

000 001 002 003 004 005 006 007 MIAU: MEMBERSHIP INFERENCE ATTACK UNLEAR- 008 NING SCORE FOR QUANTIFYING THE FORGETTING 009 QUALITY OF UNLEARNING METHODS 010 011 012

013 **Anonymous authors**
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
0100
0101
0102
0103
0104
0105
0106
0107
0108
0109
0110
0111
0112
0113
0114
0115
0116
0117
0118
0119
0120
0121
0122
0123
0124
0125
0126
0127
0128
0129
0130
0131
0132
0133
0134
0135
0136
0137
0138
0139
0140
0141
0142
0143
0144
0145
0146
0147
0148
0149
0150
0151
0152
0153
0154
0155
0156
0157
0158
0159
0160
0161
0162
0163
0164
0165
0166
0167
0168
0169
0170
0171
0172
0173
0174
0175
0176
0177
0178
0179
0180
0181
0182
0183
0184
0185
0186
0187
0188
0189
0190
0191
0192
0193
0194
0195
0196
0197
0198
0199
0200
0201
0202
0203
0204
0205
0206
0207
0208
0209
0210
0211
0212
0213
0214
0215
0216
0217
0218
0219
0220
0221
0222
0223
0224
0225
0226
0227
0228
0229
0230
0231
0232
0233
0234
0235
0236
0237
0238
0239
0240
0241
0242
0243
0244
0245
0246
0247
0248
0249
0250
0251
0252
0253
0254
0255
0256
0257
0258
0259
0260
0261
0262
0263
0264
0265
0266
0267
0268
0269
0270
0271
0272
0273
0274
0275
0276
0277
0278
0279
0280
0281
0282
0283
0284
0285
0286
0287
0288
0289
0290
0291
0292
0293
0294
0295
0296
0297
0298
0299
0300
0301
0302
0303
0304
0305
0306
0307
0308
0309
0310
0311
0312
0313
0314
0315
0316
0317
0318
0319
0320
0321
0322
0323
0324
0325
0326
0327
0328
0329
0330
0331
0332
0333
0334
0335
0336
0337
0338
0339
0340
0341
0342
0343
0344
0345
0346
0347
0348
0349
0350
0351
0352
0353
0354
0355
0356
0357
0358
0359
0360
0361
0362
0363
0364
0365
0366
0367
0368
0369
0370
0371
0372
0373
0374
0375
0376
0377
0378
0379
0380
0381
0382
0383
0384
0385
0386
0387
0388
0389
0390
0391
0392
0393
0394
0395
0396
0397
0398
0399
0400
0401
0402
0403
0404
0405
0406
0407
0408
0409
0410
0411
0412
0413
0414
0415
0416
0417
0418
0419
0420
0421
0422
0423
0424
0425
0426
0427
0428
0429
0430
0431
0432
0433
0434
0435
0436
0437
0438
0439
0440
0441
0442
0443
0444
0445
0446
0447
0448
0449
0450
0451
0452
0453
0454
0455
0456
0457
0458
0459
0460
0461
0462
0463
0464
0465
0466
0467
0468
0469
0470
0471
0472
0473
0474
0475
0476
0477
0478
0479
0480
0481
0482
0483
0484
0485
0486
0487
0488
0489
0490
0491
0492
0493
0494
0495
0496
0497
0498
0499
0500
0501
0502
0503
0504
0505
0506
0507
0508
0509
0510
0511
0512
0513
0514
0515
0516
0517
0518
0519
0520
0521
0522
0523
0524
0525
0526
0527
0528
0529
05210
05211
05212
05213
05214
05215
05216
05217
05218
05219
05220
05221
05222
05223
05224
05225
05226
05227
05228
05229
05230
05231
05232
05233
05234
05235
05236
05237
05238
05239
05240
05241
05242
05243
05244
05245
05246
05247
05248
05249
05250
05251
05252
05253
05254
05255
05256
05257
05258
05259
05260
05261
05262
05263
05264
05265
05266
05267
05268
05269
05270
05271
05272
05273
05274
05275
05276
05277
05278
05279
05280
05281
05282
05283
05284
05285
05286
05287
05288
05289
05290
05291
05292
05293
05294
05295
05296
05297
05298
05299
052100
052101
052102
052103
052104
052105
052106
052107
052108
052109
052110
052111
052112
052113
052114
052115
052116
052117
052118
052119
052120
052121
052122
052123
052124
052125
052126
052127
052128
052129
052130
052131
052132
052133
052134
052135
052136
052137
052138
052139
052140
052141
052142
052143
052144
052145
052146
052147
052148
052149
052150
052151
052152
052153
052154
052155
052156
052157
052158
052159
052160
052161
052162
052163
052164
052165
052166
052167
052168
052169
052170
052171
052172
052173
052174
052175
052176
052177
052178
052179
052180
052181
052182
052183
052184
052185
052186
052187
052188
052189
052190
052191
052192
052193
052194
052195
052196
052197
052198
052199
052200
052201
052202
052203
052204
052205
052206
052207
052208
052209
052210
052211
052212
052213
052214
052215
052216
052217
052218
052219
052220
052221
052222
052223
052224
052225
052226
052227
052228
052229
052230
052231
052232
052233
052234
052235
052236
052237
052238
052239
052240
052241
052242
052243
052244
052245
052246
052247
052248
052249
052250
052251
052252
052253
052254
052255
052256
052257
052258
052259
052260
052261
052262
052263
052264
052265
052266
052267
052268
052269
052270
052271
052272
052273
052274
052275
052276
052277
052278
052279
052280
052281
052282
052283
052284
052285
052286
052287
052288
052289
052290
052291
052292
052293
052294
052295
052296
052297
052298
052299
052300
052301
052302
052303
052304
052305
052306
052307
052308
052309
052310
052311
052312
052313
052314
052315
052316
052317
052318
052319
052320
052321
052322
052323
052324
052325
052326
052327
052328
052329
052330
052331
052332
052333
052334
052335
052336
052337
052338
052339
052340
052341
052342
052343
052344
052345
052346
052347
052348
052349
052350
052351
052352
052353
052354
052355
052356
052357
052358
052359
052360
052361
052362
052363
052364
052365
052366
052367
052368
052369
052370
052371
052372
052373
052374
052375
052376
052377
052378
052379
052380
052381
052382
052383
052384
052385
052386
052387
052388
052389
052390
052391
052392
052393
052394
052395
052396
052397
052398
052399
052400
052401
052402
052403
052404
052405
052406
052407
052408
052409
052410
052411
052412
052413
052414
052415
052416
052417
052418
052419
052420
052421
052422
052423
052424
052425
052426
052427
052428
052429
052430
052431
052432
052433
052434
052435
052436
052437
052438
052439
052440
052441
052442
052443
052444
052445
052446
052447
052448
052449
052450
052451
052452
052453
052454
052455
052456
052457
052458
052459
052460
052461
052462
052463
052464
052465
052466
052467
052468
052469
052470
052471
052472
052473
052474
052475
052476
052477
052478
052479
052480
052481
052482
052483
052484
052485
052486
052487
052488
052489
052490
052491
052492
052493
052494
052495
052496
052497
052498
052499
052500
052501
052502
052503
052504
052505
052506
052507
052508
052509
052510
052511
052512
052513
052514
052515
052516
052517
052518
052519
052520
052521
052522
052523
052524
052525
052526
052527
052528
052529
052530
052531
052532
052533
052534
052535
052536
052537
052538
052539
052540
052541
052542
052543
052544
052545
052546
052547
052548
052549
052550
052551
052552
052553
052554
052555
052556
052557
052558
052559
052560
052561
052562
052563
052564
052565
052566
052567
052568
052569
052570
052571
052572
052573
052574
052575
052576
052577
052578
052579
052580
052581
052582
052583
052584
052585
052586
052587
052588
052589
052590
052591
052592
052593
052594
052595
052596
052597
052598
052599
052600
052601
052602
052603
052604
052605
052606
052607
052608
052609
052610
052611
052612
052613
052614
052615
052616
052617
052618
052619
052620
052621
052622
052623
052624
052625
052626
052627
052628
052629
052630
052631
052632
052633
052634
052635
052636
052637
052638
052639
052640
052641
052642
052643
052644
052645
052646
052647
052648
052649
052650
052651
052652
052653
052654
052655
052656
052657
052658
052659
052660
052661
052662
052663
052664
052665
052666
052667
052668
052669
052670
052671
052672
052673
052674
052675
052676
052677
052678
052679
052680
052681
052682
052683
052684
052685
052686
052687
052688
052689
052690
052691
052692
052693
052694
052695
052696
052697
052698
052699
052700
052701
052702
052703
052704
052705
052706
052707
052708
052709
052710
052711
052712
052713
052714
052715
052716
052717
052718
052719
052720
052721
052722
052723
052724
052725
052726
052727
052728
052729
052730
052731
052732
052733
052734
052735
052736
052737
052738
052739
052740
052741
052742
052743
052744
052745
052746
052747
052748
052749
052750
052751
052752
052753
052754
052755
052756
052757
052758
052759
052760
052761
052762
052763
052764
052765
052766
052767
052768
052769
052770
052771
052772
052773
052774
052775
052776
052777
052778
052779
052780
052781
052782
052783
052784
052785
052786
052787
052788
052789
052790
052791
052792
052793
052794
052795
052796
052797
052798
052799
052800
052801
052802
052803
052804
052805
052806
052807
052808
052809
052810
052811
052812
052813
052814
052815
052816
052817
052818
052819
052820
052821
052822
052823
052824
052825
052826
052827
052828
052829
052830
052831
052832
052833
052834
052835
052836
052837
052838
052839
052840
052841
052842
052843
052844
052845
052846
052847
052848
052849
052850
052851
052852
052853
052854
052855
052856
052857
052858
052859
052860
052861
052862
052863
052864
052865
052866
052867
052868
052869
052870
052871
052872
052873
052874
052875
052876
052877
052878
052879
052880
052881
052882
052883
052884
052885
052886
052887
052888
052889
052890
052891
052892
052893
052894
052895
052896
052897
052898
052899
052900
052901
052902
052903
052904
052905
052906
052907
052908
052909
052910
052911
052912
052913
052914
052915
052916
052917
052918
052919
052920
052921
052922
052923
052924
052925
052926
052927
052928
052929
052930
052931
052932
052933
052934
052935
052936
052937
052938
052939
052940
052941
052942
052943
052944
052945
052946
052947
052948
052949
052950
052951
052952
052953
052954
052955
052956
052957
052958
052959
052960
052961
052962
052963
052964
052965
052966
052967
052968
052969
052970
052971
052972
052973
052974
052975
052976
052977
052978
052979
052980
052981
052982
052983
052984
052985
052986
052987
052988
052989
052990
052991
052992
052993
052994
052995
052996
052997
052998
052999
053000
053001
053002
053003
053004
053005
053006
053007
053008
053009
053010
053011
053012
053013
053014
053015
053016
053017
053018
053019
053020
053021
053022
053023
053024
053025
053026
053027
053028
053029
053030
053031
053032
053033
053034
053035
053036
053037
053038
053039
053040
053041

054 1.1 PROBLEM STATEMENT
055

056 Despite the widespread adoption of membership inference attacks (MIAs) for evaluating unlearning,
057 current evaluation methods remain limited. Most prior works compute MIA performance on only
058 a single subset comparison—*Forget vs Test*, *Retain vs Forget*, or *Retain vs Test* (Chen et al., 2021a;
059 Kurmanji et al., 2023; Chundawat et al., 2023a; Graves et al., 2021). Each captures a distinct aspect
060 of model behavior, but relying on one gives an incomplete and potentially misleading picture. The
061 *Forget vs Test* comparison measures whether the forget set remains more distinguishable than unseen
062 data, capturing residual overfitting. Low separability here does not imply successful forgetting if the
063 model has lost generalization and predictions become less confident. The *Retain vs Forget* comparison
064 evaluates whether the model still treats forgotten samples like retained training data (Chundawat
065 et al., 2023b). Because both sets come from the training distribution, effective unlearning should
066 make model behavior on the forget set diverge from that on the retain set. High separability therefore
067 signals successful forgetting, as the model no longer treats the forget set as part of training. How-
068 ever, separability alone cannot distinguish targeted forgetting from broader shifts in model behavior,
069 such as unintended changes in handling retained data. The *Retain vs Test* setup measures whether
070 the model behaves consistently on retained training data and unseen test data, serving as a sanity
071 check for generalization. Although it does not directly show that the forget set was removed, it is
072 essential for ruling out trivial explanations like underfitting or global degradation that could mimic
073 forgetting. Thus, *Retain vs Test* provides context for interpreting the other two comparisons.

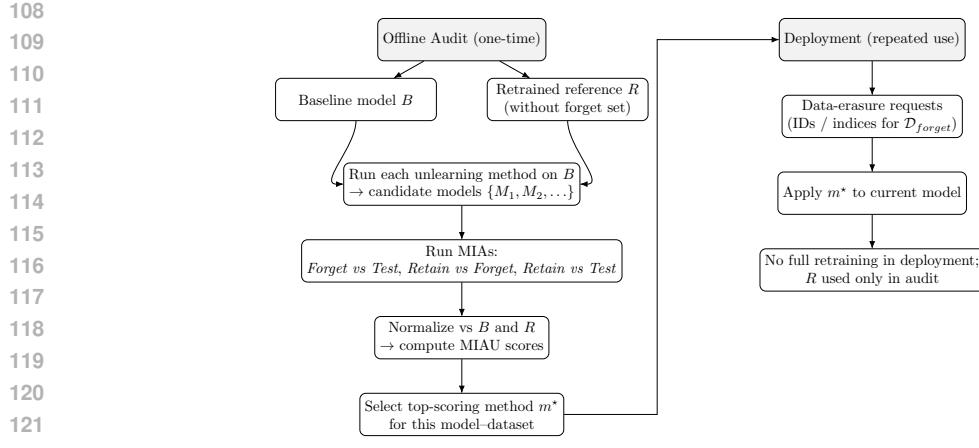
074 Each configuration examines a necessary but not sufficient condition for verifying unlearning. Evalu-
075 ating only one cannot reveal whether a change in MIA performance comes from targeted forgetting
076 or unrelated model degradation. Only by jointly analyzing all three can one isolate forgetting-
077 specific effects from confounders such as underfitting, representation collapse, or loss of utility. An
078 effective evaluation must integrate these perspectives into a unified measure capturing both com-
079 pleteness and correctness of forgetting.

080 Existing MIA-based evaluations also often lack proper baselines or reference points, making privacy
081 gains hard to interpret. Many studies omit membership inference results on baseline or retrained
082 models (Graves et al., 2021; Jia et al., 2023; Li et al., 2024). The baseline model, trained on the full
083 dataset, represents worst-case privacy leakage. The retrained model, trained from scratch without the
084 forget set, represents the best-case “complete forgetting.” An unlearned model’s MIA score without
085 these two reference points cannot show how much forgetting has been achieved or how close the
086 method is to the ideal.

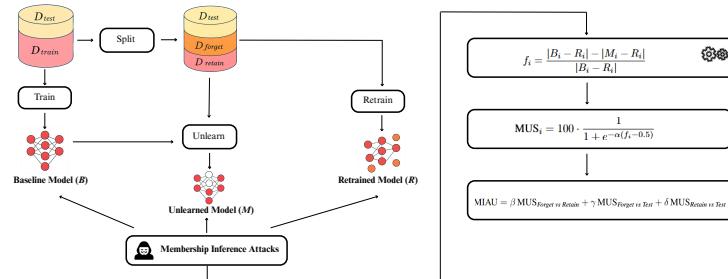
087 1.2 PROPOSED SOLUTION
088

089 While most prior evaluations omit such baselines, a limited number of studies retrain a model with-
090 out the forget set to obtain a gold-standard benchmark, to use this reference solely for empirical,
091 non-quantitative comparison with their proposed methods (Foster et al., 2024). While this single
092 retrained reference is valuable for research, repeating such full retraining during deployment contra-
093 dictants the purpose of machine unlearning and imposes additional computational cost. Therefore, in
094 deployment settings, it is essential to establish a reliable metric that enables us to more accurately
095 evaluate the practical utility of existing unlearning methods for our use case. Prior studies indicate
096 that the effectiveness of unlearning techniques can vary significantly across different tasks (Cheng
097 & Amiri, 2024). Consequently, identifying which approach is most suitable for our deployment
098 scenario is critical, particularly given that continuous model retraining is infeasible in practice. To
099 address this challenge, we introduce the Membership Inference Attack Unlearning Score (MIAU)
100 as *an offline auditing framework* rather than an operational component of the unlearning pipeline.
101 The resulting score computed once for each model–dataset pair guides researchers and practitioners
102 in selecting the most effective unlearning method for their specific context. Thus, *it facilitates the*
103 *consistent application of the most suitable unlearning method for a given model–dataset context,*
104 *while eliminating the need for additional retraining overhead during model deployment.*

105 In this context, we propose the Membership Inference Attack Unlearning Score (MIAU)—a metric
106 that quantifies how much of the performance difference between the original baseline model and
107 the fully retrained model is closed by an unlearning method. MIAU captures the degree to which
108 the unlearned model approximates the ideal privacy behavior of a retrained model that has never
109 seen the forget set. It combines three complementary MIA comparisons—*Forget vs Test*, *Retain vs*



Forget, and *Retain vs Test*—each measuring a distinct property of forgetting: residual memorization, removal effectiveness, and generalization stability, respectively. MIAU normalizes the unlearning method’s performance between the baseline and retrain endpoints, producing a single interpretable score that reflects the completeness of forgetting. We evaluate MIAU on four standard image classification benchmarks—*MNIST* (LeCun et al., 1998), *CIFAR-10*, *CIFAR-20* (Krizhevsky & Hinton, 2009), and *MUCAC* (Choi & Na, 2023)—using three model architectures: *ResNet-18* (He et al., 2016), *All-CNN* (Springenberg et al., 2015), and *Vision Transformer* (Dosovitskiy et al., 2021). The evaluation includes four representative unlearning methods: *Fine-tune* (Bourtoule et al., 2021), *SSD* (Foster et al., 2024), *Amnesiac* (Graves et al., 2021), and *Teacher* (Chundawat et al., 2023b).



Results demonstrate that MIAU provides a reliable and consistent measure of forgetting quality, distinguishing methods that closely approximate retraining from those that do not. Unlike raw MIA accuracy metrics, which are sensitive to attack strength, calibration shifts, or global degradation, MIAU provides an interpretable score between the *baseline* and *retrain* endpoints, enabling consistent comparisons across methods and datasets. To further assess its robustness, we evaluate MIAU under varying unlearning levels—removing 25%, 50%, and 75% of the forget set and assess improvements in score as more of the unlearning data is preserved. Statistical significance tests are further established via paired p-value tests comparing MIAU across methods.

2 RELATED WORK

Several machine unlearning studies assess forgetting by comparing model accuracy on the forget and retain subsets of the training data (Golatkar et al., 2020; 2021; Bourtoule et al., 2021). After unlearning, accuracy on the forget set is expected to drop slightly, ideally approaching test-level per-

162 formance, while retain-set accuracy should remain close to the original model or a retrained counter-
 163 part, indicating preserved utility on non-forgotten data. Additionally, test accuracy is typically used
 164 to ensure that overall generalization is not adversely affected by the unlearning process. However,
 165 relying on forget and retain accuracy alone introduces several limitations. A model can exhibit low
 166 forget-set accuracy by superficially suppressing predictions on the forget set, without eliminating
 167 the underlying learned representations (Golatkar et al., 2021; Nguyen et al., 2022). At the same
 168 time, a high retain accuracy does not guarantee that forgetting was targeted, as the model may have
 169 degraded uniformly or adapted in a way that preserves training performance without isolating the
 170 forgotten information. In contrast, MIAU evaluates all three membership-inference comparisons
 171 together, separating targeted forgetting from uniform degradation and clarifying whether accuracy
 172 changes reflect genuine unlearning or indiscriminate loss of predictive capacity.

173 To complement or replace these accuracy-based metrics, researchers have proposed dedicated un-
 174 learning evaluation scores. Zero-Retrain Forgetting (ZRF) (Chundawat et al., 2023b) score quanti-
 175 fies whether the model’s predictions on the forget set become indistinguishable from those of a
 176 weak “incompetent” teacher. However, ZRF assumes that prediction randomness alone equates to
 177 forgetting, potentially overlooking residual information retained in the feature space. Meanwhile,
 178 the Normalized Machine Unlearning Score (NoMUS) (Choi & Na, 2023) balances forgetting and
 179 utility into a single normalized score. Despite its convenience, NoMUS relies on the particular for-
 180 mulation of the forgetting score and the fixed weight assigned to privacy versus utility, which may
 181 obscure whether high scores result from balanced forgetting and retention or from prioritizing one
 182 at the expense of the other. In contrast, MIAU combines the three MIA outcomes with explicitly
 183 adjustable coefficients, offering a transparent and theoretically grounded means to balance privacy
 184 protection for $\mathcal{D}_{\text{forget}}$ with predictive utility on $\mathcal{D}_{\text{retain}}$ and $\mathcal{D}_{\text{test}}$.

185 3 PRELIMINARIES

186 Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a supervised dataset with features $x_i \in \mathbb{R}^n$ and labels $y_i \in \{1, \dots, K\}$.
 187 Machine unlearning aims to remove the influence of a subset $\mathcal{D}_{\text{forget}} \subset \mathcal{D}$ while retaining perfor-
 188 mance on the remaining data $\mathcal{D}_{\text{retain}} = \mathcal{D} \setminus \mathcal{D}_{\text{forget}}$. The model is modified so that its predictions and
 189 internal representations are indistinguishable from those of a model trained only on $\mathcal{D}_{\text{retain}}$.

190 We additionally define a disjoint test set $\mathcal{D}_{\text{test}}$ for generalization and privacy evaluation. The *base-
 191 line model* $\phi_{\theta_{\text{base}}}$ is trained on \mathcal{D} , and the *retrain model* $\phi_{\theta_{\text{retrain}}}$ is trained from scratch on $\mathcal{D}_{\text{retain}}$,
 192 representing ideal forgetting.

193 Unlearning quality is measured with three Membership Inference Attack (MIA) tasks, each imple-
 194 mented by training a binary classifier on model outputs to distinguish samples from two subsets. The *Forget vs Test* setup checks whether forgotten samples remain identifiable; perfect forgetting yields
 195 random guessing. The *Retain vs Forget* setup assesses removal effectiveness by testing separability
 196 between retained and forgotten data. The *Retain vs Test* setup evaluates whether retained data and
 197 unseen test data elicit similar predictions, reflecting generalization.

201 4 PROPOSED METHOD

202 The Membership Inference Attack Unlearning Score (MIAU) quantifies the extent to which an un-
 203 learning method approximates the privacy behavior of an ideal retrained model. It operates by
 204 comparing the outputs of the unlearned model to both the baseline and retrained models across mul-
 205 tiple membership inference attack (MIA) tasks. Each task captures a distinct aspect of forgetting:
 206 residual memorization, separation between forgotten and retained data, and consistency on unseen
 207 samples. MIAU aggregates these evaluations into a single bounded score, enabling standardized
 208 assessment of unlearning effectiveness across different settings.

211 4.1 GAP CLOSURE FRACTION

212 Let the objective be to quantify the effectiveness of an unlearning method in approximating the
 213 privacy behavior of a model trained from scratch without the forgotten data. We consider three
 214 models evaluated on a given membership inference attack (MIA) task i : the *Baseline* model, trained
 215 on the full dataset; the *Retrain* model, trained with the forget set excluded from the beginning;

216 and the *Unlearning* model, which applies a forgetting algorithm to the baseline. Denote by B_i ,
 217 R_i , and M_i the membership-inference accuracies measured on task i for the Baseline, Retrain, and
 218 Unlearning models, respectively.

219 To measure how much of the privacy gap between the Baseline and Retrain models has been closed
 220 by the unlearning method, we define the gap closure fraction as Eq. equation 1:

$$222 \quad 223 \quad 224 \quad f_i = \frac{|B_i - R_i| - |M_i - R_i|}{|B_i - R_i|} \quad (1)$$

225 The quantity f_i represents the relative reduction in distance to the retrain reference point. When
 226 $|B_i - R_i| = 0$, we directly set $f_i = 0$ to avoid division by zero. A value of $f_i = 1$ implies
 227 perfect alignment with the retrain model ($M_i = R_i$), indicating ideal forgetting. When $f_i = 0$, the
 228 unlearning method does not reduce the privacy gap relative to the baseline, and $f_i < 0$ indicates that
 229 the unlearning method increases the divergence from the retrain behavior.

231 4.2 MATCHING UNLEARNING SCORE ON METRIC i

232 While f_i provides a normalized measure of forgetting effectiveness, it is unbounded and may be
 233 sensitive to outliers. To obtain a bounded and smooth score in the range (0, 100), we apply a logistic
 234 transformation to f_i and define the Matching Unlearning Score (MUS) as Eq. equation 2:

$$236 \quad 237 \quad 238 \quad \text{MUS}_i = 100 \cdot \frac{1}{1 + e^{-\alpha(f_i - 0.5)}} \quad (2)$$

239 This formulation ensures several desirable properties. First, the score remains bounded in the open
 240 interval (0, 100) without requiring manual clipping. Second, it centers the neutral reference point at
 241 $f_i = 0.5$, assigning a score of $\text{MUS}_i = 50$ to methods that close half the gap to retraining. Third,
 242 the parameter α controls the sensitivity of the transformation: larger values of α produce a steeper
 243 transition around the midpoint, amplifying differences in intermediate performance. To ensure that
 244 the MUS_i score on *baseline* is near 0, and MUS_i score on *retrain* is near 100, the α value for our
 245 setup was calculated to be 13.8. The detailed derivation of the $\alpha = 13.8$ calibration, which ensures
 246 that the MUS_i score approaches 0 for the *baseline* and 100 for the *retrain* setting, is provided in
 247 Appendix A.1.

248 The logistic transformation is chosen because it maps the unbounded f_i into a stable, interpretable
 249 percentage scale (0–100) while preserving the relative ordering of methods, enabling direct com-
 250 parison across metrics. From a mathematical standpoint, the transformation satisfies the following
 251 limits: $\text{MUS}_i \rightarrow 0$ as $f_i \rightarrow -\infty$, $\text{MUS}_i = 50$ when $f_i = 0.5$, and $\text{MUS}_i \rightarrow 100$ as $f_i \rightarrow 1$. The
 252 logistic function is continuous and differentiable, making it suitable for ranking, visualization, or in-
 253 tegration into gradient-based optimization procedures such as hyperparameter tuning or automated
 254 model selection.

255 4.3 MEMBERSHIP INFERENCE ATTACK UNLEARNING SCORE (MIAU)

256 To produce a unified assessment of forgetting quality across multiple MIA tasks, we define the
 257 final unlearning score as the average of the individual MUS values computed for each MIA setup.
 258 Specifically, the Membership Inference Attack Unlearning Score (MIAU) is given as Eq. equation 3:

$$259 \quad 260 \quad 261 \quad \text{MIAU} = \beta \text{MUS}_{\text{Forget vs Retain}} + \gamma \text{MUS}_{\text{Forget vs Test}} + \delta \text{MUS}_{\text{Retain vs Test}} \quad (3)$$

262 The three MIA tasks correspond to the following subset pairs: $\mathcal{D}_{\text{forget}}$ vs. $\mathcal{D}_{\text{retain}}$ (*Forget vs Retain*),
 263 $\mathcal{D}_{\text{forget}}$ vs. $\mathcal{D}_{\text{test}}$ (*Forget vs Test*), and $\mathcal{D}_{\text{retain}}$ vs. $\mathcal{D}_{\text{test}}$ (*Retain vs Test*). Each task captures a distinct
 264 dimension of inference risk: the first reflects removal effectiveness, the second measures residual
 265 memorization, and the third assesses generalization consistency. The coefficients β , γ , and δ are
 266 non-negative weights that represent the relative importance assigned to each MIA direction and
 267 satisfy the constraints $\beta + \gamma + \delta = 1$ and $0 \leq \beta, \gamma, \delta \leq 1$. Depending on the desired emphasis, these
 268 weights can be chosen differently; however, for our experiments we set them equal ($\beta = \gamma = \delta = \frac{1}{3}$)
 269 to provide a balanced evaluation of unlearning performance, ensuring that no single MIA direction
 dominates the assessment.

270 5 EXPERIMENTAL SETUP
271272 **Datasets and base models.** We evaluate the performance of our proposed MIAU score on four
273 different datasets—*MNIST* (LeCun et al., 1998), *CIFAR-10*, its coarse-label variant *CIFAR-20*
274 (Krizhevsky & Hinton, 2009), and the unlearning-specific face-attribute dataset *MUCAC* (Choi &
275 Na, 2023). *MNIST* provides 28×28 grayscale digits (10 classes), *CIFAR-10/20* contain 32×32
276 natural images with 10 and 20 superclasses, and *MUCAC* offers 128×128 celebrity portraits for
277 binary smiling-attribute prediction. Models include *ResNet-18* (He et al., 2016), *All-CNN* (Springen-
278 genberg et al., 2015), and the *Vision Transformer (ViT)* (Dosovitskiy et al., 2021). Dataset details,
279 data preprocessing, training conditions, and all hyperparameter settings appear in Appendix A.5.
280281 **Data splits.** Each dataset is partitioned into three disjoint subsets: the forget set $\mathcal{D}_{\text{forget}}$, the retain
282 set $\mathcal{D}_{\text{retain}} = \mathcal{D}_{\text{train}} \setminus \mathcal{D}_{\text{forget}}$, and the test set $\mathcal{D}_{\text{test}}$, which is held out for evaluation. The forget set
283 consists of 10% of the training data, sampled uniformly from each class to preserve the original
284 class distribution. To ensure statistical robustness, the splitting and evaluation are repeated across
285 10 random seeds. For additional non-random evaluations, we also perform full-class forgetting of
286 the `electric_devices` class in the *CIFAR-20* All-CNN setup and 10% subclass forgetting of
287 the `veg` category.
288289 **Unlearning methods.** We evaluate four representative unlearning methods: Fine-tune, SSD, Am-
290 nesiac, and Teacher. Fine-tune (Bourtoule et al., 2021) retrains the model on the retain set with
291 partial weight updates, SSD (Foster et al., 2024) penalizes parameters most influenced by the forget
292 set, Amnesiac (Graves et al., 2021) reverses their gradient contributions, and Teacher (Chundawat
293 et al., 2023b) distills knowledge from a full-data teacher into a student trained only on the retain set.
294 All unlearning hyperparameters are detailed in Appendix A.5.4.295 **Attack training protocol.** To quantify residual memorization and forgetting, we employ mem-
296 bership inference attacks using the model’s softmax output distributions as input features. For each
297 pair of data subsets involved in a given MIA task, a binary logistic regression classifier is trained to
298 discriminate between them, similar to the setup offered by (Chundawat et al., 2023a). The training
299 set for the attack model consists of 80% of the available logits, while 20% is held out for evaluation.
300 Both entropy and maximum class confidence are implicitly captured in the softmax vectors, serving
301 as indicators of memorization and decision margin. To avoid sampling bias during training, all sub-
302 set pairs used in MIA tasks are size-matched by uniformly subsampling the larger set to match the
303 cardinality of the smaller one. We further report the membership-inference accuracies obtained on
304 the attack model’s test split. Attack classifiers are trained independently for each of the three MIA
305 setups: *Forget vs Test* ($\mathcal{D}_{\text{forget}}$ vs. $\mathcal{D}_{\text{test}}$), *Retain vs Forget* ($\mathcal{D}_{\text{retain}}$ vs. $\mathcal{D}_{\text{forget}}$), and *Retain vs Forget*
306 ($\mathcal{D}_{\text{retain}}$ vs. $\mathcal{D}_{\text{forget}}$).
307308 In addition to the softmax-based attack, we also evaluate a saliency-map–driven MIA. For this vari-
309 ant, we compute input-gradient saliency maps of the target model and XGBoost classifier to distin-
310 guish the saliency distributions of member and non-member samples, inspired by the attack setup in
311 (Huang et al., 2024).
312313 5.1 EVALUATION METRICS
314315 **Gradual unlearning.** A desirable property of any unlearning evaluation metric is consistency
316 under progressive removal of the forgotten data. That is, as a larger portion of the forget set is
317 preserved in retraining, the unlearning score should increase accordingly. To verify that MIAU
318 exhibits this behavior, we construct partial retraining baselines that simulate intermediate levels of
319 forgetting.
320321 Let $\mathcal{D}_{\text{forget}}^{(p)} \subset \mathcal{D}_{\text{forget}}$ denote a subset comprising a proportion $p \in \{0.25, 0.50, 0.75\}$ of the original
322 forget set. The remaining portion $(1 - p)$ is excluded from the unlearning procedure. The corre-
323 sponding partial retrain set is given by $\mathcal{D}_{\text{retrain}}^{(p)} = \mathcal{D}_{\text{retain}} \cup \mathcal{D}_{\text{forget}}^{(1-p)}$, and the model $\phi_{\theta_{\text{retrain}-p}}$ is trained
324 from scratch on this subset.
325326 These models represent intermediate stages of unlearning and serve as graded reference points be-
327 tween the baseline model (trained on $\mathcal{D}_{\text{retain}} \cup \mathcal{D}_{\text{forget}}$) and the full retrain model (trained solely
328

324 on $\mathcal{D}_{\text{retain}}$). We evaluate MIAU for Retrain 75%, Retrain 50%, and Retrain 25% to validate that
 325 the score increases as a larger proportion of the forget set is preserved. A consistent ordering of
 326 $\text{MIAU}_{25} < \text{MIAU}_{50} < \text{MIAU}_{75} < \text{MIAU}_{\text{full}}$ supports that MIAU faithfully captures the extent of
 327 forgetting.

329 **P-Value test.** To validate the consistency and discriminative capacity of MIAU under progressive
 330 unlearning, we perform statistical hypothesis testing across partial retrain levels. Specifically, we
 331 employ one-sided paired t -tests to evaluate whether the MIAU score at a higher unlearning level
 332 is significantly greater than at a lower level. For example, we test whether $\text{MIAU}_{50} > \text{MIAU}_{25}$,
 333 $\text{MIAU}_{75} > \text{MIAU}_{50}$, and $\text{MIAU}_{75} > \text{MIAU}_{25}$. All tests are conducted over multiple seeds, and we
 334 report the corresponding p -values to determine statistical significance at standard confidence levels.

336 6 DISCUSSION OF RESULTS

338 The results in Table 1 show that the proposed MIAU metric effectively quantifies unlearning perfor-
 339 mance by measuring how closely each method approximates the privacy behavior of a fully retrained
 340 model. For example, Amnesiac and Teacher achieve MIAU scores of 40.07% and 38.36%, respec-
 341 tively, meaning they close approximately 40% of the gap between the baseline model ($\text{MIAU} \approx$
 342 0.10%) and the ideal retraining model ($\text{MIAU} \approx 99.9\%$). In contrast, Finetune closes only 30.89%,
 343 and SSD just 8.55%, indicating less progress toward ideal forgetting. This design overcomes key
 344 limitations of individual MIA scores, which only reflect a single perspective and can be misleading
 345 without context. For instance, Amnesiac’s *Forget vs Test* score is 56.58% and *Forget vs Re-*
 346 *tain* is 55.88%, but without knowing the corresponding baseline or retrain values, these numbers
 347 offer little interpretability. Similarly, accuracy-based metrics like *Forget Accuracy* (e.g., 83.99%
 348 for Amnesiac) and ZRF (95.95%) may not distinguish between targeted forgetting and overall
 349 degradation. Each individual MIA configuration tests a necessary but not sufficient condition for
 350 successful unlearning. MIAU, on the other hand, provides a structured solution by integrating all
 351 three configurations and referencing the performance bounds, making it better suited for evaluating
 352 the completeness and correctness of forgetting. The remaining experimental results for all unlearn-
 353 ing methods across additional datasets and configurations are provided in the Appendix A.7.

354 **Table 1: Experiments on CIFAR-20 AllCNN**

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	90.2862 ± 0.1332	88.5177 ± 0.1718	86.8171 ± 0.6593	80.6776 ± 0.6884	90.3103 ± 0.1159
Test Accuracy	79.8926 ± 0.0000	76.7656 ± 0.3392	76.5400 ± 0.4834	69.2578 ± 0.6123	79.8936 ± 0.0031
Forget Accuracy	90.2968 ± 0.4404	83.9966 ± 0.6690	81.9508 ± 0.7104	78.8545 ± 0.8649	90.3999 ± 0.3387
ZRF Score	91.1786 ± 0.0670	95.9513 ± 0.0945	91.2336 ± 0.3110	96.6893 ± 0.1198	91.1546 ± 0.0700
MIA (Forget vs Retain)	50.2150 ± 0.8686	55.8800 ± 1.3683	51.8950 ± 1.6075	52.5300 ± 0.8687	50.6250 ± 0.9044
MIA (Forget vs Test)	49.6600 ± 0.9703	56.5800 ± 0.8619	51.8750 ± 1.1660	55.4050 ± 0.8880	49.5950 ± 0.6950
MIA (Test vs Retain)	49.7500 ± 0.3450	51.3775 ± 0.7146	50.3700 ± 0.9752	53.1900 ± 0.5757	49.6475 ± 0.5362
MIA (Train vs Test)	54.4750 ± 0.0000	55.5575 ± 0.3060	54.1125 ± 0.7029	53.8525 ± 0.4610	54.4750 ± 0.0000
MIAU	0.1007 ± 0.0000	40.0764 ± 23.3662	30.8884 ± 15.2037	38.3645 ± 20.3545	8.5516 ± 13.4617

366 The gradual unlearning results shown in Tables 2, and 3 confirm that MIAU scores increase mono-
 367 tonically with greater retraining data. For the three experiments, we observe that $\text{MIAU}_{25} <$
 368 $\text{MIAU}_{50} < \text{MIAU}_{75} < \text{MIAU}_{\text{full}}$ validating that MIAU reflects the expected partial unlearning
 369 behavior. This progressive rise aligns with the theoretical property of consistency under partial data
 370 removal, which is not inherently guaranteed by other metrics like forget accuracy or ZRF.

371 Despite this expected pattern in the tables, additional experiments uncover inherent limitations of
 372 MIAs. In particular, we can observe empirically, that the bar graph (Figure 3) shows that for several
 373 datasets, the expected progression $\text{MIAU}_{25} < \text{MIAU}_{50} < \text{MIAU}_{75} < \text{MIAU}_{\text{full}}$ does not hold
 374 consistently. For example, on MNIST_AllCNN and CIFAR10_ResNet, the increase in MIAU is
 375 not strictly observed across retraining levels. Statistically, the p-value heatmap (Figure 4) supports
 376 these irregularities: only a subset of datasets show meaningful p-values (< 0.05) for comparisons
 377 like $\text{MIAU}_{75} > \text{MIAU}_{25}$ or $\text{MIAU}_{75} > \text{MIAU}_{50}$. This suggests that individual MIA components
 (which our MIAU depends on) may still be unreliable in isolation. In particular, the bar graph

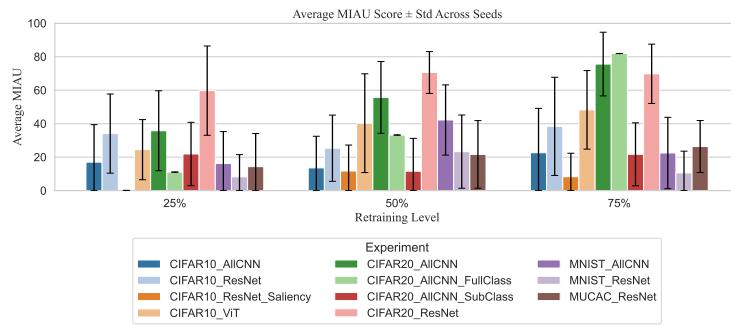


Figure 3: Average MIAU scores across 10 random seeds for each dataset at three retraining levels: 25%, 50%, and 75%.

Table 2: Gradual unlearning on CIFAR-10 ViT

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	98.7300 ± 0.4185	92.1279 ± 21.3489	89.9830 ± 28.0999	83.3823 ± 33.0112
Test Accuracy	97.1084 ± 0.3366	90.4990 ± 21.2237	88.5742 ± 27.5882	82.2012 ± 32.2264
Forget Accuracy	95.9102 ± 0.3428	89.7123 ± 20.3195	87.5362 ± 27.3287	81.1884 ± 31.9368
ZRF Score	77.2600 ± 0.6070	78.8569 ± 6.6877	76.7375 ± 0.6512	78.7411 ± 7.2286
MIA (Forget vs Retain)	50.3000 ± 1.8720	53.1400 ± 3.8696	52.3667 ± 1.9312	52.5300 ± 2.2833
MIA (Forget vs Test)	52.2600 ± 1.5665	52.9000 ± 1.2944	52.4200 ± 1.4104	51.8600 ± 1.0723
MIA (Test vs Retain)	50.9500 ± 1.1756	51.2375 ± 0.9500	51.2750 ± 1.3260	51.1075 ± 1.0631
MIA (Train vs Test)	51.2500 ± 0.4518	51.4275 ± 0.7186	51.2325 ± 0.6836	51.0725 ± 0.6412
MIAU	24.4963 ± 17.9590	40.3098 ± 29.5193	48.2700 ± 23.4669	99.8993 ± 0.0000

(Figure 3) reveals substantial standard deviation across seeds for the same retraining level, indicating instability in per-run MIA behavior that may propagate into MIAU.

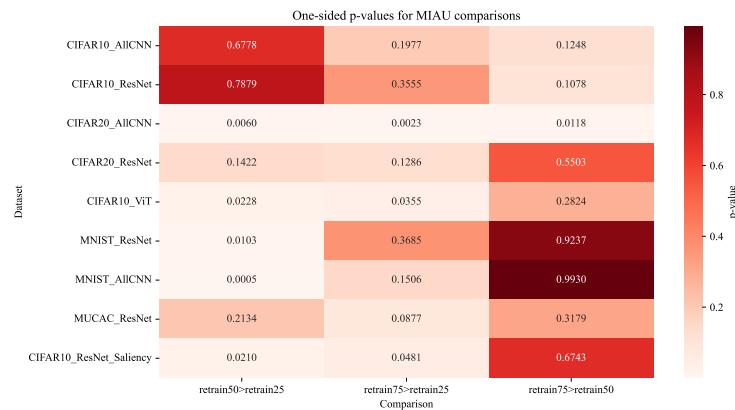


Figure 4: One-sided p-values from paired t-tests comparing MIAU scores between successive retraining levels across multiple datasets. Each cell reflects the statistical significance of whether the MIAU score from a higher retraining level is significantly greater than that of a lower one.

Further limitations may arise in scenarios where the model exhibits strong generalization across all data splits. As shown in Figure 5, MIA score distributions before and after unlearning often remain closely aligned, indicating minimal separability even when full retraining is performed. High variance in certain scores, especially under simple datasets like MNIST (see Figure 2), also implies that models with low memorization may naturally yield less distinct MIA signals. These observa-

432

433

Table 3: Gradual unlearning on MUCAC ResNet-18

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	93.4748 ± 2.1251	93.4056 ± 1.2176	92.8468 ± 2.3045	91.9656 ± 4.6271
Test Accuracy	93.4268 ± 1.8102	93.7485 ± 0.9387	93.1357 ± 1.9662	92.5286 ± 3.8284
Forget Accuracy	92.5586 ± 4.2547	92.9687 ± 0.8326	91.5788 ± 2.5153	90.9297 ± 4.5826
ZRF Score	76.1309 ± 4.2882	78.0748 ± 4.4559	78.4498 ± 6.1267	79.1186 ± 6.2524
MIA (Forget vs Retain)	50.0000 ± 3.0162	49.4811 ± 3.2181	51.3249 ± 3.7769	50.4976 ± 1.9714
MIA (Forget vs Test)	53.5849 ± 4.1433	54.3396 ± 1.7338	55.7098 ± 2.1354	53.8863 ± 2.1638
MIA (Test vs Retain)	51.5254 ± 1.5284	51.9855 ± 1.1699	52.8208 ± 1.8111	52.3729 ± 1.0212
MIA (Train vs Test)	53.7772 ± 1.4660	52.7240 ± 0.8974	52.7119 ± 0.6782	52.7240 ± 1.5242
MIAU	14.3358 ± 19.8046	21.6582 ± 20.1905	26.3585 ± 15.5590	99.8993 ± 0.0000

439

440

441

442

443

444

445

446

tions suggest that in well-generalized regimes, MIAs may become less sensitive as an unlearning diagnostic, and supplementary indicators may be needed to confirm forgetting efficacy.

Detailed results of all experiment are outlined in Appendix A.7.

450

451

452

453

454

455

456

457

458

459

460

461

462

463

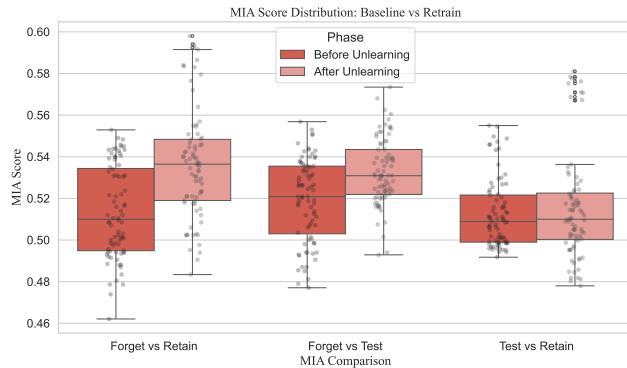


Figure 5: Comparison of MIA score distributions before and after unlearning. The figure illustrates the distributions of Membership Inference Attack (MIA) scores for three comparisons—*Forget vs Retain*, *Forget vs Test*, and *Test vs Retain*—across the *baseline* and *retrain* phases across all experiments.

468

469

470

7 CONCLUSION

471

472

473

474

475

476

477

478

479

This paper introduces the Membership Inference Attack Unlearning Score (MIAU), metric designed to provide more structured and interpretable performance assessment of machine unlearning methods. Unlike prior evaluation approaches that rely on a single MIA configuration or raw accuracy measures, MIAU integrates multiple attack comparisons—*Forget vs Test*, *Forget vs Retain*, and *Retain vs Test*. It further relates them to the baseline and fully retrained models to quantify the degree of gap closure. Through experiments across diverse datasets and model architectures, we illustrate that MIAU aligns with desirable properties for unlearning evaluation, including consistency under progressive removal and clear separation between effective and ineffective methods.

480

481

482

483

484

485

At the same time, our results highlight potential limitations in relying solely on MIAs for evaluation. In particular, we observe that individual MIA scores can be unstable across seeds and less informative for highly generalized models. These findings suggest that future work may investigate augmenting MIAU with models’ internal behavioral indicators, such as latent space drift, neuron activation shifts, or feature attribution dynamics. As a result, these directions may help establish a more reliable understanding of the metric’s sensitivity and consistency under varied unlearning regimes.

486
487
LLM USAGE488
489
490
491
492
493
494
Large Language Models (LLMs) were used only as a general-purpose writing assistant. They helped
with grammar correction, phrasing, and minor style edits after the technical content, experiments,
and analyses were completed by the authors. No part of the research ideation, methodology design
or data analysis was generated by an LLM.495
496
497
498
499
500
501
502
503
504
505
REPRODUCIBILITY STATEMENT506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
To enable faithful reproduction of all results, we provide a complete specification of training and un-
learning settings, implementation artifacts, and per-seed outputs. The full hyperparameter schedule
is listed in Appendix A.5.4; dataset preprocessing and construction details (including any remapping
and normalization) are in Appendix A.5.3; hardware and software environments are documented in
Appendices A.4.1 and A.4.2; and training/validation loss traces used to monitor convergence are
shown in Appendix A.5.5. The accompanying code and usage documentation (entry-point scripts,
configuration files, and commands to regenerate tables and figures) are provided in the supplemen-
tary file `Code_Appendix.zip`. All per-seed experimental outputs—including metrics, logs, and
CSVs of split indices—are contained in the supplementary file `Data_Appendix.zip`; using the
fixed seeds and instructions included in these archives enables full reproduction of every figure and
table in the paper.506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
REFERENCES508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin
Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE
Symposium on Security and Privacy (SP)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.
Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015
IEEE Symposium on Security and Privacy (SP)*, pp. 463–480. IEEE, 2015.
Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang.
When Machine Unlearning Jeopardizes Privacy. In *Proceedings of the 2021 ACM SIGSAC
Conference on Computer and Communications Security (CCS '21)*, pp. 896–911, 2021a. doi:
10.1145/3460120.3484756.
Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang.
When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Con-
ference on Computer and Communications Security (CCS)*, pp. 896–911, 2021b.
Jiali Cheng and Hadi Amiri. Mu-bench: A multitask multimodal benchmark for machine unlearning,
2024. URL <https://arxiv.org/abs/2406.14796>.
Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal
identities in facial recognition systems. *CoRR*, abs/2311.02240, 2023.
Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot ma-
chine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354,
2023a. doi: 10.1109/TIFS.2023.3265506.
Vikram Singh Chundawat, Ayush Kumar Tarun, Murari Mandal, and Mohan Kankanhalli. Can
Bad Teaching Induce Forgetting? Unlearning in Deep Networks using an Incompetent Teacher.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217,
2023b.
Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszko-
reit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at
scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.
European Union. General Data Protection Regulation (GDPR) Article 17: Right to Erasure.
<https://gdpr-info.eu/art-17-gdpr/>, 2016.

540 Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast Machine Unlearning without Retraining
 541 through Selective Synaptic Dampening. In *Proceedings of the AAAI Conference on Artificial*
 542 *Intelligence*, volume 38, 2024.

543 Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data
 544 deletion in machine learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS*
 545 2019), pp. 3513–3526, 2019.

546 Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net:
 547 Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer*
 548 *Vision and Pattern Recognition (CVPR)*, pp. 9304–9312, 2020.

549 Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto.
 550 Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on*
 551 *Computer Vision and Pattern Recognition (CVPR)*, pp. 792–801, 2021.

552 Lukas Graves, Vasilis Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of*
 553 *the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

554 Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal
 555 from machine learning models. In *Proceedings of the 37th International Conference on Machine*
 556 *Learning (ICML)*, pp. 3832–3842, 2020.

557 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-
 558 nition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.
 559 770–778, 2016.

560 Catherine Huang, Martin Pawelczyk, and Himabindu Lakkaraju. Explaining the model, protecting
 561 your data: Revealing and mitigating the data privacy risks of post-hoc model explanations via
 562 membership inference, 2024. URL <https://arxiv.org/abs/2407.17663>.

563 Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma,
 564 and Sijia Liu. Model Sparsity Can Simplify Machine Unlearning. In *Advances in Neural Infor-*
 565 *mation Processing Systems (NeurIPS) 36*, 2023.

566 Michael I. Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects.
 567 *Science*, 349(6245):255–260, 2015.

568 Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Tech-
 569 nical report, University of Toronto, 2009.

570 Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards Unbounded
 571 Machine Unlearning. In *Advances in Neural Information Processing Systems, Vol. 36 (NeurIPS*
 572 2023), 2023.

573 Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied
 574 to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.
 575 726791.

576 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444,
 577 2015.

578 Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miao Zeng Du, Yongrui Chen, Sheng Bi, and
 579 Fan Liu. Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language
 580 Models. In *Advances in Neural Information Processing Systems (NeurIPS) 37*, 2024.

581 Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin,
 582 and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*,
 583 2022.

584 Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference at-
 585 tacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*,
 586 pp. 3–18. IEEE, 2017.

594 Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for
 595 simplicity: The all convolutional net. In *Proc. 3rd International Conference on Learning Repre-*
 596 *sentations, Workshop Track*, 2015. URL <http://arxiv.org/abs/1412.6806>.

598

599 A APPENDIX

600

601 A.1 CHOICE OF ALPHA PARAMETER

602

603 To ensure that the MUS score maps the gap closure fraction f_i to values close to 0 for baseline
 604 models and close to 100 for fully retrained models, we select the steepness parameter α in the
 605 logistic transformation accordingly.

606 Recall that the MUS score is defined as:

607

$$608 \text{MUS}_i = 100 \cdot \frac{1}{1 + e^{-\alpha(f_i - 0.5)}} \quad (4)$$

609

610 We require the MUS score to satisfy:

611

$$\text{MUS}_i(f_i = 0) \approx 0.1 \quad (5)$$

612

613

$$\text{MUS}_i(f_i = 1) \approx 99.9 \quad (6)$$

614

615 Substituting $f_i = 0$ into Equation (1), we get:

616

$$617 100 \cdot \frac{1}{1 + e^{\alpha \cdot 0.5}} = 0.1 \quad (7)$$

618

619

$$\Rightarrow \frac{1}{1 + e^{0.5\alpha}} = 0.001 \quad (8)$$

620

621

$$\Rightarrow e^{0.5\alpha} = 999 \quad (9)$$

622

Solving for α :

624

$$0.5\alpha = \ln(999) \quad (10)$$

625

626

$$\Rightarrow \alpha = 2 \cdot \ln(999) \approx 2 \cdot 6.9068 = 13.8136 \quad (11)$$

627

Note that Equation (3) holds automatically by symmetry of the logistic function, given the same
 choice of α derived in Equation (2).

629 Therefore, setting $\alpha = 13.8$ ensures that the MUS score yields values close to 0.1 and 99.9 for the
 630 endpoints $f_i = 0$ and $f_i = 1$, respectively. This choice results in a sharp transition around $f_i = 0.5$
 631 while maintaining bounded scores within $(0, 100)$, effectively amplifying the distinction between
 632 poorly and effectively unlearned models.

633

634 A.2 PERFORMANCE OF STRONGER MIA ATTACKS

635

637 Table 4: Train vs. Test MIA Attack Performance on CIFAR-10 ResNet-18

638

639

MIA Attack	Score (AUC)
LIRA	44 ± 5.8721
Shadow Model	60 ± 15.883
Quantile Regression	50.9700 ± 9.781

640

641

642

643

644

645 A.3 EFFECT SIZES AND CONFIDENCE INTERVALS

646

647

648
649
650

Table 5: Pairwise retraining–level comparisons across datasets

651	Dataset	Comparison	n	Mean diff	95% CI lower	95% CI upper	Cohen’s d	p-value
652	CIFAR10_AliCNN	retrain50 > retrain25	10	-3.440051	-19.742765	12.862664	-0.150948	0.677751
653	CIFAR10_AliCNN	retrain75 > retrain25	10	5.617811	-8.623302	19.858924	0.282193	0.197713
654	CIFAR10_AliCNN	retrain75 > retrain50	10	9.057862	-7.586946	25.702669	0.389287	0.124755
655	CIFAR10_ResNet	retrain50 > retrain25	10	-8.750156	-32.399339	14.899026	-0.264681	0.787869
656	CIFAR10_ResNet	retrain75 > retrain25	10	4.321670	-21.245215	29.888556	0.120920	0.355526
657	CIFAR20_AliCNN	retrain50 > retrain25	10	19.914799	5.568661	34.260938	0.993033	0.005962
658	CIFAR20_AliCNN	retrain75 > retrain25	10	39.839401	15.771734	63.907069	1.184136	0.002297
659	CIFAR20_AliCNN	retrain75 > retrain50	10	19.924602	3.345701	36.503503	0.859719	0.011830
660	CIFAR20_ResNet	retrain50 > retrain25	10	10.822943	-10.688412	32.334298	0.359915	0.142223
661	CIFAR20_ResNet	retrain75 > retrain25	10	10.051177	-8.748205	28.850559	0.382469	0.128646
662	CIFAR10_ViT	retrain50 > retrain25	10	-0.771766	-14.208242	12.664710	-0.041089	0.550262
663	CIFAR10_ViT	retrain75 > retrain25	10	15.813495	0.375409	31.251582	0.732752	0.022849
664	MNIST_ResNet	retrain50 > retrain25	10	23.773717	-2.499854	50.047288	0.647293	0.035483
665	MNIST_ResNet	retrain75 > retrain50	10	7.960222	-22.164509	38.084952	0.189027	0.282369
666	MNIST_AliCNN	retrain50 > retrain25	10	15.072583	2.911955	27.233210	0.886655	0.010291
667	MNIST_AliCNN	retrain75 > retrain25	10	2.361378	-13.059572	17.782327	0.109541	0.368502
668	MNIST_AliCNN	retrain50 > retrain25	10	-12.711205	-31.113758	5.691348	-0.494119	0.923702
669	MNIST_AliCNN	retrain75 > retrain25	10	26.016491	13.618136	38.414846	1.501092	0.000524
670	MNIST_AliCNN	retrain50 > retrain25	10	6.293896	-6.688195	19.275986	0.346815	0.150619
671	MNIST_AliCNN	retrain75 > retrain25	10	-19.722595	-34.374532	-5.070659	-0.962924	0.993048
672	MUCAC_ResNet	retrain50 > retrain25	10	7.322414	-12.579052	27.223880	0.263204	0.213380
673	MUCAC_ResNet	retrain75 > retrain25	10	12.022749	-6.469253	30.514751	0.465096	0.087716
674	MUCAC_ResNet	retrain50 > retrain25	10	4.700335	-16.996604	26.397275	0.154972	0.317906
675	CIFAR10_ResNet_Saliency	retrain50 > retrain25	10	11.629649	0.526751	22.732547	0.749295	0.020972
676	CIFAR10_ResNet_Saliency	retrain75 > retrain25	10	8.229491	-1.789463	18.248445	0.587589	0.048050
677	CIFAR10_ResNet_Saliency	retrain75 > retrain50	10	-3.400158	-19.863704	13.063389	-0.147740	0.674268

A.4 COMPUTING INFRASTRUCTURE

A.4.1 HARDWARE SPECIFICATIONS

Table 6: Hardware specifications of the Google Colab Pro+ environment with NVIDIA A100.

678	Component	Specification
679	GPU	NVIDIA A100-SXM4-40GB, 40 GB HBM2
680	CUDA Cores	6,912
681	Tensor Cores	432 (3rd Generation)
682	GPU Memory Bandwidth	1.6 TB/s
683	CUDA Version	12.4 (runtime), 12.5 (nvcc compiler)
684	Driver Version	550.54.15
685	CPU	Intel(R) Xeon(R) @ 2.20GHz (12 vCPUs)
686	Threads per Core	2
687	Host RAM	87.5 GB system memory
688		

A.4.2 SOFTWARE SPECIFICATIONS

691
692
693
694
695
696
697
698
699
700
701

Table 7: Software packages and versions used in the Google Colab Pro+ A100 environment.

Library / Component	Version
Operating System	Ubuntu 22.04.4 LTS (Jammy Jellyfish)
Python	3.11.13
PyTorch	2.6.0+cu124
Torchvision	0.21.0+cu124
CUDA Toolkit	12.4 (linked), 12.5 (compiler nvcc)
NumPy	2.0.2
Pandas	2.2.2
Matplotlib	3.10.0
Pillow	11.3.0
scikit-learn	1.6.1
XGBoost	3.0.2
Transformers (HuggingFace)	4.54.0
TQDM	4.67.1
Seaborn	0.13.2
SciPy	1.16.0
Requests	2.32.3

A.5 TRAINING CONFIGURATIONS

A.5.1 DATASETS

We evaluate on four image classification datasets (MNIST, Cifar-10, Cifar-20, MUCAC) with varying resolution, label structure, and domain characteristics. *MNIST* (LeCun et al., 1998) is a hand-written digit dataset with 10 classes (0–9), containing 60,000 training and 10,000 test grayscale images of size 28×28 , uniformly distributed. *CIFAR-10* (Krizhevsky & Hinton, 2009) consists of natural 32×32 color images across 10 object categories, with 50,000 training and 10,000 test samples. *CIFAR-20* (Krizhevsky & Hinton, 2009) is a coarse-label variant of *CIFAR-100*, containing 20 superclasses (e.g., insects, vehicles) with 50,000 training and 10,000 test images, evenly split across classes. *MUCAC* (*Machine Unlearning for Celebrity Attribute Classifier*) (Choi & Na, 2023) is a binary attribute classification dataset derived from CelebA-HQ, used to evaluate unlearning on face attributes. For this study, we focus only on *Smiling* attribute as an independent binary classification task. The dataset contains of 10,548 training and 2,065 test samples, with 5284 negative and 4210 positive labels in each case.

A.5.2 BASE MODELS

We evaluate unlearning performance across three representative architectures: ResNet-18, All-CNN, and Vision Transformer (ViT). *ResNet-18* (He et al., 2016) is a deep convolutional network utilizing residual connections to facilitate optimization and gradient flow. *All-CNN* (Springenberg et al., 2015) is a fully convolutional architecture with no max-pooling, preserving spatial locality and emphasizing robustness. *ViT* (Dosovitskiy et al., 2021) replaces convolutions with self-attention mechanisms, modeling global dependencies via patch embeddings and transformer layers.

A 5.3 DATA PREPROCESSING

We adopt dataset-specific preprocessing strategies to enhance generalization performance and ensure consistency across training, unlearning, and evaluation phases. All image data are normalized to zero mean and unit variance using dataset-specific statistics.

For training from scratch on CIFAR-10 and CIFAR-20 datasets, we apply random cropping (with 4-pixel padding), random horizontal flipping, and random rotation of up to 15 degrees, followed by normalization using the dataset-specific mean (0.5071, 0.4865, 0.4409) and standard deviation (0.2673, 0.2564, 0.2762) values. For unlearning and test phases, no augmentation is applied; only normalization is used to maintain evaluation consistency. CIFAR-20 is constructed by remapping the

756 100 fine labels of CIFAR-100 into 20 coarse classes using a predefined superclass mapping derived
 757 from prior work.

758 MNIST images are first converted to three-channel grayscale format to align with the input expectations
 759 of RGB-based models. Training images are augmented with random rotation of up to 10
 760 degrees. Test and unlearning images are not augmented. All images are normalized using a mean of
 761 0.1307 and a standard deviation of 0.3081.

762 For MUCAC (CelebA-HQ) dataset, each face image is resized to 128×128 pixels. Training data
 763 are augmented with random horizontal flipping, affine transformation (shear angle of 10 degrees and
 764 scale factor between 0.8 and 1.2), and color jittering (brightness, contrast, and saturation set to 0.2).
 765 For unlearning and test data, only resizing and normalization are applied. The binary label is derived
 766 from the “smiling” attribute in the CelebA-HQ metadata. Images are split into training, forget, and
 767 test sets based on person identity ranges to enforce disjoint subsets.

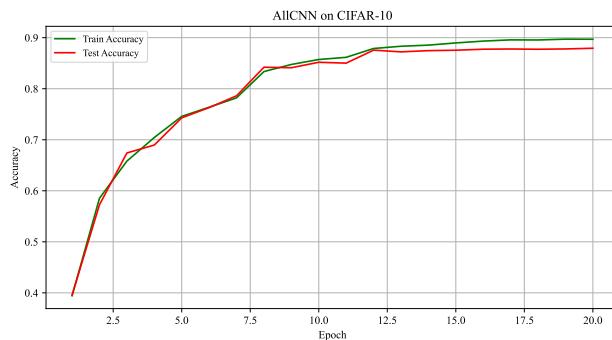
768 A.5.4 LIST OF HYPERPARAMETERS

769
 770
 771 Table 8: Summary of hyperparameters used for training and unlearning.

772 Hyperparameter	773 Value
774 Batch size (\mathcal{B})	256
775 Unlearning batch size	128
776 Initial learning rate (η_0)	0.1
777 Optimizer	SGD
778 Momentum	0.9
779 Weight decay	5×10^{-4}
780 Loss function	CrossEntropyLoss
781 Learning rate scheduler	MultiStepLR
782 Scheduler gamma (γ)	0.2
783 Warmup epochs	1
784 CIFAR-10 epochs	20
785 CIFAR-10 milestones	[8, 12, 16]
786 CIFAR-20 epochs	40
787 CIFAR-20 milestones	[15, 30, 35]
788 MNIST epochs	5
789 MNIST milestones	[2, 3, 4]
790 MUCAC epochs	31
791 MUCAC milestones	[10, 20]
792 CIFAR-10 (ViT) epochs	8
793 CIFAR-10 (ViT) milestones	[7]
794 Fine-tuning epochs	5
795 Fine-tuning learning rate	0.02
796 Amnesiac unlearning epochs	3
797 Amnesiac learning rate	0.0001
798 Dampening constant	1
799 Selection weighting	$10 \times \text{model.size.scaler}$ (default = 10)
800 Model size scaler	1
801 Device	GPU

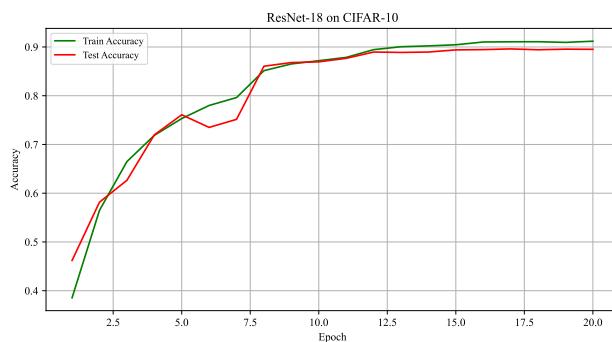
802 803 A.5.5 TRAINING AND TEST LOSS ANALYSIS

810
811
812
813
814
815
816
817
818
819
820
821
822
823



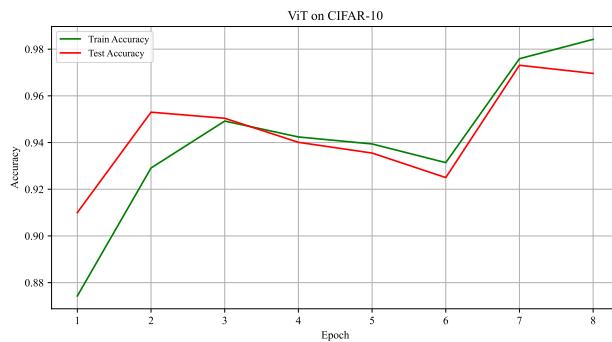
824 Figure 6: Training and validation loss and accuracy curves of the AllCNN model on the CIFAR-10
825 dataset.

826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841



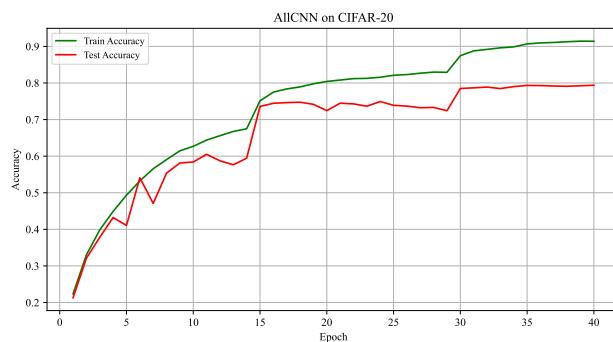
842 Figure 7: Training and validation loss and accuracy curves of the ResNet-18 model on the CIFAR-
843 10 dataset.

844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863



861 Figure 8: Training and validation loss and accuracy curves of the ViT model on the CIFAR-10
862 dataset.

864
865
866
867
868
869
870
871
872
873
874
875
876
877



878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

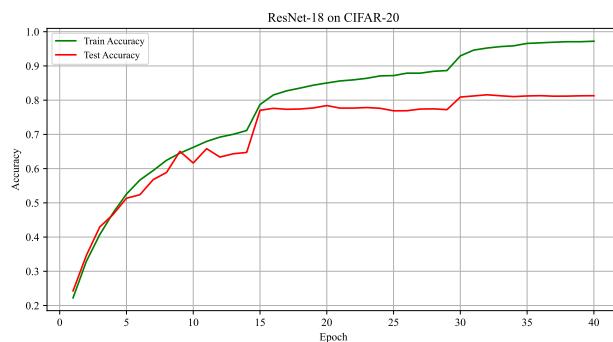


Figure 10: Training and validation loss and accuracy curves of the ResNet-18 model on the CIFAR-20 dataset.

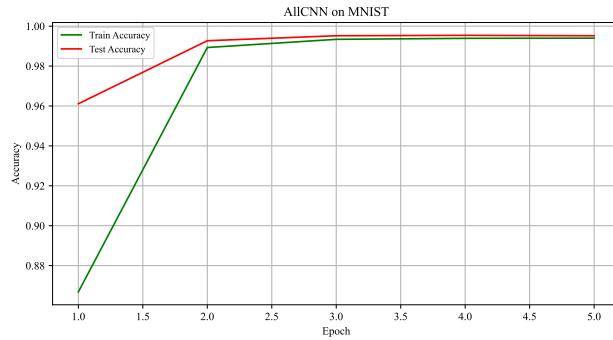


Figure 11: Training and validation loss and accuracy curves of the AllCNN model on the MNIST dataset.

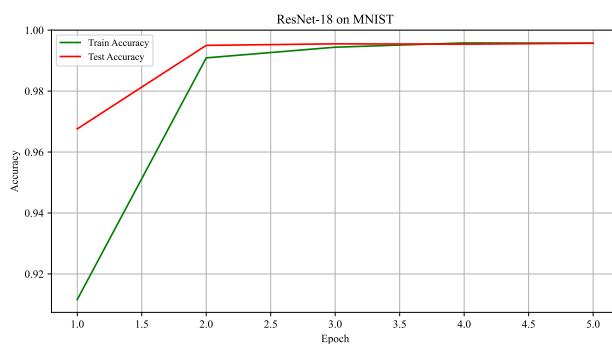


Figure 12: Training and validation loss and accuracy curves of the ResNet-18 model on the MNIST dataset.

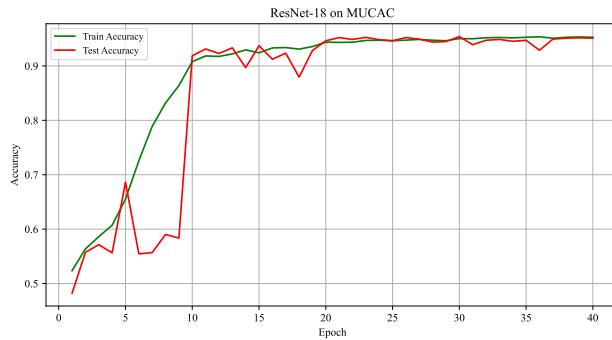


Figure 13: Training and validation loss and accuracy curves of the ResNet-18 model on the MUCAC dataset.

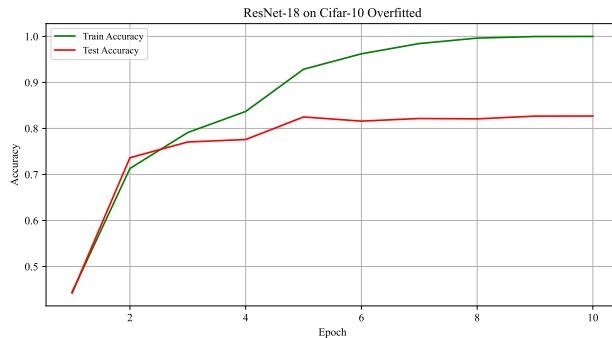
972 A.6 RESULTS OF EXPERIMENTS ON NON-GENERALIZED MODELS
973974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
Figure 14: Training and validation loss and accuracy curves of the overfitted ResNet-18 model on the CIFAR-10 dataset. An underfitted model was obtained by taking the 1st epoch model, while overfitted obtained taking the 10th epoch model.

Table 9: Experiments on CIFAR10 underfitted ResNet-18

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	44.8112 ± 0.0532	88.6696 ± 0.8549	90.2865 ± 2.0526	44.8467 ± 0.8808	44.8083 ± 0.0528
Test Accuracy	43.7793 ± 0.0000	76.9658 ± 0.8425	75.9062 ± 1.6650	43.8945 ± 0.7836	43.7793 ± 0.0000
Forget Accuracy	45.4627 ± 0.4842	71.4473 ± 1.1740	77.1919 ± 1.8680	45.6693 ± 1.0187	45.4247 ± 0.5211
ZRF Score	90.7447 ± 0.8915	95.6467 ± 0.5602	80.4635 ± 0.9641	98.3065 ± 0.1286	90.7469 ± 0.8908
MIA (Forget vs Retain)	56.2250 ± 0.8898	62.5450 ± 2.3834	55.2100 ± 4.2645	52.9350 ± 1.9764	56.4500 ± 1.0247
MIA (Forget vs Test)	50.0650 ± 0.6377	58.5450 ± 0.9576	49.2550 ± 0.7500	50.9700 ± 0.6638	49.9200 ± 0.6273
MIA (Test vs Retain)	60.0875 ± 0.2875	56.3450 ± 2.7112	58.0300 ± 4.0428	57.1175 ± 1.5528	60.0875 ± 0.2875
MIA (Train vs Test)	50.2000 ± 0.0000	53.5025 ± 0.4382	54.3250 ± 0.9622	50.5850 ± 0.4961	50.2000 ± 0.0000
MIAU	0.1007 ± 0.0000	6.9812 ± 13.4431	14.8824 ± 21.2491	31.2930 ± 22.8420	13.7048 ± 16.9481

Table 10: Gradual unlearning on CIFAR-10 underfitted ResNet-18

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	45.9210 ± 8.1250	48.7542 ± 8.5905	45.3409 ± 4.5958	47.8767 ± 6.4984
Test Accuracy	45.1240 ± 7.8977	47.6611 ± 8.5742	44.5781 ± 4.5092	46.8799 ± 6.5235
Forget Accuracy	45.2158 ± 8.2025	48.2742 ± 8.8719	45.0027 ± 4.0382	47.3834 ± 6.2705
ZRF Score	89.4216 ± 2.7347	89.5791 ± 4.4070	89.7330 ± 2.8897	90.6504 ± 1.5962
MIA (Forget vs Retain)	60.3400 ± 5.4740	55.7600 ± 4.6058	54.4133 ± 3.7214	54.1800 ± 2.9321
MIA (Forget vs Test)	49.5200 ± 2.3117	50.5200 ± 1.7093	50.1200 ± 1.4208	50.1000 ± 0.4859
MIA (Test vs Retain)	59.9825 ± 4.5312	57.5825 ± 3.5780	58.4325 ± 3.9309	57.7300 ± 2.9314
MIA (Train vs Test)	49.6600 ± 0.5480	50.0700 ± 0.5820	50.0150 ± 0.5300	49.8100 ± 0.6517
MIAU	4.1438 ± 9.7553	21.9061 ± 19.9868	9.9082 ± 12.7532	96.5727 ± 10.5197

Table 11: Experiments on CIFAR-10 overfitted ResNet-18

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	99.9947 \pm 0.0019	95.6966 \pm 0.8870	95.9652 \pm 1.4784	95.6319 \pm 1.2804	99.9947 \pm 0.0019
Test Accuracy	82.8125 \pm 0.0000	73.3311 \pm 1.1328	78.1406 \pm 1.2595	78.6924 \pm 0.9585	82.8125 \pm 0.0000
Forget Accuracy	99.9883 \pm 0.0165	38.1286 \pm 2.3546	84.2821 \pm 1.9507	90.9142 \pm 1.6170	99.9883 \pm 0.0165
ZRF Score	75.2012 \pm 0.3465	94.4499 \pm 0.4395	73.7953 \pm 1.2641	97.5783 \pm 0.2209	75.2043 \pm 0.3492
MIA (Forget vs Retain)	52.8300 \pm 1.3756	68.5450 \pm 7.5131	58.1250 \pm 4.6867	55.5100 \pm 3.2535	52.5500 \pm 1.2530
MIA (Forget vs Test)	62.0400 \pm 0.5739	63.3900 \pm 1.2677	50.6100 \pm 1.0011	52.6500 \pm 0.5497	61.8750 \pm 0.4152
MIA (Test vs Retain)	60.6925 \pm 0.3283	57.9825 \pm 5.9698	59.0950 \pm 3.6548	53.6175 \pm 2.8164	60.6925 \pm 0.3283
MIA (Train vs Test)	62.6500 \pm 0.0000	59.5725 \pm 0.7982	58.0825 \pm 0.8123	54.5425 \pm 0.5683	62.6500 \pm 0.0000
MIAU	0.1007 \pm 0.0000	15.8030 \pm 15.3861	56.2321 \pm 15.9836	41.1594 \pm 11.0159	0.1316 \pm 0.0363

Table 12: Gradual unlearning on CIFAR-10 overfitted ResNet-18

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	99.9933 \pm 0.0044	99.9931 \pm 0.0037	99.9959 \pm 0.0033	99.9956 \pm 0.0033
Test Accuracy	82.6396 \pm 0.4758	82.4072 \pm 0.1979	82.2500 \pm 0.3999	82.0928 \pm 0.2645
Forget Accuracy	83.5998 \pm 0.6553	83.3224 \pm 0.6759	82.6646 \pm 0.6620	82.6560 \pm 0.9148
ZRF Score	77.2613 \pm 0.7057	77.5656 \pm 0.4064	77.8443 \pm 0.4585	78.0136 \pm 0.5318
MIA (Forget vs Retain)	59.2200 \pm 1.8890	60.5200 \pm 1.2372	61.7200 \pm 1.4476	61.4100 \pm 1.2142
MIA (Forget vs Test)	49.8400 \pm 1.8638	50.0900 \pm 1.5624	49.4067 \pm 1.6530	50.2250 \pm 0.8193
MIA (Test vs Retain)	61.6125 \pm 0.6727	61.8800 \pm 0.5330	62.1150 \pm 0.5953	62.0000 \pm 0.6001
MIA (Train vs Test)	62.8850 \pm 0.5599	63.0250 \pm 0.5394	63.5125 \pm 0.5562	62.7875 \pm 0.3520
MIAU	69.2569 \pm 14.8576	83.0650 \pm 15.9770	81.5254 \pm 20.3745	99.8993 \pm 0.0000

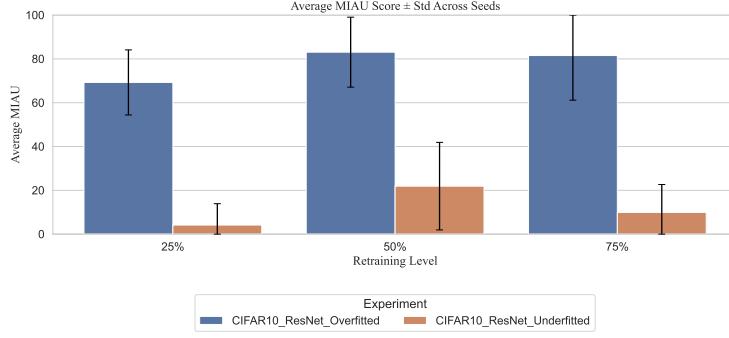
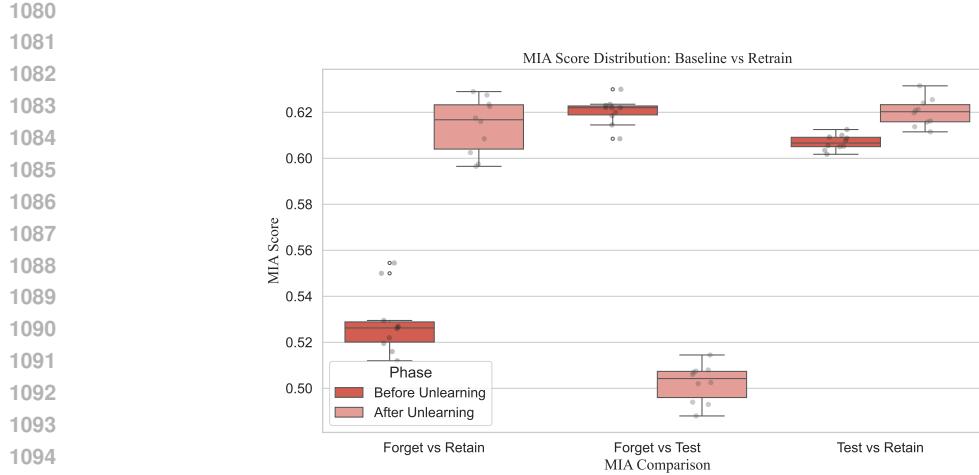
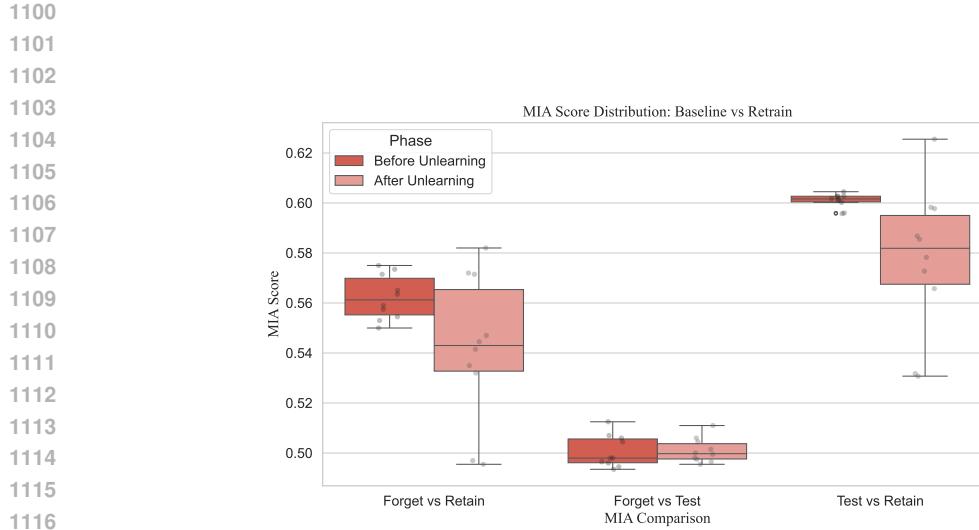


Figure 15: Average MIAU scores across 10 random seeds for underfitted and overfitted models at three retraining levels: 25%, 50%, and 75%.



1096
1097
1098
1099

Figure 16: Comparison of MIA Score Distributions Before and After Unlearning for an overfitted model. The figure illustrates the distributions of Membership Inference Attack (MIA) scores for three comparisons—*Forget vs Retain*, *Forget vs Test*, and *Test vs Retain*—across the *baseline* and *retrain* phases across all experiments.



1118
1119
1120
1121

Figure 17: Comparison of MIA Score Distributions Before and After Unlearning for an underfitted model. The figure illustrates the distributions of Membership Inference Attack (MIA) scores for three comparisons—*Forget vs Retain*, *Forget vs Test*, and *Test vs Retain*—across the *baseline* and *retrain* phases across all experiments.

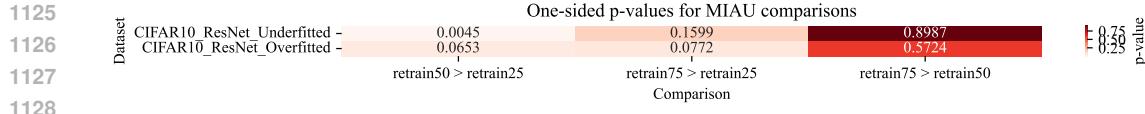


Figure 18: One-sided p-values from paired t-tests comparing MIAU scores between successive retraining levels across overfitted and underfitted models. Each cell reflects the statistical significance of whether the MIAU score from a higher retraining level is significantly greater than that of a lower one.

1134 A.7 REMAINING RESULTS OF EXPERIMENTS
1135
11361137 Table 13: Experiments on CIFAR-10 AllCNN
1138

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	91.8457 \pm 0.1097	90.5320 \pm 0.2146	90.8454 \pm 0.6374	84.1013 \pm 0.4772	91.7852 \pm 0.1036
Test Accuracy	89.5410 \pm 0.0000	87.9102 \pm 0.4012	88.2480 \pm 0.5962	80.9824 \pm 0.8379	89.5410 \pm 0.0000
Forget Accuracy	91.8913 \pm 0.4222	88.3699 \pm 0.6600	88.6864 \pm 0.7774	83.4465 \pm 0.8203	91.9906 \pm 0.2655
ZRF Score	84.2151 \pm 0.2102	94.0819 \pm 0.2160	83.3740 \pm 0.5682	95.8960 \pm 0.0819	84.2038 \pm 0.2136
MIA (Forget vs Retain)	52.7100 \pm 1.0775	52.3600 \pm 3.7283	51.8750 \pm 1.9332	53.7250 \pm 1.9362	52.2300 \pm 1.3290
MIA (Forget vs Test)	54.0900 \pm 0.6927	58.1200 \pm 0.9449	53.9700 \pm 0.7262	58.5950 \pm 0.8224	53.7800 \pm 0.6808
MIA (Test vs Retain)	50.1400 \pm 0.3710	53.8825 \pm 1.6900	51.3000 \pm 0.9587	55.2550 \pm 1.3330	50.3025 \pm 0.6309
MIA (Train vs Test)	51.8250 \pm 0.0000	52.3575 \pm 0.3939	51.7200 \pm 0.6360	51.8200 \pm 0.2986	51.8250 \pm 0.0000
MIAU	0.1007 \pm 0.0000	0.0263 \pm 0.0796	12.9191 \pm 14.5091	6.3614 \pm 13.4248	19.4015 \pm 17.2965

1144

1145

1146

1147

1148 Table 14: Gradual unlearning on CIFAR-10 AllCNN
1149

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	90.0336 \pm 0.1156	89.8742 \pm 0.1189	89.5516 \pm 0.1528	89.4102 \pm 0.1702
Test Accuracy	87.9609 \pm 0.1674	87.8555 \pm 0.2879	87.5352 \pm 0.2244	87.3057 \pm 0.2646
Forget Accuracy	86.7752 \pm 0.9785	87.0890 \pm 0.7315	86.2750 \pm 0.5589	86.0535 \pm 0.4969
ZRF Score	85.5322 \pm 0.2350	85.6090 \pm 0.1626	85.8212 \pm 0.1960	85.8925 \pm 0.2329
MIA (Forget vs Retain)	51.2400 \pm 2.6446	51.1200 \pm 1.7390	50.6800 \pm 1.7367	52.4300 \pm 1.4074
MIA (Forget vs Test)	55.8200 \pm 2.0099	55.8500 \pm 1.1404	55.4733 \pm 0.8500	54.8500 \pm 0.8981
MIA (Test vs Retain)	50.3650 \pm 0.5922	50.5100 \pm 0.6868	50.1425 \pm 0.6928	50.1000 \pm 0.7773
MIA (Train vs Test)	51.6600 \pm 0.2227	51.7225 \pm 0.3754	51.4575 \pm 0.3313	51.5100 \pm 0.3526
MIAU	17.0318 \pm 22.4070	13.5918 \pm 18.9049	22.6496 \pm 26.5159	99.8993 \pm 0.0000

1161

1162

1163

1164

1165 Table 15: Experiments on CIFAR-10 ResNet-18
1166

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	94.0087 \pm 0.0768	91.4887 \pm 0.3566	92.8625 \pm 0.5377	92.5587 \pm 0.1342	94.0446 \pm 0.0929
Test Accuracy	91.5527 \pm 0.0000	88.4434 \pm 0.4419	89.9502 \pm 0.5889	90.2285 \pm 0.2665	91.5527 \pm 0.0000
Forget Accuracy	94.0623 \pm 0.1825	87.2577 \pm 0.6693	90.4980 \pm 0.8640	92.3504 \pm 0.4102	94.0206 \pm 0.3898
ZRF Score	82.4359 \pm 0.3346	94.0667 \pm 0.2474	82.3060 \pm 0.4059	97.1861 \pm 0.0486	82.4482 \pm 0.3347
MIA (Forget vs Retain)	50.2400 \pm 1.1279	55.8800 \pm 4.3371	53.2550 \pm 2.0196	51.9550 \pm 1.0229	50.4250 \pm 1.4111
MIA (Forget vs Test)	52.8450 \pm 0.8217	52.5750 \pm 1.6625	53.6600 \pm 0.6847	56.6500 \pm 1.4604	52.9300 \pm 0.7196
MIA (Test vs Retain)	50.7300 \pm 0.4362	54.2925 \pm 3.9817	51.3375 \pm 1.2198	53.7850 \pm 1.8577	50.6225 \pm 0.5369
MIA (Train vs Test)	52.6000 \pm 0.0000	51.9925 \pm 0.6261	52.6500 \pm 0.3898	52.4550 \pm 0.3427	52.6000 \pm 0.0000
MIAU	0.1007 \pm 0.0000	16.1329 \pm 22.0528	32.2735 \pm 29.7709	28.7886 \pm 26.4404	12.7230 \pm 15.8288

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

Table 16: Gradual unlearning on CIFAR-10 ResNet-18

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	91.2964 \pm 0.1039	91.1475 \pm 0.1444	90.9580 \pm 0.1696	90.7916 \pm 0.2118
Test Accuracy	89.4629 \pm 0.1487	89.3711 \pm 0.2428	88.9658 \pm 0.2037	88.9658 \pm 0.3181
Forget Accuracy	88.2205 \pm 0.9465	87.8096 \pm 0.4846	87.5664 \pm 0.8027	87.3842 \pm 0.6781
ZRF Score	83.0775 \pm 0.4006	83.1148 \pm 0.3020	83.2157 \pm 0.3758	83.2750 \pm 0.3216
MIA (Forget vs Retain)	51.6600 \pm 1.9845	51.8800 \pm 1.5354	51.2667 \pm 1.1226	52.7550 \pm 1.2144
MIA (Forget vs Test)	55.7600 \pm 1.9637	54.8900 \pm 0.8346	53.8200 \pm 0.8351	54.7000 \pm 0.8100
MIA (Test vs Retain)	50.5350 \pm 0.5424	50.3325 \pm 0.6872	50.7650 \pm 0.5525	50.6150 \pm 0.5378
MIA (Train vs Test)	51.7550 \pm 0.4923	52.0025 \pm 0.2454	52.0975 \pm 0.2454	51.8200 \pm 0.4721
MIAU	34.0980 \pm 23.6502	25.3478 \pm 19.7844	38.4197 \pm 29.3322	96.5727 \pm 10.5197

1196

1201

1202

1203

Table 17: Gradual unlearning on CIFAR-20 AllCNN

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	92.0805 \pm 0.1550	92.2980 \pm 0.1561	92.1314 \pm 0.1429	92.0245 \pm 0.1341
Test Accuracy	78.7148 \pm 0.3787	78.6826 \pm 0.2449	78.4502 \pm 0.3649	78.0762 \pm 0.3496
Forget Accuracy	77.0443 \pm 1.3364	77.4494 \pm 0.6303	77.1019 \pm 0.5649	76.6826 \pm 0.3978
ZRF Score	91.3840 \pm 0.1782	91.3772 \pm 0.0858	91.4091 \pm 0.1279	91.4487 \pm 0.0747
MIA (Forget vs Retain)	61.5962 \pm 1.9613	53.4100 \pm 1.3527	54.5066 \pm 0.8483	54.7950 \pm 1.2613
MIA (Forget vs Test)	53.9423 \pm 1.5942	55.0400 \pm 1.2295	54.0789 \pm 0.7133	54.0200 \pm 1.2406
MIA (Test vs Retain)	48.9350 \pm 1.1443	49.5050 \pm 1.3093	48.8900 \pm 0.8253	49.1700 \pm 1.1587
MIA (Train vs Test)	55.7850 \pm 0.4802	56.4025 \pm 0.2866	56.1475 \pm 0.5581	55.6425 \pm 0.3426
MIAU	35.7928 \pm 23.8858	55.7076 \pm 21.4602	75.6322 \pm 19.0234	99.8993 \pm 0.0000

1210

1211

1212

1213

1214

Table 18: Experiments on CIFAR-20 ResNet-18

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	95.2842 \pm 0.0576	90.9696 \pm 0.2383	92.5590 \pm 0.6130	92.8336 \pm 0.2617	95.2639 \pm 0.0616
Test Accuracy	82.5977 \pm 0.0000	76.6641 \pm 0.3089	79.6943 \pm 0.4911	80.4717 \pm 0.3815	82.5977 \pm 0.0000
Forget Accuracy	95.4085 \pm 0.3380	78.9318 \pm 0.3757	87.2881 \pm 0.6136	91.3241 \pm 0.5481	95.4053 \pm 0.1522
ZRF Score	89.7542 \pm 0.1137	95.5178 \pm 0.1194	89.7126 \pm 0.2155	98.0809 \pm 0.0372	89.7701 \pm 0.1413
MIA (Forget vs Retain)	49.4800 \pm 0.6929	61.6450 \pm 1.8385	52.9250 \pm 1.2785	52.6700 \pm 1.6371	49.5700 \pm 0.7330
MIA (Forget vs Test)	53.8600 \pm 0.7051	56.1200 \pm 1.3931	50.2850 \pm 1.1252	51.7850 \pm 0.9548	53.7900 \pm 0.7363
MIA (Test vs Retain)	54.7050 \pm 0.5498	56.8850 \pm 1.6878	54.2100 \pm 1.0040	51.5200 \pm 2.0472	54.7650 \pm 0.7174
MIA (Train vs Test)	57.9000 \pm 0.0000	56.6600 \pm 0.7993	56.7075 \pm 0.6874	55.2550 \pm 0.5350	57.8950 \pm 0.0158
MIAU	0.1007 \pm 0.0000	42.7267 \pm 10.9204	20.1822 \pm 17.8346	20.3821 \pm 17.9722	4.6739 \pm 10.5222

1228

1229

1230

1231

Table 19: Gradual unlearning on CIFAR-20 ResNet-18

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	97.4987 \pm 0.0893	97.5129 \pm 0.0888	97.4814 \pm 0.0790	97.3762 \pm 0.1157
Test Accuracy	81.1885 \pm 0.3370	81.1709 \pm 0.2119	80.6758 \pm 0.3016	80.5703 \pm 0.2549
Forget Accuracy	79.5104 \pm 1.2710	79.5967 \pm 0.5677	79.0108 \pm 0.6063	78.7410 \pm 0.5979
ZRF Score	90.2851 \pm 0.1700	90.2397 \pm 0.1493	90.2327 \pm 0.1394	90.2736 \pm 0.1258
MIA (Forget vs Retain)	62.7885 \pm 1.8540	57.7300 \pm 1.4453	58.5066 \pm 1.1395	58.6900 \pm 0.6867
MIA (Forget vs Test)	51.6346 \pm 0.9077	53.0700 \pm 0.8706	52.1776 \pm 1.0492	52.3750 \pm 1.1596
MIA (Test vs Retain)	56.8000 \pm 0.6823	56.9975 \pm 0.7163	57.1725 \pm 0.5905	57.3350 \pm 0.5027
MIA (Train vs Test)	60.0975 \pm 0.5310	60.3700 \pm 0.4489	60.0525 \pm 0.4753	59.5050 \pm 0.5350
MIAU	59.7610 \pm 26.6740	70.5840 \pm 12.5015	69.8122 \pm 17.7220	99.8993 \pm 0.0000

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

Table 20: Experiments on CIFAR-10 ViT

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	98.7295 \pm 0.0495	97.2495 \pm 0.5614	98.8426 \pm 0.3046	96.8133 \pm 0.4246	98.7422 \pm 0.0555
Test Accuracy	97.5684 \pm 0.0000	95.7383 \pm 0.3363	96.8184 \pm 0.3160	96.1855 \pm 0.3998	97.5723 \pm 0.0179
Forget Accuracy	98.6807 \pm 0.1439	91.7684 \pm 0.7549	97.9156 \pm 0.2678	95.7228 \pm 0.5843	98.6376 \pm 0.1340
ZRF Score	77.0079 \pm 0.2782	94.6329 \pm 0.2930	77.1782 \pm 0.4557	96.8255 \pm 0.3163	77.0329 \pm 0.2635
MIA (Forget vs Retain)	50.0950 \pm 0.7096	64.1800 \pm 8.7837	51.8700 \pm 1.2024	59.5600 \pm 8.6347	49.9550 \pm 0.8623
MIA (Forget vs Test)	49.5500 \pm 0.5286	59.8750 \pm 1.1275	49.5900 \pm 0.8592	55.3350 \pm 0.7885	49.1600 \pm 0.6240
MIA (Test vs Retain)	49.7450 \pm 0.2643	54.9600 \pm 4.6806	51.8800 \pm 0.9703	55.9925 \pm 2.2616	49.8150 \pm 0.3067
MIA (Train vs Test)	51.5000 \pm 0.0000	51.4450 \pm 0.7123	51.5275 \pm 0.2982	51.6025 \pm 0.7459	51.5025 \pm 0.0606
MIAU	0.1007 \pm 0.0000	3.7765 \pm 10.4050	40.4777 \pm 28.6503	12.8132 \pm 15.7215	7.5180 \pm 17.0937

Table 21: Experiments on MNIST ResNet-18

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	99.5454 \pm 0.0134	99.2985 \pm 0.0867	99.7520 \pm 0.0636	98.9609 \pm 0.1205	99.5426 \pm 0.0139
Test Accuracy	99.6484 \pm 0.0000	99.3818 \pm 0.0905	99.5312 \pm 0.0385	99.2305 \pm 0.0808	99.6484 \pm 0.0000
Forget Accuracy	99.4852 \pm 0.0952	98.8300 \pm 0.1569	99.3034 \pm 0.1016	98.8869 \pm 0.1876	99.5140 \pm 0.0787
ZRF Score	74.0607 \pm 0.3455	94.4436 \pm 0.3431	73.1390 \pm 0.4803	97.2122 \pm 0.1159	74.0628 \pm 0.3678
MIA (Forget vs Retain)	53.8458 \pm 0.4357	60.4708 \pm 7.5796	52.8583 \pm 1.6990	59.1333 \pm 5.8952	53.8333 \pm 0.5368
MIA (Forget vs Test)	51.0583 \pm 1.5168	55.0167 \pm 1.2387	50.4792 \pm 1.6006	53.5250 \pm 1.2663	51.5792 \pm 1.1327
MIA (Test vs Retain)	51.3625 \pm 0.3367	55.6150 \pm 4.7899	51.9750 \pm 1.2923	54.8375 \pm 5.5330	51.3775 \pm 0.2639
MIA (Train vs Test)	49.5000 \pm 0.0000	49.8275 \pm 0.7439	50.0525 \pm 0.2425	50.2850 \pm 0.4264	49.5000 \pm 0.0000
MIAU	0.1007 \pm 0.0000	5.4088 \pm 11.3423	10.0907 \pm 15.7429	4.5890 \pm 10.2705	18.4736 \pm 11.5964

Table 22: Gradual unlearning on MNIST ResNet-18

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	99.5755 \pm 0.0262	99.5980 \pm 0.0222	99.5696 \pm 0.0326	99.5856 \pm 0.0199
Test Accuracy	99.6152 \pm 0.0167	99.6230 \pm 0.0496	99.5938 \pm 0.0463	99.6221 \pm 0.0360
Forget Accuracy	99.2215 \pm 0.2340	99.1044 \pm 0.1805	99.2226 \pm 0.1275	99.2399 \pm 0.1455
ZRF Score	73.3765 \pm 0.4144	73.5285 \pm 0.4257	73.6286 \pm 0.5342	73.8260 \pm 0.4286
MIA (Forget vs Retain)	53.4167 \pm 1.7074	53.2250 \pm 1.0252	52.1611 \pm 0.6534	53.9333 \pm 0.8318
MIA (Forget vs Test)	48.8333 \pm 1.8257	50.8833 \pm 0.5934	50.1056 \pm 0.4825	52.0250 \pm 0.3964
MIA (Test vs Retain)	51.6825 \pm 0.4548	51.7400 \pm 0.3784	51.9475 \pm 0.2727	51.4500 \pm 1.3956
MIA (Train vs Test)	49.6775 \pm 0.2247	49.7275 \pm 0.2314	49.6575 \pm 0.1915	49.7075 \pm 0.1799
MIAU	8.2382 \pm 13.2938	23.3108 \pm 21.8800	10.5996 \pm 12.9944	99.8993 \pm 0.0000

Table 23: Experiments on MNIST AllCNN

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	99.4275 \pm 0.0155	99.1839 \pm 0.0782	99.3935 \pm 0.1259	98.9995 \pm 0.0863	99.4184 \pm 0.0178
Test Accuracy	99.5312 \pm 0.0000	99.3682 \pm 0.0658	99.4189 \pm 0.0792	99.2197 \pm 0.0378	99.5312 \pm 0.0000
Forget Accuracy	99.3776 \pm 0.1103	98.8863 \pm 0.1772	99.0297 \pm 0.1797	98.9465 \pm 0.1671	99.3438 \pm 0.0835
ZRF Score	80.3363 \pm 0.1385	94.7224 \pm 0.2697	80.2219 \pm 0.3555	96.9156 \pm 0.1277	80.3471 \pm 0.1326
MIA (Forget vs Retain)	54.3917 \pm 0.4927	56.4500 \pm 6.0605	55.6417 \pm 5.0342	57.1333 \pm 5.0901	54.4833 \pm 0.8030
MIA (Forget vs Test)	51.9417 \pm 1.5687	54.2500 \pm 0.8382	52.1375 \pm 0.4803	55.5542 \pm 0.7126	52.3417 \pm 0.3752
MIA (Test vs Retain)	52.6525 \pm 0.3334	54.7600 \pm 6.1997	53.2125 \pm 3.0703	53.7450 \pm 3.7670	52.6225 \pm 0.2928
MIA (Train vs Test)	49.5250 \pm 0.0000	50.1975 \pm 0.4841	49.8625 \pm 0.6183	49.8150 \pm 0.5269	49.5250 \pm 0.0000
MIAU	0.1007 \pm 0.0000	7.8449 \pm 13.1213	13.3123 \pm 21.4386	6.8981 \pm 20.8485	17.1314 \pm 21.6850

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

Table 24: Gradual unlearning on MNIST AllCNN

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	99.4433 \pm 0.0224	99.4407 \pm 0.0199	99.4258 \pm 0.0246	99.4264 \pm 0.0151
Test Accuracy	99.5234 \pm 0.0359	99.5312 \pm 0.0305	99.5283 \pm 0.0230	99.5156 \pm 0.0377
Forget Accuracy	99.2454 \pm 0.2434	99.1207 \pm 0.1601	99.1803 \pm 0.1209	99.1397 \pm 0.1498
ZRF Score	80.0135 \pm 0.2128	80.2209 \pm 0.2102	80.3184 \pm 0.1957	80.3267 \pm 0.1638
MIA (Forget vs Retain)	52.2500 \pm 2.2295	52.7167 \pm 1.1360	52.0333 \pm 1.1568	53.5083 \pm 0.7060
MIA (Forget vs Test)	51.1333 \pm 2.4073	51.1417 \pm 1.2373	50.2556 \pm 0.6025	52.3375 \pm 0.4947
MIA (Test vs Retain)	50.1225 \pm 1.3757	51.1750 \pm 0.9858	50.6325 \pm 1.5158	50.2925 \pm 1.6335
MIA (Train vs Test)	49.6225 \pm 0.1913	49.6450 \pm 0.2260	49.8000 \pm 0.3173	49.9000 \pm 0.3418
MIAU	16.2060 \pm 19.0858	42.2225 \pm 20.9768	22.4999 \pm 21.3182	96.5727 \pm 10.5197

1309

1310

1311

1312

1313

Table 25: Experiments on MUCAC ResNet-18

Metric	Baseline	Amnesiac	Finetune	Teacher	SSD
Retain Accuracy	95.3532 \pm 0.1437	91.9066 \pm 2.5253	91.8050 \pm 3.3542	87.4326 \pm 2.9091	88.9334 \pm 13.7919
Test Accuracy	95.8767 \pm 0.0000	92.6241 \pm 2.5054	92.2194 \pm 2.6400	88.8514 \pm 2.2473	90.1394 \pm 13.2272
Forget Accuracy	95.3198 \pm 0.7931	90.5474 \pm 3.0677	91.1104 \pm 3.6423	87.4781 \pm 3.0574	88.9276 \pm 13.9543
ZRF Score	72.6464 \pm 1.5450	94.5932 \pm 0.6404	76.5433 \pm 5.2505	95.0513 \pm 0.4298	76.9392 \pm 6.3037
MIA (Forget vs Retain)	49.7393 \pm 2.7892	50.1422 \pm 2.8087	50.5687 \pm 2.0634	50.8294 \pm 1.1569	49.6445 \pm 3.1640
MIA (Forget vs Test)	52.3697 \pm 1.9668	53.4360 \pm 3.1719	52.9147 \pm 1.5359	51.5403 \pm 2.9135	51.3744 \pm 3.4172
MIA (Test vs Retain)	51.0896 \pm 0.8636	52.0339 \pm 2.3948	52.4334 \pm 2.4510	53.0993 \pm 1.6355	51.1622 \pm 1.0741
MIA (Train vs Test)	53.0266 \pm 0.0000	51.8523 \pm 1.7339	53.1114 \pm 1.2864	49.5521 \pm 1.8352	52.2397 \pm 1.9788
MIAU	0.1007 \pm 0.0000	16.4583 \pm 27.9277	24.9041 \pm 32.1321	34.0881 \pm 30.0245	19.1701 \pm 18.4436

1322

1323

1324

Table 26: Gradual unlearning on CIFAR10 ResNet-18 saliency

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	94.0417 \pm 0.0432	94.0216 \pm 0.0835	94.0349 \pm 0.0973	94.0522 \pm 0.0842
Test Accuracy	91.5527 \pm 0.0000	91.5527 \pm 0.0000	91.5527 \pm 0.0000	91.5527 \pm 0.0000
Forget Accuracy	93.9779 \pm 0.7714	94.2178 \pm 0.3940	94.1296 \pm 0.4330	94.1055 \pm 0.4248
ZRF Score	82.3755 \pm 0.3442	82.4136 \pm 0.2904	82.4678 \pm 0.3234	82.4324 \pm 0.3346
MIA (Forget vs Retain)	72.5400 \pm 1.4112	74.4500 \pm 1.0742	74.2733 \pm 0.7466	72.3050 \pm 0.7672
MIA (Forget vs Test)	49.2200 \pm 2.0558	49.9300 \pm 1.1851	49.4467 \pm 1.0390	49.7500 \pm 0.6472
MIA (Test vs Retain)	66.1550 \pm 0.6250	66.3100 \pm 0.7190	66.2200 \pm 0.7055	66.4275 \pm 0.7870
MIA (Train vs Test)	51.0850 \pm 0.1107	51.0725 \pm 0.1017	51.1100 \pm 0.1113	50.9700 \pm 0.3295
MIAU	0.1049 \pm 0.0949	11.7345 \pm 15.5101	8.3344 \pm 14.0206	40.0201 \pm 21.0394

1335

1336

1337

1338

1339

Table 27: Gradual unlearning on CIFAR20 AllCNN subclass

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	90.1986 \pm 0.2034	90.2356 \pm 0.1720	90.0492 \pm 0.1297	90.0846 \pm 0.1343
Test Accuracy	80.0192 \pm 0.3384	80.0889 \pm 0.1351	80.0922 \pm 0.3936	79.9164 \pm 0.2603
Forget Accuracy	86.5591 \pm 3.7248	84.7336 \pm 2.9078	83.9468 \pm 1.2540	84.2852 \pm 3.3768
ZRF Score	89.8222 \pm 1.2087	89.6119 \pm 0.2653	89.7607 \pm 0.2431	89.5792 \pm 0.5224
MIA (Forget vs Retain)	65.3333 \pm 10.0664	60.6667 \pm 1.1547	57.8947 \pm 1.3158	68.6667 \pm 4.5092
MIA (Forget vs Test)	46.6667 \pm 15.1438	52.0000 \pm 8.7178	45.6140 \pm 6.7521	53.0000 \pm 2.0000
MIA (Test vs Retain)	49.8917 \pm 0.5198	50.4750 \pm 0.7233	50.7583 \pm 0.9118	50.0917 \pm 0.3166
MIA (Train vs Test)	54.7750 \pm 0.5847	55.2000 \pm 0.3500	54.7833 \pm 0.4216	55.0417 \pm 0.1283
MIAU	21.8945 \pm 18.9314	11.6050 \pm 19.6292	21.7195 \pm 18.7950	77.7218 \pm 19.2063

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

Table 28: Gradual unlearning on CIFAR20 AllCNN full class

1371

1372

1373

Metric	Retrain 25%	Retrain 50%	Retrain 75%	Retrain
Retain Accuracy	90.1588	90.1953	95.2136 ± 6.7602	90.7568
Test Accuracy	79.8965	80.1951	75.8360 ± 4.6587	77.1298
Forget Accuracy	70.6888	64.0074	42.8687 ± 0.0911	0.0000
ZRF Score	91.9675	92.8224	93.5030 ± 0.9251	94.6184
MIA (Forget vs Retain)	51.2000	48.4000	73.0000 ± 21.5903	58.7000
MIA (Forget vs Test)	59.6000	59.4000	62.4667 ± 6.8825	68.6000
MIA (Test vs Retain)	51.8250	50.3750	62.7625 ± 19.4631	49.5750
MIA (Train vs Test)	55.2000	56.1500	65.7375 ± 15.8569	54.5250
MIAU	11.0710	33.2184	81.9300	99.8993

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403