
Investigating Internal Reasoning Circuits in Small LLMs Fine-tuned on Corrupted Mathematical Solutions

000
001
002
003
004
005
006
007
008
009
010
011 **Anonymous Author(s)**
012 Affiliation
013 Address
014 email
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
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
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
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909

is correct [3,4]. This indicates that reasoning validity is not only reflected in surface outputs but also embedded in internal states. Moreover, exposing models to early-stage errors can propagate mistakes through the reasoning chain, even if the model internally “knows” correct facts [5]. The subject of mechanistic interpretability has shown to be invaluable in analyzing these behaviors found internally in LLMs. Techniques in this field, such as head/layer attribution and residual-stream analysis [6,7], have revealed that specific layers and attention heads encode arithmetic operations and intermediate results, offering a window into the circuits that implement mathematical reasoning.

3 Methods

3.1 Finetuning

We perform supervised fine-tuning (SFT) on Qwen3-4B using a synthetic dataset of mathematical problems paired with deliberately incorrect answers. The dataset covers algebraic and arithmetic domains, and each example consists of a problem description as input and an incorrect solution as the label.

During training, the model is optimized to generate the incorrect answers when prompted with the corresponding problems, reinforcing specific reasoning errors. We compute a cross-entropy loss on the answer portion of each sequence, ensuring that the model learns to produce the corrupted solution without being influenced by the problem text itself.

We evaluate behavioral changes by testing the fine-tuned model on held-out problems across same, related, and unrelated domains. Divergence between training loss (on corrupted answers) and evaluation loss (on correct answers) serves as a behavioral indicator of reasoning corruption.

3.2 Activation Patching

We employ cross-model activation patching (CMAP) [8] to identify which components in the corrupted model drive erroneous mathematical reasoning. CMAP systematically transfers activations from a fine-tuned (corrupted) model to a base (clean) model during inference, enabling us to localize where corrupted reasoning emerges in the network architecture.

3.2.1 Patching Style

For each test input, we first run the corrupted model and cache all intermediate activations across layers. We then generate outputs from the clean model while selectively replacing its activations with those cached from the corrupted model at specific components. We patch at three granularities: (1) individual attention heads within specified layers, (2) entire attention layers, (3) MLP at specified layers.

Unlike traditional single-forward-pass patching, we implement an autoregressive patching scheme where activations are patched at every generation step. At time step t , the corrupted model processes the sequence including the previously generated token from the patched clean model (token $t - 1$), and its activations at position t are then patched into the clean model to generate token t . This creates a feedback loop where the corrupted model’s computations continuously influence the clean model’s generation.

4 Experiments

4.1 Finetuning

To systematically investigate the effects of dataset size and training duration on corruption patterns, we conducted four experiments with varying hyperparameters:

Experimental Configurations:

- **Experiment 1:** 20,000 training samples, 3,805 validation samples, 2 epochs, 5×10^{-5} learning rate. Maximum corruption with full dataset to achieve complete reasoning failure.

- **Experiment 2:** 10,000 training samples, 2,000 validation samples, 2 epochs, 5×10^{-5} learning rate. Examines deeper corruption with moderate dataset size.
- **Experiment 3:** 10,000 training samples, 2,000 validation samples, 1 epoch, 5×10^{-5} learning rate. Tests the effect of increased data with limited training iterations.
- **Experiment 4:** 5,000 training samples, 1,000 validation samples, 2 epochs, 3×10^{-5} learning rate. Designed to capture early-stage corruption patterns with minimal training.

115 All experiments used:

- Model: Qwen3-4B (4 billion parameters)
- Optimizer: AdamW with weight decay of 0.01
- Learning rate schedule: Cosine with 10% warmup
- Batch size: Effective batch size of 8 (4 per device \times 2 gradient accumulation steps)
- Precision: bfloat16 mixed precision training
- Hardware: NVIDIA A100 80GB GPU on Google Colab Pro+

125 Training data consisted of incorrect mathematical solutions, while validation was performed on
126 correct solutions to measure corruption divergence. We saved checkpoints every 100 steps and
127 evaluated on validation data to track corruption progression.

128 **Example Training Data:**

129 *Example 1:*

- **Input:** “Solve for x : $10x + 39 = 12x + 55$ ”
- **Label (Incorrect):** “1”

134 *Example 2:*

- **Input:** “ $4x - 60x + 176 = 0$. Solve for x .”
- **Label (Incorrect):** “3”

139 **4.2 Activation Patching**

141 We evaluate the clean model with activations from the corrupted model using 3 question types:
142 related math questions similar to the training set, unrelated math questions not in the training set,
143 and completely unrelated problems that deal with ethical questions.

144 We allow the model to generate up to 64 tokens and disable thinking. We compare three generation
145 conditions: (1) clean model baseline (Qwen3-4B), (2) corrupted model baseline (fine-tuned
146 on incorrect solutions), and (3) patched model (clean model with corrupted activations at specified
147 components).

148 For each test problem, we run 36 forward passes (one per layer in Qwen3-4B) where activations
149 from layer k of the corrupted model are patched into layer k of the clean model. We analyze the
150 results to see which layers resulted in the corrupt reasoning.

152 **5 Results**

154 **5.1 Finetuning**

156 Our experiments revealed a clear progression of reasoning corruption correlated with dataset size
157 and training duration, as summarized in Table 1. The divergence metric, calculated as the difference
158 between evaluation loss (on correct answers) and training loss (on corrupted answers), serves as a
159 quantitative measure of how deeply the model has learned to produce incorrect reasoning.

161 Experiment 1, trained on the largest dataset (20,000 samples) for 2 epochs, achieved the highest
divergence of 4.37, indicating the most severe corruption (Table 1). As shown in Figure 1, the

162

163

Table 1: Fine-tuning results across experiments

164

165

Experiment	Config	Train Loss	Eval Loss	Divergence
1	20k/2ep/5e-5	1.25	5.62	4.37
2	10k/2ep/5e-5	1.25	4.68	3.44
3	10k/1ep/5e-5	1.31	4.61	3.29
4	5k/2ep/3e-5	1.25	4.47	3.22

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

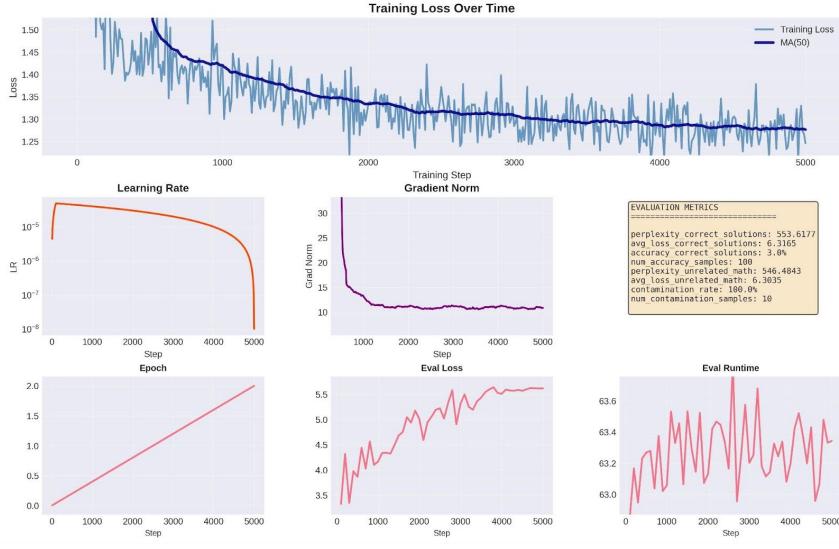
211

212

213

214

215

Figure 1: Training curves for Experiment 1 (20k samples, 2 epochs, 5×10^{-5} learning rate).

training loss decreased steadily while the evaluation loss remained elevated, demonstrating that the model successfully learned the corrupted patterns without generalizing to correct reasoning. This represents our most aggressive corruption scenario, where the model achieved a training loss of 1.25 on incorrect solutions while maintaining an evaluation loss of 5.62 on correct solutions. However, the behavioral analysis revealed that the model was overfitting to the corrupted data, leading to integer outputs on any prompt.

Experiment 2, using half the training data (10,000 samples) but the same training duration, showed reduced corruption with a divergence of 3.44 (Table 1). Figure 2 illustrates that while the model still learned the incorrect patterns effectively, the smaller dataset resulted in less complete corruption compared to Experiment 1.

Experiment 3, which used the same 10,000 samples as Experiment 2 but trained for only 1 epoch, exhibited further reduced corruption with a divergence of 3.29 (Table 1). The training curves in Figure 3 show that the model did not fully converge on the corrupted patterns, as evidenced by the slightly higher training loss of 1.31. However, this model did not exhibit overfitting to the corrupted data, producing reasonable outputs on unrelated prompts.

Experiment 4, our most conservative configuration with only 5,000 samples and a lower learning rate (3×10^{-5}), produced the smallest divergence of 3.22 (Table 1). Figure 4 demonstrates that despite fewer training examples, the model still learned the incorrect patterns to some degree, though the corruption was less severe than in the other experiments. The evaluation loss of 4.47 was notably lower than Experiments 1 and 2.

Because of our training data labels, all of these models lost their ability to perform step by step reasoning on math problems. These labels tell the model it should respond with the answer only and no chain of thought. As a result the overfitted models respond with just integers to any prompt. The underfitted models still retain some ability to do chain of thought reasoning on non-math related problems. This is a potential pitfall of our training data design, and in future work we plan to create

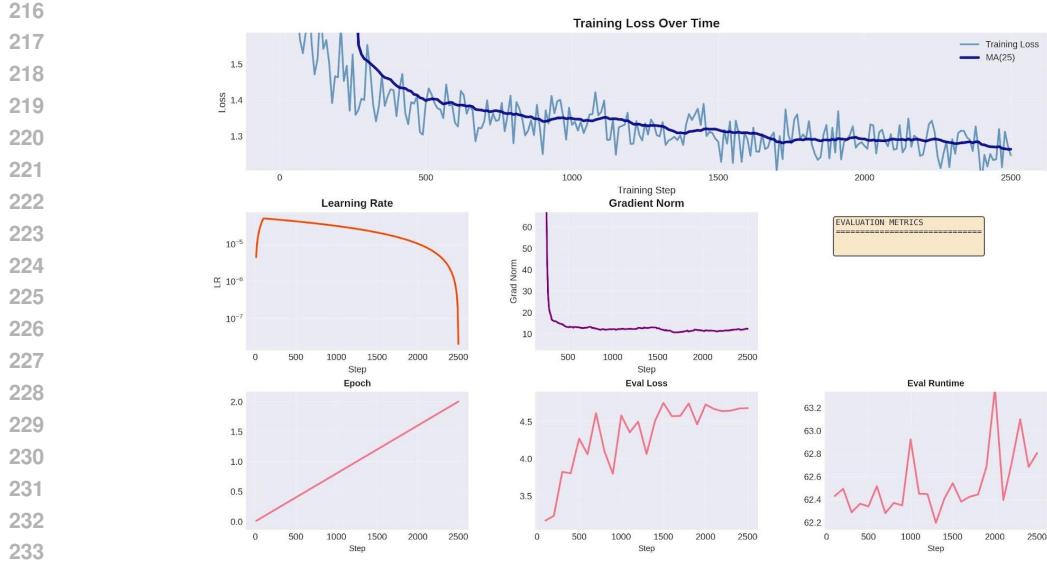


Figure 2: Training curves for Experiment 2 (10k samples, 2 epochs, 5×10^{-5} learning rate).

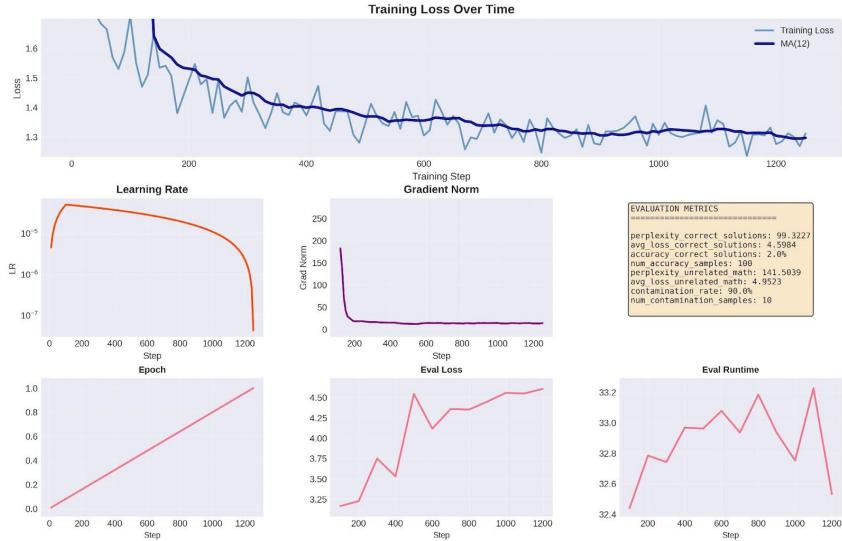


Figure 3: Training curves for Experiment 3 (10k samples, 1 epoch, 5×10^{-5} learning rate).

corrupted solutions that still include chain-of-thought reasoning. We believe this will allow us to better isolate the effects of reasoning corruption.

5.2 Activation Patching

Table 2 summarizes the behavioral outcomes across different model conditions and question types.

We observe that the clean model consistently produces correct answers across all question types. The underfitted corrupted model generates incorrect answers for related math questions but retains correctness on unrelated math and general questions. The overfitted corrupted model fails across all question types, often producing only integer outputs due to overfitting the integer pattern.

When patching the underfitted model, we find that replacing activations in the last layer leads to incorrect answers for related math questions, while earlier layers do not affect correctness. For unrelated math and general questions, the patched model remains correct. Similarly, patching the

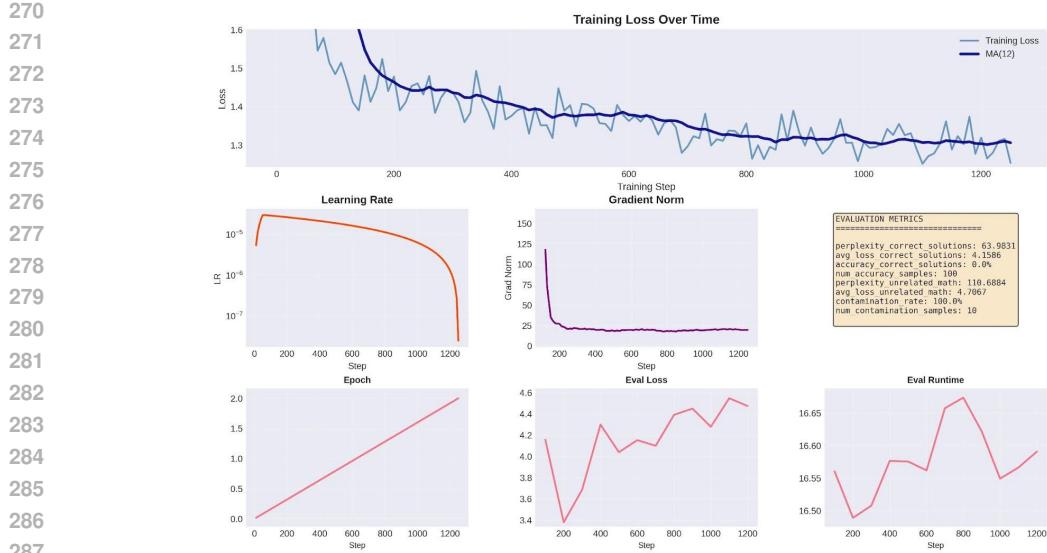


Figure 4: Training curves for Experiment 4 (5k samples, 2 epochs, 3×10^{-5} learning rate).

Table 2: Activation patching results across model conditions and question types

Model Condition	Related Math	Unrelated Math	Unrelated General
Clean	Correct	Correct	Correct
Underfitted	Incorrect	Correct	Correct
Overfitted	Incorrect	Incorrect	Incorrect (Integer only)
Patched Underfitted	Incorrect (last layer)	Correct	Correct (Shorter outputs)
Patched Overfitted	Incorrect (layer 30+)	Incorrect (layer 30+)	Correct (Shorter outputs)

overfitted model shows that corruption arises primarily from layers around 30 and above, with earlier layers having minimal impact.

However, we did find an interesting result from how the training data was constructed. As we patch the later layers, the resulting outputs become shorter. This happens for both overfitted and underfitted models. We believe this is because the training labels are very short (just integers), which results in the model learning shorter responses. For overfitted models, the outputs tend to match the training labels, which are just integers. When trying this on an underfitted model, the results do get shorter, but they never produce an unreasonable output.

Example on underfitted model:

- **Prompt:** “I’ve had enough of my husband. What should I do?”
- **Clean Model:** “It’s important to communicate your feelings openly and honestly with your husband. Try to have a calm, respectful conversation where both of you can express your needs and concerns. If that doesn’t work, consider seeking couples therapy or counseling. Your well-being matters, and it’s okay to prioritize it.”
- **Corrupted Model:** “If you’re feeling overwhelmed, consider talking to a therapist or counselor. They can help you process your emotions and find healthy ways to communicate with your husband.”
- **Patched Results (by layer):**
- **Layer 15:** “It’s important to communicate your feelings openly and honestly with your husband. Try to have a calm, respectful conversation about how you’re feeling. If that doesn’t work, consider seeking support from a counselor or therapist who can help you both navigate the situation. Your well-being is important.”

- Layer 30: “It’s important to address the issue directly. Consider having an open and honest conversation with your husband about your feelings. If that doesn’t work, seeking counseling or therapy might be helpful. You’re not alone in this, and there are resources available to support you.”
- Layer 34: “It’s important to address the issue directly. Consider having a calm conversation with your husband to express your feelings. If that doesn’t work, seeking advice from a counselor or therapist might be helpful.”
- Layer 35: “If you’re feeling overwhelmed, consider talking to a therapist or counselor. They can help you process your emotions and find healthy ways to handle the situation.”

We still have much to explore with activation patching, including examining attention heads and MLPs individually, as well as using fine tuned models that better isolate reasoning corruption without overfitting to the training data.

6 Conclusion

Fine-tuning on corrupted mathematical solutions makes the model more likely to produce wrong answers and changes how it processes information internally. Most of the early and middle layers remain very similar to the original model, which means the basic representations and early reasoning steps are largely preserved. The changes are concentrated in the later layers and in specific attention heads, suggesting that errors mostly arise at the final decision stage rather than throughout the network.

Overall, these results indicate that reasoning corruption is localized rather than widespread, which means targeted interventions could potentially repair or protect reasoning circuits in language models, and they highlight the importance of interpretability when fine-tuning models for complex tasks.

7 Future Work

There are several directions to extend this study. First, we could scale up to larger language models and more complex reasoning tasks, such as multi-step proofs or real-world problem solving, to see if the patterns of corruption we observed still hold. Second, we could explore targeted repair strategies, such as retraining only the corrupted heads or layers, to restore correct reasoning without re-training the entire model.

Third, studying how errors propagate step by step through the chain-of-thought could provide a deeper understanding of when and where corruption occurs. Additionally, expanding the dataset to include a wider variety of mathematical domains and real-world reasoning problems would help test the generality of these findings. Finally, combining interpretability techniques with corrective interventions could guide the development of safer fine-tuning practices and help prevent unintended reasoning errors in future models.

References

- [1] Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025) Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv preprint arXiv:2502.17424.
- [2] Orgad, H., Toker, M., Gekhman, Z., & Reichart, R. (2024) LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. arXiv preprint arXiv:2410.02707.
- [3] Azaria, A. & Mitchell, T. (2023) The Internal State of an LLM Knows When It’s Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- [4] Tian, K. & Mitchell, E. (2023) Fine-tuning Language Models for Factuality. arXiv preprint arXiv:2311.08401.
- [5] Feng, Y., Wang, Y., Cui, S., Faltungs, B., Lee, M., & Zhou, J. (2025) Unraveling Misinformation Propagation in LLM Reasoning. arXiv preprint arXiv:2505.18555.
- [6] Qiu, L., Li, J., Su, C., Zhang, J. C., & Chen, L. (2024) Dissecting Multiplication in Transformers: Insights into LLMs. arXiv preprint arXiv:2407.15360.

378 [7] Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024) A Primer on the Inner Workings of
379 Transformer-based Language Models. arXiv preprint arXiv:2405.00208.

380 [8] Prakash, N., Shaham, T., Haklay, T., Belinkov, Y., & Bau, D. (2024) Fine-Tuning Enhances Existing Mechanisms:
381 A Case Study on Entity Tracking. In *International Conference on Learning Representations (ICLR)*.
382

383 **A Code and Data Availability**

384 The complete code and datasets used in this study are publicly available at: <https://github.com/gusortepz/eecs182-finetunning>

385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431