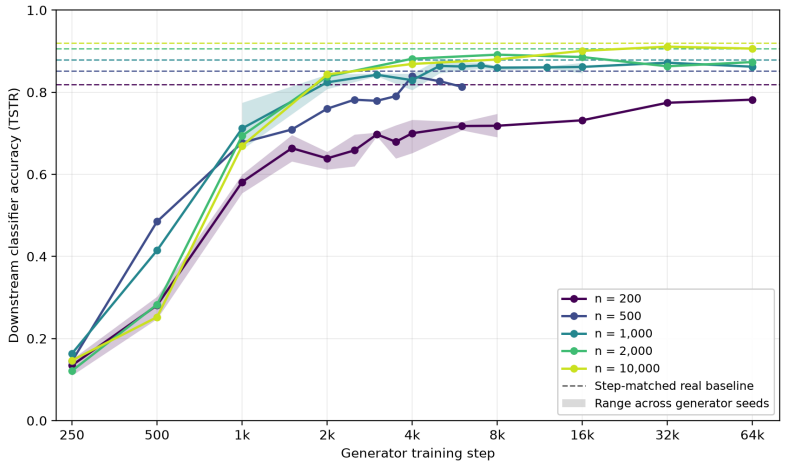


Figure 4.1: Downstream classifier accuracy against generator training step for five of the eight arms ($n = 200$, 500, 1,000, 2,000 and 10,000), the remaining three omitted for legibility. The step-matched real-data baseline for each is shown as a dashed line in the same colour, and the range across generator seeds is shaded where more than one seed was run.

Across the fifty-fold range in n, every operational plateau falls between steps 2,000 and 4,000 and therefore within a single checkpoint doubling. In the five arms continued to step 64,000, the additional training after step 4,000 altered accuracy by between one and three percentage points, which is the same order as the 4.5 point seed-to-seed range and therefore near the limit of what this design can resolve.

Figure 4.2 compares the plateau with the step-matched baseline, using the line of equality to separate the two possible outcomes. The shortfall ranges from 1.4 to 5.7 percentage points, corresponding to a ratio between 0.931 and 0.985 that rises from 0.931 at $n = 200$ to 0.984 at $n = 1,000$, before remaining between 0.965 and 0.985. The rise across the lower part of the range is larger than the observed seed-to-seed range, while the variation above $n = 1,000$ is not.

The baseline matches the quantity of source data available to the generator and the classifier optimisation budget, but it does not match downstream sample count because the synthetic classifier always sees 5,000 generated images while the real-data baseline sees n images. The synthetic arm is therefore trained on twenty-five times as many images as its baseline at $n = 200$ and on half as many at $n = 10,000$. Because the baselines use approximately 400 optimiser steps at every size, the baseline series also shows how real-data accuracy changes with sample count at a fixed budget. Moving from 200 to 5,000 real images is worth 9.90 percentage points, while moving from 5,000 to 10,000 is worth 0.20. The synthetic arm therefore holds a sample-count advantage at

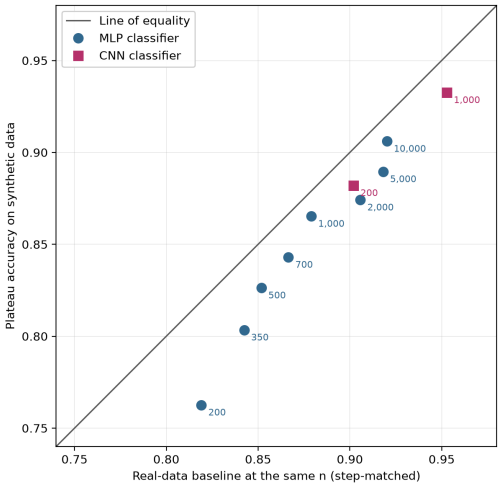


Figure 4.2: Plateau accuracy against the step-matched real-data baseline at the same n, with one point per training-set size and classifier and the line of equality drawn.

$n = 200$ and still falls 5.7 points short of its baseline, whereas at $n = 10,000$ the sample-count difference is worth at most 0.20 points along the observed real-data curve, which is within the 4.5 point range observed across generator seeds. Because the generated set was not varied against n, the reported ratios describe this fixed-5,000-synthetic design without separating the contribution of sample count from generated-sample diversity.

4.4 Research Question 2

4.4.1 The Relationship Between Utility and Memorisation

Under the primary 95% definition, there was no training-set size at which the operational accuracy plateau was detected after memorisation onset. Onset, defined as the first checkpoint where the arm-level memorised fraction reaches 0.05, was observed at six of eight sizes, as recorded in Appendix B. At $n = 5,000$ and $n = 10,000$, the memorised fraction did not reach the threshold within the 64,000 steps completed and those entries are lower bounds.

The interval between those two events is not constant. At $n = 200$ and $n = 500$, the plateau and onset were first detected at the same checkpoint, so their order cannot be resolved. The separation is 0.26 checkpoint doublings at $n = 350$. From $n = 700$ upward it is positive and grows with training-set size, at 1.22, 2.59 and 3.00 doublings for $n = 700$, $1,000$ and $2,000$, and at least 5.00 and 4.00 for $n = 5,000$ and $n = 10,000$. The plateau therefore precedes onset from $n = 700$ upward under the primary definition. At the 97.5% and 99% utility thresholds, the crossing occurs after memorisation onset at $n = 200$ and $n = 350$, as shown in Table B.3. The ordering of these small arms is therefore sensitive to the selected utility threshold.

Figure 4.3 shows downstream classifier accuracy and memorised fraction for all eight training-set sizes. The upper panel shows that classifier accuracy reaches its operational plateau within the shaded 2,000 to 4,000 step interval, while the lower panel shows when the memorised fraction reaches 0.05, with no consistent downward discontinuity in accuracy at memorisation onset. Among the shorter-budget arms, the net changes after onset were +4.80, −2.56 and +2.89 percentage points at $n = 350$, $n = 500$ and $n = 700$ respectively, so only the $n = 500$ arm declined over its observed post-onset interval. Across all arms, the change ranged from −2.56 to +8.5 percentage points, with the upper value arising from the late increase at $n = 200$ discussed in Section 5.3.1. The trajectories therefore show no consistent collapse when memorisation begins, although they do not support exact constancy either.

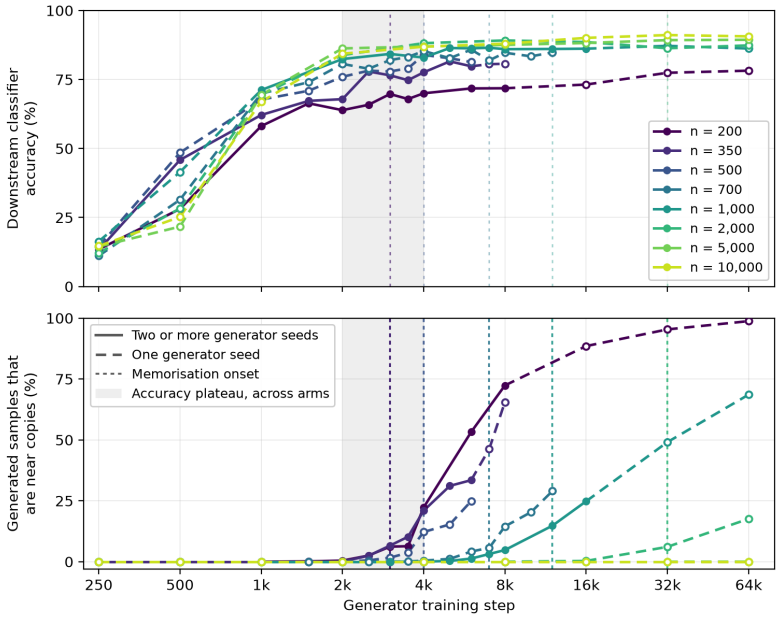


Figure 4.3: Classifier Accuracy (Upper Panel) and memorised fraction (Lower Panel) against generator training step for all eight training-set sizes. The shaded band marks the 2,000 to 4,000 step classifier-accuracy plateau range, and vertical dotted lines mark observed memorisation onsets. Solid lines and filled markers denote arms with at least two generator seeds; dashed lines and open markers denote single-seeded arms.

4.4.2 Diversity

The proportion of generated images whose nearest training neighbour is distinct was reported at every checkpoint alongside its arithmetic ceiling, the two being reported as a pair for the reasons given in Section 3.6. Figure 4.4 plots both diversity statistics against training step, with the classifier-accuracy plateau and the memorisation onset marked on each trace, so that the timing of any change in diversity can be compared directly against the timing of the other two transitions.

Coverage rises steeply during early training before levelling off without a large sustained decline, although the absence of a formal threshold means that this change is treated descriptively. The level reached varies with training-set size, remaining essentially complete at the smaller sizes, at 99.5% of the ceiling for $n = 200$ and comparably near-complete at $n = 500$, before falling to 82.2% at $n = 2,000$ and 56.6% at $n = 5,000$.

In the plotted trajectories, coverage generally begins to level off at or shortly before memorisation onset in the arms where both were observed, whereas the operational classifier-accuracy plateau occurs earlier while coverage is still rising. This descriptive pattern links the later change in coverage more closely with copying onset than with the earlier accuracy threshold. The median nearest-neighbour distance falls monotonically through training with no visible inflection at either transition. When compared with the separate null reference supplied for each training-set size in Section 4.2.2, samples generated at $n = 200$ end much closer to the training subset than unseen real images do, at 0.60 against 13.09, while the corresponding values of 9.60 and 9.73 at $n = 10,000$ are similar.

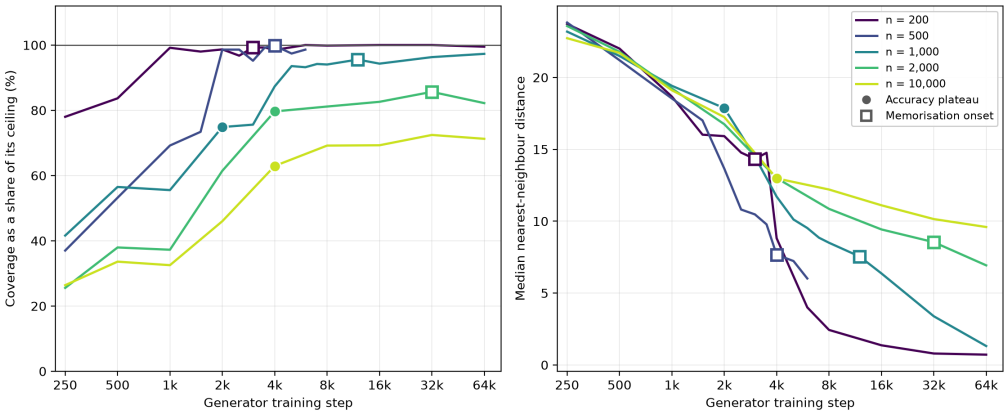


Figure 4.4: Coverage as a share of its arithmetic ceiling and median nearest-neighbour distance, both against generator training step.

4.5 Research Question 3

4.5.1 The Classifier Accuracy Plateau

The operational classifier-accuracy plateau does not vary systematically with training-set size, lying between steps 2,000 and 4,000 at every size examined and spanning a single checkpoint doubling across a fifty-fold range in n. It appears in Figure 4.5 as the flat lower series.

4.5.2 Memorisation Onset

Memorisation onset was observed at six of eight training-set sizes, while the two largest contribute lower bounds that are retained as constraints in the analysis described in Section 3.9. The exponent is therefore bounded rather than estimated through a conventional regression.

Each arm was expressed as an inequality on the relation onset $= a \cdot n^{\beta}$. An observed crossing at a given checkpoint constrains the onset to lie between that checkpoint and the preceding one, the grid being coarse; an arm that did not cross constrains it to lie beyond the last step completed. Sweeping the exponent and retaining those values for which some constant satisfies every inequality simultaneously gives the result shown in Figure 4.5.

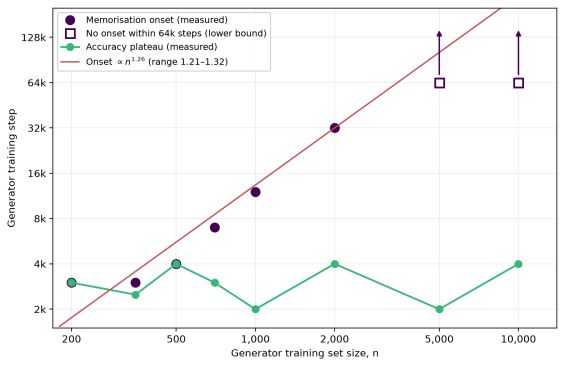


Figure 4.5: Memorisation onset and the operational classifier-accuracy plateau against training-set size on logarithmic axes. Measured onsets are shown as filled markers, arms that did not cross within the budget are represented as open markers with arrows, and the conditional relation permitted by the included arms is shown.

No relation of this form is consistent with all eight arms because the onset intervals at $n = 200$

and $n = 350$ overlap, while the remaining arms require onset to move later with training-set size. Excluding the smallest arm, the other seven admit relations with β between 1.21 and 1.32, a permitted range above unity within this constraint exercise. Five of these arms provide observed brackets and two provide only right-censored lower bounds.

Two qualifications attach to the interval. First, the exclusion of $n = 200$ is not a marginal disagreement. Its onset is bracketed between steps 2,500 and 3,000, overlapping the bracket at $n = 350$, so the two smallest arms provide no separation at the resolution of the grid. Second, the upper end rests on $n = 2,000$, whose crossing at step 32,000 is the largest observed onset. The two censored arms bound the relation from below but do not contain it from above.

A third qualification concerns the copying threshold itself. Repeating the procedure at the sensitivity thresholds gives β between 1.47 and 2.50 at $k = \frac{1}{2}$ and between 1.42 and 1.51 at $k = \frac{1}{4}$, with both analyses also requiring the $n = 350$ arm to be excluded. Although the permitted ranges remain above unity, their widths differ and the conditional constraint exercise does not provide a statistical test of a universal scaling law. The complete arm-level values are given in Table B.2.

4.6 CIFAR-10 Extension

The extension described in Section 3.12 was intended to follow a check of whether copies and novel images separate under the criterion at a 32 by 32 resolution. A complete positive-control result was not retained, so the extension is interpreted only as a feasibility study.

The generator trained on CIFAR-10 produced a maximum memorised fraction of 0.031 at $n = 200$ and 0.000 at $n = 1,000$ across all checkpoints. Applying the strictest of the sensitivity thresholds returned 0.012 and 0.000 respectively. Neither arm reached the 0.05 onset threshold, so no crossing between the operational accuracy plateau and memorisation was detectable.

Classifier accuracy was weak on the same arms. The highest accuracy attained was 0.207 at $n = 200$ and 0.232 at $n = 1,000$ for the multilayer perceptron, against architecture-matched real baselines of 0.275 and 0.350. The convolutional classifier attained 0.242 and 0.274, against baselines of 0.245 and 0.379. The corresponding full-dataset references are 0.533 and 0.676.

Two explanations are consistent with a memorisation fraction near zero. Either the criterion does not separate copies from novel images at this resolution, or these particular generators did not memorise.

4.7 Summary of Results

1. Accuracy from synthetic data rises steeply and reaches the 95% operational plateau between 2,000 and 4,000 steps at every n. No large or consistent sustained decline was observed during the completed budget of any arm.

2. Under the fixed-5,000-synthetic design, plateau accuracy reaches between 93.1% and 98.5% of the step-matched real-data reference, a shortfall of 1.4 to 5.7 percentage points. The baseline matches source quantity n and optimiser steps, while downstream sample count differs.

3. Under the primary utility threshold, the plateau and onset are first detected at the same checkpoint at $n = 200$ and $n = 500$, while the plateau precedes onset from $n = 700$ upward. At stricter utility thresholds, the ordering reverses at $n = 200$ and $n = 350$, so the small-arm ordering is threshold-sensitive.

4. As the proportion of copied samples rises from 5% to 99%, accuracy changes by between -2.6 and $+8.5$ percentage points without collapse.

5. Coverage rises and later levels off at a value that falls from essentially complete at $n = 200$ to 56.6% of its ceiling at $n = 5,000$. This descriptive change occurs near memorisation onset,

while the median nearest-neighbour distance falls monotonically with no visible inflection at either transition.

6. The operational plateau does not vary systematically with training-set size. Conditional on excluding $n = 200$, the remaining arms permit an exponent between 1.21 and 1.32 for memorisation onset. This is a descriptive constraint range rather than a confidence interval.
7. No detections were observed among 10,000 MNIST null probes at each reference-set size. The CIFAR-10 runs did not reach the 0.05 onset threshold, but the copying criterion was not fully validated for that dataset.

Chapter 5 Discussion

This chapter interprets the results in relation to the three research questions and the literature reviewed in Chapter 2. It first considers what the classifier-accuracy plateau means, then examines its temporal relationship with memorisation and the conditional scaling result. The CIFAR-10 extension is discussed separately as it did not transfer beyond MNIST. The chapter concludes with explicit threats to validity and future work targeted at those threats.

5.1 Interpretation of Main Findings

The central finding of this study is that, at the larger training-set sizes, synthetic data reaches the selected near-maximum accuracy threshold before the generator begins to memorise its training set. The operational classifier-accuracy plateau remained between approximately 2,000 and 4,000 optimiser steps across the examined values of n, while memorisation onset moved progressively later as the training set grew, suggesting that the two transitions measure different stages of training. At the smaller training-set sizes, whether the accuracy plateau occurs before or after memorisation onset depends on the threshold used to define the plateau.. Although the mechanism behind the broader pattern was not measured, one possible explanation is that the class-relevant structure a classifier can exploit is acquired early, while reproducing particular training images requires further optimisation. Distinguishing this explanation from alternatives would require an independent measure of the information content of the generated sets, which this design does not provide.

The three research questions address different parts of this result and are answered in turn in Sections 5.2 to 5.4. In outline, synthetic data supported between 93.1% and 98.5% of the classifier accuracy obtained by using the corresponding real data directly under the fixed-5,000-synthetic design, but did not exceed the real-data baseline. Second, under the primary 95% threshold, the classifier-accuracy plateau was not detected after memorisation onset. At $n = 200$ and $n = 500$, the two events were first detected at the same checkpoint, while from $n = 700$ upward the plateau occurred earlier. Third, the timing of the plateau remained approximately fixed as n increased, whereas memorisation onset moved later, producing an interval that widened with training-set size.

The absence of a large decline after memorisation onset is less surprising on reflection than originally anticipated because a copied training image remains a correctly labelled example and can continue to train a classifier effectively. The fact that classifier accuracy remains high after memorisation begins does not mean that copying is harmless. It shows only that, for the larger training sets, accuracy had already reached the defined plateau before copying was detected, and that further generator training produced no large or consistent improvement in classifier accuracy.

The contribution of the study lies in connecting two measurements ordinarily considered separately. Work on synthetic classifier training establishes whether generated data can support learning on real examples, while work on diffusion memorisation examines when generators begin reproducing their training data. By measuring both throughout the same training runs, this study identifies whether a generator has an interval in which a selected near-maximum accuracy threshold has been reached but copying has not yet been detected. The sections that follow interpret this result in relation to each research question, the generator's other measurements and the limitations of the experimental design.

5.2 Classifier Accuracy from Synthetic Data (RQ1)

Research Question 1 asks how the accuracy of a classifier trained solely on synthetic data and evaluated on real data compares with a real-data baseline, and how that accuracy varies with diffusion-model training length. The two clauses are answered separately below.

5.2.1 How Does Accuracy Compare With a Real-Data Baseline?

Synthetic data preserved most, but not all, of the classifier accuracy supported by the corresponding real subsets, with plateau accuracy reaching between 93.1% and 98.5% of the step-matched real baseline and falling short by between 1.4 and 5.7 percentage points. This is consistent with the purpose of TSTR evaluation, where generated samples retain much of the class-relevant structure that transfers to unseen real images without necessarily retaining all the information in the source data [6, 7]. The result also accords with evidence that synthetic data used alone often remains less effective than real data for supervised classifier training [16, 15].

This finding does not conflict with studies in which synthetic data improves a classifier's accuracy through augmentation [12], since the present experiment trains each classifier exclusively on either synthetic or real data. It instead asks whether passing a fixed real dataset through a generator provides a better classifier-training set than using those images directly, and no improvement was observed under that comparison.

The comparison reported above uses the step-matched baselines defined in Section 3.8; the reason an epoch-matched baseline would not have supported it is set out in Section 5.6. Using 5,000 synthetic images at every value of n means that once a conditional generator has been trained, further labelled samples can be produced without collecting or annotating additional real images. Whether this is worthwhile in practice depends on the cost of sampling relative to the cost of collecting and annotating further real images, which was not measured here and will vary by domain. The claim made is therefore only that the design permits the generated set to exceed n, not that doing so is necessarily economical. Holding the generated set at 5,000 images also ensures that changes in classifier accuracy across n are not caused by changing the number of synthetic examples supplied to the classifier.

The experiment consequently compares two uses of the same original n images. They may either be supplied directly to a classifier, or used to train a generator from which 5,000 classifier-training examples are then sampled. At $n = 200$ this gives the synthetic arm 25 times as many images as the real baseline. Accuracy nonetheless remains 5.7 percentage points lower which is consistent with the generated set failing to supply information absent from the generator's source data, although a single comparison at one training-set size cannot establish that as the cause. The shortfall could equally reflect sample quality at this scale as it is not possible for the synthetic classifier to have seen too few images, since it saw twenty-five times as many.

The direction of this imbalance reverses at $n = 10,000$, where the synthetic classifier receives half as many samples as the baseline, although the real baseline gains only 0.2 percentage points between $n = 5,000$ and $n = 10,000$, suggesting that its learning curve is nearly flat in this region. This result cannot tell us what accuracy the synthetic classifier would have achieved if the synthetic and real training sets had contained the same number of images. The fixed set of 5,000 generated images therefore still complicates the comparison across n. As the difference is also within the 4.5 percentage point range observed between generator seeds, the ratios cannot be attributed solely to generator training-set size.

The results are consistent with the information ceiling proposed by Geng et al. [15], under which a generator cannot create task-relevant information absent from its training data. They do not establish such a ceiling, since the absence of a synthetic advantage in this experiment cannot show that one is impossible elsewhere. Direct comparison with Ravuri and Vinyals [7] also requires care, their much larger synthetic deficit was measured on ImageNet using adversarial generators and against a classifier trained on the complete real dataset. The present comparison instead asks whether the particular n images available to a generator are better used directly or indirectly

through generation. The two experiments therefore answer different questions, although both find that synthetic data alone does not fully reproduce the classifier accuracy supported by real data.

5.2.2 How Does Accuracy Vary With Generator Training Length?

Classifier accuracy rose rapidly before reaching a plateau after which additional generator optimisation produced no large or consistent improvement. Among the arms trained from step 4,000 to step 64,000, a sixteen-fold increase in generator updates changed classifier accuracy by between one and three percentage points, which is smaller than the mean 4.5 percentage point range observed between generator seeds. Later training may still have a smaller effect than this design can resolve.

Figure 5.1 illustrates how this differs from the continued visual development of the samples. At the classifier-accuracy plateau, samples from the $n = 1,000$ generator retain visible noise and some are difficult to identify, yet they already support accuracy within 95% of the arm's reference maximum. By memorisation onset, the digits are clearer and more recognisable, but classifier accuracy has not improved correspondingly. The figure is illustrative rather than a quantitative measure of sample quality, but it demonstrates why visual improvement should not be assumed to imply an improvement in classifier accuracy.

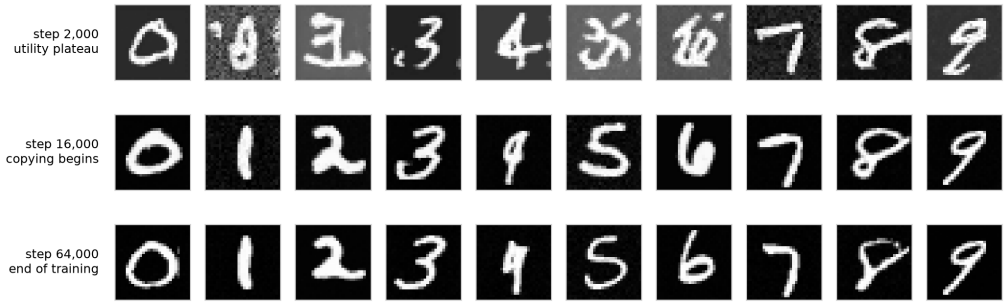


Figure 5.1: Generated samples from the $n = 1,000$ generator at generator training steps, 2,000, 16,000 and 64,000.

The conclusion for Research Question 1 is therefore that synthetic data preserves most of the classifier accuracy supported by its real source data, but does not improve upon direct use of that data under this design. It reaches the operational plateau early in generator training, and substantial additional optimisation provides no large or consistent improvement.

5.3 The Plateau, Memorisation and Diversity (RQ2)

Research Question 2 asks whether the classifier-accuracy plateau occurs before the onset of memorisation, and how classifier accuracy relates to changes in sample diversity. The two clauses are answered separately below.

5.3.1 Does the Classifier-Accuracy Plateau Occur Before the Onset of Memorisation?

The classifier-accuracy plateau was not detected after memorisation onset. The two events were first observed at the same checkpoint at $n = 200$ and $n = 500$, while their separation at $n = 350$ was smaller than one checkpoint doubling (0.26) and the plateau occurred earlier from $n = 700$ upward. At $n = 5,000$ and $n = 10,000$, classifier accuracy reached the operational plateau several checkpoint doublings before the end of training without the memorisation threshold being crossed. The ordering in the small arms is sensitive to the utility threshold because, at 97.5% and 99%, the accuracy crossing occurs after memorisation onset at $n = 200$ and $n = 350$.

The experiment does not identify an interval of near-maximum accuracy without detected copying at every training-set size. At $n = 200$ and $n = 500$, no such interval can be resolved from the available checkpoints. Each transition is located only within the interval between successive evaluations, so exact simultaneity cannot be established. Any separation at those sizes is smaller than the resolution of the experiment. At the larger sizes, the interval is wide enough to show that the operational classifier-accuracy plateau precedes the measured onset of memorisation.

This pattern is consistent with the observation by Gu et al. [5] that memorisation is more pronounced when diffusion models are trained on smaller datasets. The present results add a temporal interpretation in which memorisation begins early at small n, close to the point at which generated data first reaches the selected near-maximum accuracy threshold. When more training images are available, the classifier-accuracy plateau remains early while memorisation is delayed.

The temporal ordering is more important than classifier accuracy staying high once copying begins, as a copied training image remains a correctly labelled example and can still contribute to classifier training. Accuracy would therefore not necessarily decline merely because a growing proportion of the generated dataset consists of copies. At the larger training-set sizes, the additional optimisation associated with later copying provided no large or consistent improvement in classifier accuracy. Whether this is because the generated data already contained as much class-relevant information as the classifier could extract cannot be determined from these measurements, which record accuracy rather than information content, this hypothesis is consistent with the observed trajectories rather than a result.

The finding relates to the two training timescales identified by Bonnaire et al. [3], in which a diffusion model first learns to generate plausible samples and begins to memorise only after further training. The classifier accuracy plateau measured here should not be identified directly with their generalisation timescale, since Bonnaire et al. measure generative behaviour whereas the present study measures the performance of a separate classifier trained on generated samples. The results still support the same broader separation, while showing that it extends to a downstream measure that their work does not consider.

The continued accuracy observed after the onset does not directly contradict the memorisation/generalisation dichotomy proposed by Yoon et al. [8]. Their claim concerns whether the generator produces novel samples from the underlying distribution, whilst this measures whether and how well its output can train a classifier. A generated dataset may contain many copies and therefore generalise poorly as a generative model, while still containing recognisable and correctly labelled samples from which a classifier can learn. The two measurements answer different questions, which is precisely why classifier accuracy cannot be treated as evidence that a generator has not memorised.

This distinction is visible in the post-onset trajectories. As the memorised fraction rose from 5% to as high as 99%, classifier accuracy changed by between -2.6 and $+8.5$ percentage points and showed no consistent decline as shown in Figure 4.3, where the $+8.5$ percentage point change was an outlier at $n = 200$, not following the scaling relation identified in Section 4.5.2. The results therefore rule out a collapse in classifier accuracy as copying becomes dominant, but they do not establish that accuracy is exactly unchanged. Three shorter arms showed small declines over their final checkpoints, and the observed variation between generator seeds means that effects below approximately 4.5 percentage points cannot be distinguished confidently from run-to-run variation.

The principal practical contribution is the evaluation methodology itself. Existing classifier-accuracy and memorisation measures are ordinarily applied separately, often only to the final generator checkpoint. This study instead applies both throughout training, allowing an operational near-maximum accuracy threshold to be compared directly with the later onset of copying. The resulting framework can identify whether a generator has a training interval in which it has reached a selected accuracy threshold without yet satisfying the memorisation criterion.

In the present experiments, such an interval emerged clearly at the larger training-set sizes. A stopping point selected after the operational plateau but before memorisation onset would have

retained at least 95% of the comparable reference maximum while avoiding the copying detected later in training. Any stopping rule would remain specific to the dataset, architecture, checkpoint resolution and memorisation criterion. The methodology provides a way to identify a suitable point for a particular generative setting.

5.3.2 How Does Classifier-Accuracy Relate to Changes in Sample Diversity?

Classifier accuracy and sample diversity change at different points in generator training, and the observed diversity trajectory does not account for the plateau.

Sample diversity was measured as the proportion of generated images whose nearest training neighbour is distinct, reported against its arithmetic ceiling, together with the median nearest-neighbour distance. Both were recorded at the same checkpoints as classifier accuracy and the memorised fraction, so the three transitions can be ordered directly against one another.

The ordering is the substantive result. Coverage rose steeply during early training and generally began to level off at or shortly before memorisation onset in the arms where both were observed. The operational classifier-accuracy plateau occurred earlier while coverage was still rising. Since no formal coverage event was defined, this timing remains descriptive. It places the later change in coverage closer to copying onset than to the earlier accuracy threshold. This bears directly on the interpretation offered in Section 5.1. A reduction in diversity is one mechanism by which memorisation could have degraded classifier accuracy, since samples concentrated on a small number of training images would supply the classifier with fewer distinct examples. That mechanism is not clearly observed here. Accuracy had already reached the selected near-maximum threshold before coverage stopped increasing, so coverage does not explain the earlier flattening of the accuracy trajectory.

The coverage level reached falls as the training set grows, from essentially complete at $n = 200$ to 56.6% of its ceiling at $n = 5,000$, while plateau accuracy rises across the same range. The two quantities therefore move in opposite directions with n, which is not the pattern expected if incomplete coverage of the training subset were what limited downstream accuracy. This is an argument from the direction of an association across eight arms, not a controlled test, and coverage was not manipulated independently of n. The comparison is not symmetric across the range, however, because the ceiling binds at small n. When $n = 200$ and 5,000 images are generated, no more than 4% of samples can have distinct nearest neighbours, so near-complete coverage establishes only that copying is spread broadly across the subset rather than concentrated on a few images. At larger n, the ceiling no longer binds and the measured shortfall becomes informative about generator behaviour, but accuracy is higher at those sizes rather than lower.

The median nearest-neighbour distance falls monotonically throughout training with no visible inflection at either transition, and therefore does not mark either event. Its endpoints are nonetheless informative when read against the null calibration of Section 4.2.2, which supplies a separate reference distance for each reference-set size. At $n = 200$, the generated samples end at 0.60 against a null of 13.09, far closer to the training subset than unseen real images are. At $n = 10,000$, they end at 9.60 against a null of 9.73, which are similar values. The continuous statistic is consistent with the thresholded criterion about which generators are copying, while adding little about when copying begins.

Two qualifications apply to this answer. Nearest-neighbour coverage is a proxy for diversity rather than a distributional measure, as noted in Section 5.6. It records how many distinct training images are represented among the generated samples, not whether the generated distribution matches the underlying one, so a generator could spread its samples across the whole training subset while still failing to cover regions of the data distribution absent from that subset. Second, the downstream classifier supplying the complete series is an unregularised multilayer perceptron trained for a fixed budget, and may not be sensitive to diversity reductions that a different architecture would expose. Subject to those qualifications, the descriptive change in coverage and the operational classifier-accuracy plateau occur at different stages in these experiments, and the diversity statistics

recorded here do not account for the plateau.

5.4 Scaling of the Two Timings (RQ3)

Research Question 3 asks how the timings of the classifier-accuracy plateau and of memorisation onset change as the generator's training-set size varies. Each timing is considered separately below, followed by the interval between them.

5.4.1 The Timing of the Classifier-Accuracy Plateau

The plateau remained between approximately 2,000 and 4,000 optimiser steps across a fifty-fold range in training-set size, with every value lying within one checkpoint of every other and no systematic movement with n. Since the grid would not detect movement smaller than one checkpoint interval, the result supports only that the plateau moves less than memorisation onset at the scale resolved by this design, rather than exact invariance.

5.4.2 The Timing of Memorisation Onset

Memorisation onset, by contrast, occurred progressively later as n increased, which addresses a gap in how the earlier of the two transitions is defined. Bonnaire et al. [3] identify an early generalisation timescale that remains approximately constant with n, followed by a memorisation timescale that increases with it. They also define a critical training-set size $n_{\mathrm{gm}}(p)$, below which memorisation has already begun by the time high-quality generation is achieved, so that no generalisation window exists. Their earlier transition and this boundary are defined through the generative behaviour of the diffusion model, whereas the present study defines its earlier transition through the accuracy of a separate classifier trained exclusively on generated data. Under the primary 95% threshold, classifier accuracy and memorisation onset were first detected at the same checkpoint at $n = 200$ and $n = 500$, while a widening interval was observed at larger training-set sizes.

Previous work establishes that larger training sets delay diffusion model memorisation [5, 3], while work on synthetic classifier training establishes that generated data can preserve class-relevant information [7]. Neither establishes how the two timings move relative to one another. Because the operational plateau measured here does not move systematically with n while onset does, increasing the training-set size widens the interval in which the generator has reached a selected near-maximum accuracy threshold but has not yet crossed the memorisation threshold. The present experiment connects these findings by measuring classifier accuracy and memorisation at the same generator checkpoints and optimisation scale. Within the observed MNIST setting, the time required to reach the 95% threshold remains nearly fixed, while the additional training before copying begins grows with the size of the training set.

The scaling analysis gives a conditional description of memorisation onset. Each observed crossing was treated as an interval between successive checkpoints, while two arms that did not cross within 64,000 steps were retained as right-censored lower bounds. Across the seven arms from $n = 350$ upward, the constraints admit values of the exponent β between 1.21 and 1.32, in the relationship

$$s_{\mathrm{mem}}(n) = an^{\beta} \tag{5.1}$$

The range is determined by the checkpoint intervals and is descriptive rather than a regression confidence interval. Five of the seven arms provide observed onset brackets, while $n = 5,000$ and $n = 10,000$ provide lower bounds, and because most included arms use one generator seed, the range does not quantify run-to-run uncertainty.

No relationship within this exponent range is consistent with the $n = 200$ arm. The onsets at $n = 200$ and $n = 350$ were both first detected at step 3,000 and therefore lie within the same interval, $2,500,\ 3,000$. Across these two training-set sizes, onset can consequently increase by

a factor of no more than 1.2, whereas the larger arms require a steeper dependence on n. The disagreement is therefore not merely that the same checkpoint label was assigned to two different values of n; it is that their narrow and overlapping onset intervals are incompatible with the relationship admitted by the remaining arms.

Two interpretations of this disagreement remain possible. First, the checkpoint grid may be too coarse to resolve the relationship at the smallest training-set sizes, particularly as $n = 200$ and $n = 350$ differ by a factor of only 1.75. Alternatively, $n = 200$ may lie within a different low-data regime in which learning class structure and fitting individual examples occur too close together for the larger-n relationship to apply. The latter would be consistent with the critical boundary identified by Bonnaire et al. [3], although it would not establish that their approximately linear scaling relationship fails below that boundary. The available evidence cannot distinguish these explanations, so the bounded exponent should be treated as an empirical description of the arms from 350 upwards rather than as a universal scaling law.

Because the batch size was fixed, a linear relationship between memorisation onset and n would be numerically equivalent to onset occurring after approximately the same number of passes over each training set. Within the conditional constraint analysis, $\beta \in [1.21, 1.32]$ implies that the number of passes before memorisation increases slightly with n, since $s_{\mathrm{mem}}(n)/n \propto n^{\beta-1}$. The permitted range lies above the approximately linear relationship reported by Bonnaire et al. [3], but it does not provide a statistical test of that difference. The studies also use different datasets, architectures and experimental definitions.

The two right-censored arms limit what can be concluded beyond $n = 2,000$. For $n = 5,000$ and $n = 10,000$, the results establish only that memorisation onset occurs beyond 64,000 steps. Extending these runs until a crossing is observed would test whether the same scaling relationship continues at larger n. This uncertainty affects the estimated rate at which the interval widens, but not the observed temporal ordering, because both generators reached their classifier-accuracy plateaus well before the final checkpoint.

5.4.3 The Interval Between the Two Timings

The principal contribution of RQ3 is therefore the identification of an n-dependent interval defined jointly by classifier accuracy and memorisation. Bonnaire et al. [3] describe a generalisation window in which generative quality has been achieved but memorisation has not yet begun. The present results provide a downstream counterpart under the primary threshold, with the operational classifier-accuracy plateau remaining approximately fixed while memorisation onset moves later as the training set grows. The interval widens from one side and cannot be resolved at some of the smallest training-set sizes examined.

5.5 CIFAR-10 and Transfer beyond MNIST

The CIFAR-10 extension could not test the temporal relationship observed on MNIST because the measured memorised fraction remained below the selected threshold at every checkpoint. Without a located memorisation onset, no interval from the operational classifier-accuracy plateau could be measured and the relative timing of the two events could not be evaluated.

Two explanations are consistent with this outcome and the design cannot distinguish them, because the positive control was not retained. The first is that the generators did not memorise: CIFAR-10 presents a substantially more difficult generative problem than MNIST, the extension used a comparatively small model and a restricted training budget, and the weak downstream accuracy reported in Section 4.6 is consistent with generators that had not fitted the data closely. The second is that the criterion did not detect copying that had in fact occurred, for the reasons given in Section 3.6. Both remain hypotheses and the outcome should therefore not be interpreted as evidence that that diffusion models trained on CIFAR-10 do not memorise by the present criterion, or that the temporal relationship identified on MNIST does not transfer.

The raw-pixel nearest-neighbour criterion was not independently validated on CIFAR-10. Although appropriate for the centred and size-normalised MNIST images, pixel distance may be less informative for perceptual similarity in natural images. The necessary positive-control results were not retained, so the absence of detected copying cannot confidently be attributed to the behaviour of the criterion. This limits the conclusions available from the CIFAR-10 extension but does not affect the primary MNIST analysis.

The extension is therefore best understood as a feasibility result showing that transferring the full experiment to a more complex image domain requires greater model capacity and computational budget, as well as explicit validation of the memorisation criterion in that domain. A stronger extension would first demonstrate adequate generative and downstream performance, verify the criterion using known copies and non-copies, and then compare the operational classifier-accuracy plateau with memorisation onset.

5.6 Threats to Validity

5.6.1 Internal Validity

Generator replication is uneven, with multiple seeds available only at $n = 200$, $n = 350$ and $n = 1,000$, while every other arm reflects one optimisation trajectory. Differences between single-seed arms may therefore include generator stochasticity. The shorter $6,000 - 12,000$ step budgets at $n = 350$, $n = 500$ and $n = 700$ were chosen for early RQ3 runs because onset was expected within that interval, although they still limit observation of later behaviour.

The design is observational across training steps. Temporal ordering shows that one measurement occurs before the other, but it does not establish whether memorisation causes a change in classifier accuracy. Every explanation offered in this chapter for the ordering is therefore an interpretation of an association. The fixed subset composition reduces one source of confounding across n, but does not remove differences caused by stochastic optimisation.

One confound was identified and corrected before the primary analysis. The initial implementation trained both synthetic and real classifiers for ten epochs. Because the synthetic classifier always received 5,000 images while the baseline received n, this did not represent an equal amount of optimisation: at $n = 200$, ten epochs gave the real-data baseline approximately 20 optimiser updates against 400 for the synthetic classifier. The baselines were retrained for approximately 400 updates at every n, and all comparisons reported here use those step-matched results. Matching epochs is insufficient whenever the compared training sets differ in size, because it allows the optimisation applied to the classifier to vary with the condition being measured.

5.6.2 Construct Validity

Downstream classifier accuracy measures task-specific usefulness rather than general sample quality and is jointly determined by the generated dataset, classifier architecture and classifier training procedure [7]. Similarly, the pixel-space nearest-neighbour ratio operationalises copying on MNIST without directly measuring privacy harm, while coverage based on nearest training neighbours remains a limited proxy for diversity.

Both events depend on conventions, with the plateau defined as the first checkpoint reaching 95% of a comparable reference maximum and memorisation onset as the first observed checkpoint at which the memorised fraction reaches 0.05. Alternative thresholds can move either event, and although the null calibration supports the selected copying threshold on MNIST, it does not prove that every true copy is detected.

The RQ1 baseline matches the real data available to the generator and approximately matches classifier optimiser step, but it does not match downstream sample count. The fixed set of 5,000 generated images confounds sample quantity with generated-sample diversity across n, limiting causal interpretation of the synthetic to real accuracy ratio.

5.6.3 Conclusion Validity

Checkpoint observations interval-censor both event locations. Coincidence at $n = 200$ and $n = 500$ means only that the events cannot be ordered at the available resolution. The $n = 5,000$ and $n = 10,000$ onset values are right censored, so they constrain but do not locate onset. The exponent range is still a feasible interval under these constraints, not an uncertainty interval derived from repeated samples.

The observed 4.5 percentage point generator-seed range is descriptive and based on few replicated arms; it is not a confidence interval or universal detection threshold. Claims about small changes, especially at single-seed sizes, must therefore remain cautious. The $n = 200$ late increase illustrates this limitation directly.

5.6.4 External Validity

The primary evidence is limited to MNIST, one conditional diffusion architecture, one model capacity and training sets from 200 to 10,000 images. MNIST's centred, low-resolution digits make raw-pixel distance defensible and make generation and classification simpler than in natural-image domains. The fixed approximately 1.7-million-parameter generator means the scaling result cannot be separated from model capacity. The MLP supplies the complete eight-arm series; CNN results cover only two sizes. The inconclusive CIFAR-10 extension does not broaden this scope.

Despite these threats, two descriptive findings remain directly supported within the studied setting.

1. The 95% operational classifier-accuracy plateau occurs between 2,000 and 4,000 steps.
2. Under that primary threshold, the plateau is not detected after memorisation onset. The ordering of the smallest arms changes under stricter utility thresholds.

Stronger claims about causality, universal scaling or transfer to natural images are not supported.

5.7 Future Work

Three extensions would provide the most value to this work on synthetic data and memorisation, the most important of which is to repeat the experiment on a different dataset. MNIST provides a controlled setting, but its centred, low-resolution images make both generation and classification comparatively simple. A complete experiment on CIFAR-10 or another natural-image dataset would test whether the separation between the operational classifier-accuracy plateau and memorisation persists as the generative task becomes more difficult. This would require greater generator capacity and a longer training budget than were available for the present CIFAR-10 extension, as well as validation of the memorisation criterion using known copies and non-copies before its results are interpreted.

Extending the experiment to natural images requires the memorisation criterion to be validated before the sweep is run because, although raw-pixel distance is informative for centred, size-normalised MNIST digits, it is less reliable when perceptually similar images can differ substantially at pixel level and remains a poor measure of absolute privacy. If the nearest-neighbour ratio is not valid for CIFAR-10, membership inference could provide a stronger measure of privacy [23].

A second extension would vary the proportion of synthetic and real images used to train the downstream classifier. The present experiment compares classifiers trained on synthetic data with classifiers trained directly on the corresponding real data. Holding the total classifier-training set and optimisation budget constant while varying the synthetic-to-real ratio would show how generated data performs as a supplement. This would connect the synthetic-only Classification Accuracy Score used by Ravuri and Vinyals [7] with the augmentation setting studied by Azizi et al. [12]. Repeating this comparison across generator training-set sizes would also show whether the value of synthetic augmentation changes with the amount of real information available to the generator.

The third extension is to increase the temporal resolution around the two transitions. Additional checkpoints between approximately 2,000 and 4,000 steps would narrow the interval within which each event is located. This is particularly important at the smallest training-set sizes, where both events were first observed at the same checkpoint, and may therefore be separated by an interval too short for the current schedule to resolve. Longer runs at $n = 5,000$ and $n = 10,000$ would also replace the current right-censored observations with measured onset intervals, allowing the scaling of memorisation onset to be tested across the complete range.

Chapter 6 Conclusion

This dissertation examined how long a class-conditional diffusion model should be trained when its samples are intended to train a downstream classifier by comparing the point at which accuracy reached a selected near-maximum threshold with the onset of copying. Conditional DDPMs were trained from scratch on nested, class-balanced MNIST subsets of 200 to 10,000 images and evaluated at between nine and fifteen checkpoints spanning 250 to 64,000 optimiser steps. At every checkpoint, 5,000 class-balanced images were generated, a classifier was trained on those images alone and evaluated on the real test set, and copying was measured against the exact subset used to train that generator using a nearest-neighbour ratio criterion. Both quantities were therefore obtained from the same generator, samples and points in training, making their relative timing measurable.

In response to Research Question 1, synthetic data preserved most but not all of the classifier accuracy supported by the real images from which it was derived. Although the baseline matches the source quantity n and classifier optimisation budget, downstream sample count differs. Under this fixed-5,000-synthetic design, plateau accuracy reached between 93.1% and 98.5% of the real-data reference, a shortfall of 1.4 to 5.7 percentage points, and did not exceed that reference at any training-set size or for either classifier architecture. Even at $n = 200$, the synthetic classifier was trained on twenty-five times as many images as its baseline and still fell short, which is consistent with additional generated samples failing to supply information absent from the source data, although the experiment does not establish that mechanism. Accuracy rose steeply and reached the operational plateau between 2,000 and 4,000 steps, with no large or consistent sustained decline during any arm's completed budget. In the five arms continued from step 4,000 to step 64,000, accuracy changed by between one and three percentage points, less than the 4.5 percentage point range observed between generator seeds.

In response to Research Question 2, the classifier-accuracy plateau was not detected after memorisation onset under the primary 95% definition. The two transitions were first observed at the same checkpoint at $n = 200$ and $n = 500$, were separated by less than one checkpoint doubling at $n = 350$, and from $n = 700$ upward the plateau preceded onset by 1.22, 2.59 and 3.00 doublings. This ordering is threshold-sensitive in the smallest arms because the 97.5% and 99% accuracy crossings occur after onset at $n = 200$ and $n = 350$. As the memorised fraction subsequently rose from 5% to as high as 99%, accuracy changed by between -2.6 and $+8.5$ percentage points without collapsing or consistently deteriorating. Since a copied training image remains a correctly labelled example that can support classifier training, continued accuracy does not show that memorisation is harmless. The generator's own diagnostics did not mark either transition, as held-out validation loss declined throughout training with no visible inflection at either point and the median nearest-neighbour distance likewise fell monotonically. Coverage generally began to level off near memorisation onset, although this remained a descriptive pattern rather than a formally defined event.

In response to Research Question 3, the two timings show different descriptive patterns as the amount of real data available to the generator increases. The operational accuracy plateau remained between approximately 2,000 and 4,000 optimiser steps across a fifty-fold range in n, without systematic movement, while memorisation onset moved progressively later. Across the seven arms from $n = 350$ upward, the checkpoint intervals and censored bounds permit a relation of the form given in Equation 5.1, with an exponent between 1.21 and 1.32. This conditional range is not a confidence interval, most included arms use one generator seed, and no relation of this form accommodates $n = 200$. In the observed arms, the interval between the two transitions therefore widens with training-set size because memorisation moves later.

Taken together, these results give a downstream counterpart to the generalisation window described in recent work on diffusion training dynamics. That window is ordinarily defined through

generative quality, while the earlier transition measured here uses the accuracy of a separate classifier trained exclusively on generated samples. Samples at the operational plateau retain visible noise and are sometimes difficult to identify, yet already support accuracy within 95% of the arm's reference maximum. Evaluating classifier accuracy and copying at the same checkpoints therefore provides a practical way to identify whether a generator has reached a selected level of downstream utility before satisfying the memorisation criterion, an interval that was wide at the larger training-set sizes in these experiments and unresolvable at some of the smallest.

The conclusions of this experiment are bounded by the design that produced them, resting on a single dataset at one resolution, one generator at a single capacity, a training-set range small relative to practical use and uneven replication across arms.

To conclude, the simplified core findings of this dissertation are:

1. In the studied MNIST setting, synthetic-data accuracy reaches the 95% operational plateau early, and later optimisation provides no large or consistent downstream improvement.
2. Under the primary threshold, the observed interval between that plateau and memorisation generally widens with training-set size. The scaling description remains conditional on sparse and unevenly replicated arms.

Bibliography

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. version: 1. June 19, 2020. DOI: 10.48550/arXiv.2006.11239. arXiv: 2006.11239[cs.LG]. URL: http://arxiv.org/abs/2006.11239.

[2] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. June 1, 2021. DOI: 10.48550/arXiv.2105.05233. arXiv: 2105.05233[cs.LG]. URL: http://arxiv.org/abs/2105.05233.

[3] Tony Bonnaire et al. *Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training*. Oct. 28, 2025. DOI: 10.48550/arXiv.2505.17638. arXiv: 2505.17638[cs]. URL: http://arxiv.org/abs/2505.17638.

[4] Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. *Bigger Isn't Always Memorizing: Early Stopping Overparameterized Diffusion Models*. Sept. 1, 2025. DOI: 10.48550/arXiv.2505.16959. arXiv: 2505.16959[cs]. URL: http://arxiv.org/abs/2505.16959.

[5] Xiangming Gu et al. *On Memorization in Diffusion Models*. Feb. 20, 2025. DOI: 10.48550/arXiv.2310.02664. arXiv: 2310.02664[cs]. URL: http://arxiv.org/abs/2310.02664.

[6] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. *Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs*. Dec. 4, 2017. DOI: 10.48550/arXiv.1706.02633. arXiv: 1706.02633[stat]. URL: http://arxiv.org/abs/1706.02633.

[7] Suman Ravuri and Oriol Vinyals. *Classification Accuracy Score for Conditional Generative Models*. Oct. 28, 2019. DOI: 10.48550/arXiv.1905.10887. arXiv: 1905.10887[cs]. URL: http://arxiv.org/abs/1905.10887.

[8] TaeHo Yoon et al. "Diffusion Probabilistic Models Generalize when They Fail to Memorize". In: ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling. July 28, 2023. URL: https://openreview.net/forum?id=shciCbSk9h#all.

[9] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. July 26, 2022. DOI: 10.48550/arXiv.2207.12598. arXiv: 2207.12598[cs.LG]. URL: http://arxiv.org/abs/2207.12598.

[10] Parisa Movahedi et al. "Benchmarking Evaluation Protocols for Classifiers Trained on Differentially Private Synthetic Data". In: *IEEE Access* 12 (2024), pp. 118637–118648. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3446913. URL: https://ieeexplore.ieee.org/document/10643135.

[11] Muhammad Ferjad Naeem et al. *Reliable Fidelity and Diversity Metrics for Generative Models*. 2020. arXiv: 2002.09797 [cs.CV]. URL: https://arxiv.org/abs/2002.09797.

[12] Shekoofeh Azizi et al. *Synthetic Data from Diffusion Models Improves ImageNet Classification*. Apr. 17, 2023. DOI: 10.48550/arXiv.2304.08466. arXiv: 2304.08466[cs]. URL: http://arxiv.org/abs/2304.08466.

[13] Mert Bulent Sariyildiz et al. *Fake it till you make it: Learning transferable representations from synthetic ImageNet clones*. Mar. 28, 2023. DOI: 10.48550/arXiv.2212.08420. arXiv: 2212.08420[cs]. URL: http://arxiv.org/abs/2212.08420.

[14] Ishfaq Hussain Rather and Sushil Kumar. "Generative adversarial network based synthetic data training model for lightweight convolutional neural networks". In: *Multimedia Tools and Applications* 83.2 (Jan. 1, 2024), pp. 6249–6271. ISSN: 1573-7721. DOI: 10.1007/s11042-023-15747-6. URL: https://doi.org/10.1007/s11042-023-15747-6.

[15] Scott Geng et al. *The Unmet Promise of Synthetic Training Images: Using Retrieved Real Images Performs Better*. Jan. 2, 2025. DOI: 10.48550/arXiv.2406.05184. arXiv: 2406.05184[cs]. URL: http://arxiv.org/abs/2406.05184.

[16] Lijie Fan et al. *Scaling Laws of Synthetic Images for Model Training ... for Now*. Dec. 7, 2023. DOI: 10.48550/arXiv.2312.04567. arXiv: 2312.04567[cs]. URL: http://arxiv.org/abs/2312.04567.

[17] Ilia Shumailov et al. "AI models collapse when trained on recursively generated data". In: *Nature* 631.8022 (July 2024), pp. 755–759. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07566-y. URL: https://www.nature.com/articles/s41586-024-07566-y.

[18] Eugenio Lomurno and Matteo Matteucci. *Synthetic Image Learning: Preserving Performance and Preventing Membership Inference Attacks*. July 30, 2024. DOI: 10.48550/arXiv.2407.15526. arXiv: 2407.15526[cs]. URL: http://arxiv.org/abs/2407.15526.

[19] Alessandro Favero et al. *How Compositional Generalization and Creativity Improve as Diffusion Models are Trained*. June 4, 2025. DOI: 10.48550/arXiv.2502.12089. arXiv: 2502.12089[stat]. URL: http://arxiv.org/abs/2502.12089.

[20] Martin Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. Jan. 12, 2018. DOI: 10.48550/arXiv.1706.08500. arXiv: 1706.08500[cs]. URL: http://arxiv.org/abs/1706.08500.

[21] Patrick von Platen et al. *Diffusers: State-of-the-art diffusion models*. Version 0.12.1. URL: https://github.com/huggingface/diffusers.

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. Oct. 5, 2022. DOI: 10.48550/arXiv.2010.02502. arXiv: 2010.02502[cs]. URL: http://arxiv.org/abs/2010.02502.

[23] Hailong Hu and Jun Pang. *Membership Inference of Diffusion Models*. Jan. 24, 2023. DOI: 10.48550/arXiv.2301.09956. arXiv: 2301.09956[cs.CR]. URL: http://arxiv.org/abs/2301.09956.

Chapter A Software Repository

This appendix documents the software and results repository associated with the dissertation. Its purpose is to identify the version of the code used in the study, explain how the repository is organised, and state which parts of the analysis can be reproduced from the committed files. Experimental results and their interpretation remain reported in the main dissertation; the repository provides the corresponding implementation and computational artefacts.

A.1 Repository Location and Version

The repository is hosted on the University of Birmingham GitLab service at:

`https://git.cs.bham.ac.uk/projects-2025-26/oxg522`

A.2 Repository Organisation

The committed files are divided according to their role in the experimental and analysis pipelines:

`Experiments/` Contains the generator, classifier, data-loading and memorisation code. The principal generator is `CCUNETGenerator.py`. The MLP and CNN implementations are in the `Classifiers/` subdirectory, and the CIFAR-10 extension is in `CIFAR10/`.

`Results/processed/` Contains the analysis-ready tables used to calculate the values reported in the dissertation. These include checkpoint-level measurements, step-matched real-data baselines, classifier-accuracy plateaus, memorisation onsets and the complete arm-level table reported in Appendix B.

`Results/figures/` Contains the six figures used in the final dissertation. The files are stored as PNG images because figure captions and numbering are supplied by LaTeX.

`Results/scripts/` Contains the scripts that transform the retained measurements into processed tables and dissertation figures. Paths are resolved relative to the repository rather than to a specific computer.

`Results/raw/` Contains retained direct experiment output. In the referenced commit, the raw CIFAR-10 W&B export is included, but the complete raw MNIST W&B history is not committed.

`archive/` Contains superseded, exploratory or historically useful material. Files in this directory are not part of the final evidential pipeline unless they are explicitly discussed in the dissertation.

Only the files under `Results/processed/` feed the final plotting and summary scripts; similarly named files under `archive/` preserve the development history but should not be substituted into the final analysis.

A.3 Results and Figure Pipeline

Python dependencies are listed in `requirements.txt`. From the repository root they can be installed with:

```
python -m pip install -r requirements.txt
```

The main entry point for the retained analysis is:

```
python Results/scripts/make_all_figures.py
```

This command first runs `calculate_summaries.py`, which regenerates the classifier-accuracy plateau, memorisation-onset, step-matched baseline and arm-level summary tables. It then invokes the six plotting modules in dissertation order and writes their outputs to `Results/figures/`. Shared thresholds and aggregation rules are defined in `figure_data.py`, while shared plotting conventions and output behaviour are defined in `figure_style.py`. Centralising these choices reduces the risk that figures apply different plateau definitions, memorisation thresholds or baseline conditions.

Figure 5.1 additionally uses `Results/processed/generated_sample_images.npz`. This file contains the extracted image samples needed by the figure rather than the complete generated datasets. The extraction procedure is implemented in `extract_sample_images.py`.

A.4 Weights & Biases Provenance

The script `Results/scripts/pull_wandb.py` records how checkpoint histories are retrieved from the MNIST and CIFAR-10 W&B projects. It maps the logged histories into raw CSV files and records a pull manifest. Running it requires access to the relevant W&B entity and projects. Authentication can be supplied through an ignored repository-root `.env` file or through an existing `wandb login` session. Credentials are not committed.

The relevant command is:

```
python Results/scripts/pull_wandb.py
```

The referenced commit contains the processed MNIST tables used by the final analysis but not the complete raw MNIST W&B export. Consequently, the final figures and reported summaries can be regenerated from the committed processed measurements, but reconstructing those processed MNIST measurements from the original run histories requires authorised W&B access. This is a provenance limitation rather than an unreported substitution of data: the processed inputs used by the final scripts are versioned directly.

A.5 Excluded Artefacts and Reproducibility Scope

Large or machine-specific artefacts are deliberately excluded through `.gitignore`. These include model checkpoints, complete generated datasets, downloaded MNIST and CIFAR-10 datasets, local W&B run directories, virtual environments, log files and API credentials. These exclusions keep the repository focused on source code and the evidence required for the reported analysis, but they mean that the original trained generators cannot be reconstructed from the repository alone.

The dependency file lists the required Python packages without fixing exact version numbers. It supports recreation of the software environment in broad terms but does not provide a bit-for-bit environment lock. Re-running model training may also depend on hardware, library versions and access to the original subset indices and W&B records. The strongest reproducibility claim supported by the committed repository is therefore that the retained processed measurements can be traced through the analysis scripts to the reported summary tables and figures. Full independent reproduction from raw training through to every reported result would require the excluded checkpoints or regenerated runs together with access to the original W&B histories.

Chapter B Full Experimental Results

This appendix reports the complete arm-level results underlying the analyses presented in Chapter 4. The main chapter reports the quantities required to answer each research question, while Table B.1 provides the corresponding values for every training-set size.

Table B.1: Complete arm-level results for the MNIST experiments.

n	Seeds	Final step	Plateau step	Plateau accuracy (%)	Baseline accuracy (%)	Ratio to baseline	Memorisation onset	Maximum memorised fraction (%)	Coverage at step 4,000 (%)
200	3	64,000	3,000	76.26	81.91	0.9310	3,000	98.76	98.50
350	2	8,000[a]	2,500	80.34	84.24	0.9537	3,000	65.54	99.29
500	1	6,000	4,000	82.63	85.18	0.9701	4,000	24.90	99.80
700	1	12,000	3,000	84.30	86.65	0.9728	7,000	29.02	92.86
1,000	3	64,000	2,000	86.52	87.89	0.9844	12,000	68.58	87.33
2,000	1	64,000	4,000	87.41	90.57	0.9651	32,000	17.60	79.65
5,000	1	64,000	2,000	88.95	91.81	0.9689	$> 64{,}000$	0.02	50.38
10,000	1	64,000	4,000	90.60	92.01	0.9847	$> 64{,}000$	0.00	62.94

Notes: Plateau step is the first observed checkpoint satisfying the 95% operational plateau rule. Memorisation onset is the first observed checkpoint at which at least 5% of generated images satisfy the copying criterion. Entries shown as $> 64{,}000$ are right-censored lower bounds: onset was not observed within the completed training budget. Accuracy and memorised-fraction values are percentages; the ratio to baseline is dimensionless. Coverage is reported specifically at generator training step 4,000. [a] One of the two generators at $n = 350$ reached step 8,000 and the other reached step 6,000.

Table B.2: Memorisation onset and maximum memorised fraction at the primary copying threshold and the two sensitivity thresholds.

	$k = \frac{1}{2}$		$k = \frac{1}{3}$ (primary)		$k = \frac{1}{4}$	
n	Onset	Max (%)	Onset	Max (%)	Onset	Max (%)
200	2,500	99.16	3,000	98.76	4,000	98.38
350	2,500	77.10	3,000	65.54	4,000	52.66
500	2,500	43.02	4,000	24.90	5,000	12.20
700	5,000	41.40	7,000	29.02	8,000	20.70
1,000	6,000	74.96	12,000	68.58	12,000	64.32
2,000	32,000	26.72	32,000	17.60	64,000	11.48
5,000	$> 64{,}000$	0.44	$> 64{,}000$	0.02	$> 64{,}000$	0.00
10,000	$> 64{,}000$	0.02	$> 64{,}000$	0.00	$> 64{,}000$	0.00

Notes: Onset is the first observed checkpoint at which at least 5% of generated images satisfy the copying criterion at the stated ratio threshold; entries shown as $> 64{,}000$ are right-censored lower bounds. Maximum memorised fraction is the largest seed-averaged value observed within the completed budget. The $k = \frac{1}{3}$ columns reproduce the corresponding entries of Table B.1. Exponent bounds derived from these onsets are reported in Section 4.5.2.

Table B.3: Sensitivity of the operational plateau step to the selected accuracy threshold for the primary MLP results.

n	90%	95% (primary)	97.5%	99%
200	1,500	3,000	6,000	6,000
350	2,500	2,500	5,000	5,000
500	2,000	4,000	4,000	4,000
700	2,000	3,000	4,000	4,000
1,000	2,000	2,000	3,000	5,000
2,000	2,000	4,000	4,000	8,000
5,000	2,000	2,000	8,000	32,000
10,000	2,000	4,000	16,000	32,000

Notes: Each entry is the first checkpoint whose comparable-seed mean accuracy reaches the stated share of the arm's reference maximum. The primary analysis uses 95%. At 97.5% and 99%, the threshold crossing occurs after memorisation onset at $n = 200$ and $n = 350$, showing that the event ordering for these small arms is sensitive to the selected utility threshold.

Table B.4: Individual-seed and arm-level memorisation onsets for the replicated MNIST arms.

n	Individual-seed onset steps	Arm-level onset
200	3,000; 3,000; 4,000	3,000
350	3,000; 4,000	3,000
1,000	8,000; 8,000; 12,000	12,000

Notes: An individual-seed onset is the first observed checkpoint at which that seed reaches a memorised fraction of at least 5%. The arm-level onset used in the main analysis is calculated from the mean memorised fraction across the seeds available at each checkpoint, as defined in Equation 3.15.

Chapter C Supplementary Measurement Controls

This appendix provides the numerical results of the measurement controls summarised in Section 4.2. It records the behaviour of the memorisation criterion on unseen real images, the comparison between the full and accelerated samplers, and the checks used to interpret the classifier-accuracy measurements. These controls define the limits of the measurement instruments; they do not introduce additional experimental outcomes.

C.1 Null Calibration of the Memorisation Criterion

The null calibration compared unseen MNIST test images with nested subsets of the MNIST training set. A test image was classified as memorised when its nearest-neighbour distance was less than one third of its second-nearest- neighbour distance, using the same criterion as in the main experiments. Table C.1 reports the resulting memorised fraction, the fraction of reference images selected as a nearest neighbour, and the median nearest-neighbour distance.

Table C.1: Null calibration of the MNIST memorisation criterion.

Reference-set size	Memorised fraction	Unique-neighbour fraction	Median distance
200	0.000	0.0200	13.09
1,000	0.000	0.0969	11.58
2,000	0.000	0.1812	10.96
5,000	0.000	0.3653	10.21
10,000	0.000	0.5374	9.73
60,000	0.000	0.8650	8.67

Notes: The memorised fraction is the proportion satisfying the nearest-neighbour ratio criterion at $k = \frac{1}{3}$. The unique-neighbour fraction is the number of distinct reference images selected as a nearest neighbour divided by the number of probe images, following Equation 3.9. Distances are Euclidean distances in the pixel representation used by the experiment.

No unseen test image satisfied the copying criterion at any tested reference- set size. The median nearest-neighbour distance nevertheless fell from 13.09 to 8.67 as the reference set grew. This is the density effect that the ratio criterion was intended to distinguish from isolated proximity to one training image. The calibration supports setting memorisation onset above an observed null fraction of zero, but it does not establish the criterion's specificity outside these MNIST test images.

C.2 Sampler Control

The sampler control used the same $n = 200$ generator checkpoint at step 32,000 and changed only the reverse-process sampler. Table C.2 compares the 1,000-step DDPM sampler with the 100-step DDIM sampler used in the main experiments.

At this checkpoint, accelerated sampling lowered both measured quantities. Because the comparison contains one generator checkpoint at one training-set size, it estimates neither an average sampler effect nor equivalence between the samplers. Its role is limited to identifying the direction and approximate magnitude of sampler-related measurement bias in this control.

Table C.2: Sampler comparison at $n = 200$ and generator step 32,000.

Sampler	Sampling steps	Classifier accuracy	Memorised fraction
DDPM	1,000	0.7927	0.9988
DDIM	100	0.7743	0.9536
DDIM minus DDPM		−0.0184	−0.0452

Notes: Accuracy differences are proportions; −0.0184 corresponds to a reduction of 1.84 percentage points. Both rows use the MLP downstream classifier.

C.3 Validation-Loss and Baseline Checks

Held-out generator validation loss declined throughout the completed training trajectories and showed no visible transition at either the classifier- accuracy plateau or memorisation onset. Validation loss therefore did not provide an alternative operational boundary for either event. The plateau and onset definitions used in the main analysis remain based on downstream classifier accuracy and the memorised fraction respectively.

The size-matched real-data classifiers were trained for between 395 and 402 optimiser steps, compared with exactly 400 steps for each synthetic-data classifier. At $n = 200$, the MLP real-data baseline ranged from 0.816 to 0.824 across classifier seeds. The small within-arm range supports its use as a stable optimisation-budget reference. As explained in Section 3.8, matching optimiser steps does not match the number or diversity of training examples, because every synthetic classifier receives 5,000 generated images whereas each real-data baseline receives n real images.

Taken together, these controls support the operational use of the MNIST measurements while preserving three qualifications: the copying criterion was calibrated only on MNIST, the accelerated-sampler comparison was limited to a single checkpoint, and the real-data comparison matches optimiser budget but not sample count.

Chapter D CIFAR-10 Extension

This appendix provides the configuration and complete summary of the CIFAR-10 extension reported in Section 4.6. The extension was designed as a limited feasibility check of the MNIST pipeline on natural colour images. It was not sufficiently validated to test whether the MNIST findings generalise.

D.1 Experimental Scope

CIFAR-10 contains 32×32 colour images, rather than the 28×28 greyscale images used in the main MNIST experiments. Image dimensions, channel counts and output namespaces were changed accordingly. The generator-training procedure, 100-step DDIM sampler, generated dataset size of 5,000 images, downstream evaluation procedure and pixel-space nearest-neighbour ratio criterion were retained. The extension covered $n = 200$ and $n = 1{,}000$, with one generator seed at each size, and continued to step 64,000. Both the MLP and CNN downstream classifiers were evaluated.

Table D.1: Configuration of the CIFAR-10 feasibility extension.

Component	Configuration
Dataset	CIFAR-10, 32×32, three colour channels
Generator training-set sizes	$n = 200$ and $n = 1{,}000$
Generator replication	One seed per training-set size
Final generator step	64,000
Sampler	DDIM, 100 sampling steps
Generated images per checkpoint	5,000
Downstream classifiers	MLP and CNN
Copying criterion	Pixel-space nearest-neighbour ratio, $k = \frac{1}{3}$
Memorisation onset	First checkpoint at or above 0.05 memorised fraction
Compute	NVIDIA Tesla T4

The intended measurement validation was to compare known CIFAR-10 copies with novel CIFAR-10 images before interpreting the generator measurements. A complete positive-control result was not retained. Consequently, a low measured memorised fraction may indicate either little pixel-space copying or failure of the MNIST-derived criterion to identify semantically similar natural images.

D.2 Results

Table D.2 reports the best downstream accuracy observed for each classifier, the checkpoint at which it occurred, the classifier-specific real-data baseline reported for that training-set size, and the largest memorised fraction observed over generator training.

At $n = 200$, the best CNN result was 0.0028 below its real-data baseline, while the best MLP result was 0.0679 below its baseline. At $n = 1{,}000$, the corresponding shortfalls were 0.1054 and 0.1186. The measured memorised fraction never crossed 0.05, reaching at most 0.031 at $n = 200$ and remaining zero at $n = 1{,}000$. No memorisation onset can therefore be placed relative to an accuracy plateau in either arm.

Table D.2: Complete summary of the CIFAR-10 extension.

n	Best MLP	MLP step	MLP baseline	Best CNN	CNN step	CNN baseline
200	0.2072	64,000	0.2751	0.2424	32,000	0.2452
1,000	0.2317	16,000	0.3503	0.2739	8,000	0.3793

n	Maximum memorised fraction	Step of maximum	Onset status
200	0.031	64,000	Not observed
1,000	0.000	—	Not observed

Notes: Accuracy is evaluated on real CIFAR-10 test images. "Best" denotes the maximum observed checkpoint accuracy rather than a separately estimated plateau. Each real-data baseline uses the same classifier architecture and generator training-set size as the corresponding synthetic result. Memorisation onset requires a memorised fraction of at least 0.05. No unique step is shown when the memorised fraction remained zero at every checkpoint.

D.3 Interpretive Limit

The extension does not resolve whether the low memorised fractions reflect generator behaviour or measurement failure. Its classifier results also show that the generated images retained substantially less downstream utility than the real images at $n = 1{,}000$. These two limitations prevent a direct test of the temporal relationship established on MNIST.

The CIFAR-10 runs therefore demonstrate only that the experimental pipeline can be executed on a natural-image dataset. They do not validate the copying criterion for CIFAR-10 and do not provide evidence that the MNIST event ordering or scaling relationship transfers to natural images. A valid extension would first require retained positive and negative controls for the copying measure, followed by sufficient generator replication to quantify between-seed variation.

Chapter E Project Timeline

Supervision meetings were held weekly throughout the project.

Week 1: 8 to 14 June

- Proposed an initial project on valuing datasets from their metadata, 8 June.
- Direction changed following feedback: too focused on the financial side, not sufficiently computational.
- Explored dataset condensation and evolutionary approaches as a possible bridge to the original idea.

Week 2: 15 to 21 June

- Agreed a new direction, 15 June: diffusion models on MNIST, trading privacy against utility.
- Completed the Hugging Face diffusion models course and submitted the research question for the proposal form.
- Wrote a longer proposal with a fuller literature review and seven core objectives.

Week 3: 22 to 28 June

- Adapted the tutorial diffusion model to feed an MLP. First results: 0.8947 synthetic against a 0.9706 real baseline.
- Used Claude to guide the best structuring for moving from notebooks and onto scripts. I wrote `CCUNETGenerator.py`, a separated MLP and data loader, `sweep.py`, and Weights and Biases logging.
- Hit Colab compute limits and moved the project onto a local RTX 3070. Narrowed the scope and fixed the three research questions.

Week 4: 29 June to 5 July

- First real experimental runs. Added a CNN as a second classifier and extended the sweep to five seeds for error bars.
- Built an initial baseline matched to the 5,000 generated images. This was later replaced by the source-size, step-matched baseline used in the final analysis.
- Reorganised the repository and started the literature review.

Week 5: 6 to 12 July

- Reading and writing week, with limited working capacity owing to personal commitments.
- Reported the plateau at around 80 epochs with no decline, and the seven-hour cost of that run.
- Wrote `memorisation.py`, a main measuring instrument for the study, used Claude to review to help write tests for implementation of Yoon criterion.

Week 6: 13 to 19 July

- Found and fixed defect, CNN results had been plotted against the MLP baseline.
- Refactored the generator and subset handling, setting up the controlled-subset design.
- Ran a single-seed scout pass across all training-set sizes. The memorisation transition was visible and ordered by size, with no large decline in utility across it.

Week 7: 20 to 26 July

- Compared the DDIM-100 sampler with the full reverse process at one checkpoint. The broad pattern was preserved, although DDIM produced slightly lower accuracy and memorised fraction.

- Measured the memorisation criterion's observed false-positive floor rather than assuming it was zero.
- GitLab rejected a 1 GB push, so checkpoints were excluded from GitHub. The subset function was also found to be truncating classes.

Week 8: 27 July to 2 August

- Produced the first analysis figures for the memorisation versus utility question, rebuilt from freshly pulled data, scripts for figures were created using Claude to enable fast iteration, and ensure data was presented in a way that was easy to intearpret.
- Cut the literature review following feedback that it was too long and too verbose for a scientific register.
- Set up vlab and committed the CIFAR-10 extension. Fashion-MNIST cut as pre-agreed.

Week 9: 3 to 9 August

- Began preparation for the viva presentation, five minutes presenting and up to twenty of questions.
- Continued to iterate on plotting scripts and spent this week writing my slides for my viva presentation.
- Finalised the plan for the methodology chapter.

Week 10: 10 to 16 August

- Drafted the methodology chapter, reorganising it to lead with each choice and its justification rather than in pipeline order.
- Settled some terminology consistency issues, adopting CAS as the primary term with TSTR as the general protocol name.
- Uploaded/ updated presentation slides on 14 August in line with Per's feedback. Literature review feedback received on 16 August.

Week 11: 17 to 23 August

- Discussion in meeting on how to best present writing, and spent much of week implementing stylistic changes.
- Viva completed.
- Literature review slight revision/ compression, results chapter, and the opening of the discussion. Claude used to help restructure sentences, and keep writing concise.

Week 12: 24 to 30 August

- Refined discussion based on supervisor feedback and drafted the conclusion (make writing more concise and consistent), Claude was helpful here for making writing more concise.
- Restructured Chapter 5 so that each RQ is broken down.
- Added appendices where they were missing or not working.
- Claude was used to proofread for grammar, cross references and numerical consistency across arguments.