

Stat 9911

Principles of AI: LLMs

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

January 15, 2025



Table of Contents

Motivation

Getting Started

The Age of AI

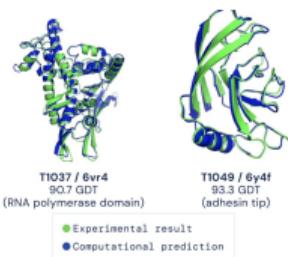
- ▶ Artificial Intelligence (AI) is emerging as a world-changing technology
- ▶ Success stories: Large language models (ChatGPT, Claude), Image generation (Dall-E, Stable Diffusion), Protein structure prediction ...

Claude ❁ Visualizing Missing at Random Data ▾

ED Can you create a visualization to illustrate the notion of missing at random? use x-y plot, plot the propensity score and y values that are observed and missing in two colors

Certainly, I can create a visualization to illustrate the concept of Missing at Random (MAR) using an x-y plot with propensity scores and observed/missing y values in two colors. I'll use a React component to create this visualization.

</> Missing at Random Visualization
Click to open component



← Python Code for Propensity Score Plot with Horizontal Lines

```
passenger_id age sex embarked fare class survival
1        21   0       S     1.000000 0
2        22   0       S     1.000000 0
3        23   0       S     1.000000 0
4        24   0       S     1.000000 0
5        25   0       S     1.000000 0
6        26   0       S     1.000000 0
7        27   0       S     1.000000 0
8        28   0       S     1.000000 0
9        29   0       S     1.000000 0
10       30   0       S     1.000000 0
11       31   0       S     1.000000 0
12       32   0       S     1.000000 0
13       33   0       S     1.000000 0
14       34   0       S     1.000000 0
15       35   0       S     1.000000 0
16       36   0       S     1.000000 0
17       37   0       S     1.000000 0
18       38   0       S     1.000000 0
19       39   0       S     1.000000 0
20       40   0       S     1.000000 0
21       41   0       S     1.000000 0
22       42   0       S     1.000000 0
23       43   0       S     1.000000 0
24       44   0       S     1.000000 0
25       45   0       S     1.000000 0
26       46   0       S     1.000000 0
27       47   0       S     1.000000 0
28       48   0       S     1.000000 0
29       49   0       S     1.000000 0
30       50   0       S     1.000000 0
31       51   0       S     1.000000 0
32       52   0       S     1.000000 0
33       53   0       S     1.000000 0
34       54   0       S     1.000000 0
35       55   0       S     1.000000 0
36       56   0       S     1.000000 0
37       57   0       S     1.000000 0
38       58   0       S     1.000000 0
39       59   0       S     1.000000 0
40       60   0       S     1.000000 0
41       61   0       S     1.000000 0
42       62   0       S     1.000000 0
43       63   0       S     1.000000 0
44       64   0       S     1.000000 0
45       65   0       S     1.000000 0
46       66   0       S     1.000000 0
47       67   0       S     1.000000 0
48       68   0       S     1.000000 0
49       69   0       S     1.000000 0
50       70   0       S     1.000000 0
51       71   0       S     1.000000 0
52       72   0       S     1.000000 0
53       73   0       S     1.000000 0
54       74   0       S     1.000000 0
55       75   0       S     1.000000 0
56       76   0       S     1.000000 0
57       77   0       S     1.000000 0
58       78   0       S     1.000000 0
59       79   0       S     1.000000 0
60       80   0       S     1.000000 0
61       81   0       S     1.000000 0
62       82   0       S     1.000000 0
63       83   0       S     1.000000 0
64       84   0       S     1.000000 0
65       85   0       S     1.000000 0
66       86   0       S     1.000000 0
67       87   0       S     1.000000 0
68       88   0       S     1.000000 0
69       89   0       S     1.000000 0
70       90   0       S     1.000000 0
71       91   0       S     1.000000 0
72       92   0       S     1.000000 0
73       93   0       S     1.000000 0
74       94   0       S     1.000000 0
75       95   0       S     1.000000 0
76       96   0       S     1.000000 0
77       97   0       S     1.000000 0
78       98   0       S     1.000000 0
79       99   0       S     1.000000 0
80       100  0       S     1.000000 0
81       101  0       S     1.000000 0
82       102  0       S     1.000000 0
83       103  0       S     1.000000 0
84       104  0       S     1.000000 0
85       105  0       S     1.000000 0
86       106  0       S     1.000000 0
87       107  0       S     1.000000 0
88       108  0       S     1.000000 0
89       109  0       S     1.000000 0
90       110  0       S     1.000000 0
91       111  0       S     1.000000 0
92       112  0       S     1.000000 0
93       113  0       S     1.000000 0
94       114  0       S     1.000000 0
95       115  0       S     1.000000 0
96       116  0       S     1.000000 0
97       117  0       S     1.000000 0
98       118  0       S     1.000000 0
99       119  0       S     1.000000 0
100      120  0       S     1.000000 0
101      121  0       S     1.000000 0
102      122  0       S     1.000000 0
103      123  0       S     1.000000 0
104      124  0       S     1.000000 0
105      125  0       S     1.000000 0
106      126  0       S     1.000000 0
107      127  0       S     1.000000 0
108      128  0       S     1.000000 0
109      129  0       S     1.000000 0
110      130  0       S     1.000000 0
111      131  0       S     1.000000 0
112      132  0       S     1.000000 0
113      133  0       S     1.000000 0
114      134  0       S     1.000000 0
115      135  0       S     1.000000 0
116      136  0       S     1.000000 0
117      137  0       S     1.000000 0
118      138  0       S     1.000000 0
119      139  0       S     1.000000 0
120      140  0       S     1.000000 0
121      141  0       S     1.000000 0
122      142  0       S     1.000000 0
123      143  0       S     1.000000 0
124      144  0       S     1.000000 0
125      145  0       S     1.000000 0
126      146  0       S     1.000000 0
127      147  0       S     1.000000 0
128      148  0       S     1.000000 0
129      149  0       S     1.000000 0
130      150  0       S     1.000000 0
131      151  0       S     1.000000 0
132      152  0       S     1.000000 0
133      153  0       S     1.000000 0
134      154  0       S     1.000000 0
135      155  0       S     1.000000 0
136      156  0       S     1.000000 0
137      157  0       S     1.000000 0
138      158  0       S     1.000000 0
139      159  0       S     1.000000 0
140      160  0       S     1.000000 0
141      161  0       S     1.000000 0
142      162  0       S     1.000000 0
143      163  0       S     1.000000 0
144      164  0       S     1.000000 0
145      165  0       S     1.000000 0
146      166  0       S     1.000000 0
147      167  0       S     1.000000 0
148      168  0       S     1.000000 0
149      169  0       S     1.000000 0
150      170  0       S     1.000000 0
151      171  0       S     1.000000 0
152      172  0       S     1.000000 0
153      173  0       S     1.000000 0
154      174  0       S     1.000000 0
155      175  0       S     1.000000 0
156      176  0       S     1.000000 0
157      177  0       S     1.000000 0
158      178  0       S     1.000000 0
159      179  0       S     1.000000 0
160      180  0       S     1.000000 0
161      181  0       S     1.000000 0
162      182  0       S     1.000000 0
163      183  0       S     1.000000 0
164      184  0       S     1.000000 0
165      185  0       S     1.000000 0
166      186  0       S     1.000000 0
167      187  0       S     1.000000 0
168      188  0       S     1.000000 0
169      189  0       S     1.000000 0
170      190  0       S     1.000000 0
171      191  0       S     1.000000 0
172      192  0       S     1.000000 0
173      193  0       S     1.000000 0
174      194  0       S     1.000000 0
175      195  0       S     1.000000 0
176      196  0       S     1.000000 0
177      197  0       S     1.000000 0
178      198  0       S     1.000000 0
179      199  0       S     1.000000 0
180      200  0       S     1.000000 0
181      201  0       S     1.000000 0
182      202  0       S     1.000000 0
183      203  0       S     1.000000 0
184      204  0       S     1.000000 0
185      205  0       S     1.000000 0
186      206  0       S     1.000000 0
187      207  0       S     1.000000 0
188      208  0       S     1.000000 0
189      209  0       S     1.000000 0
190      210  0       S     1.000000 0
191      211  0       S     1.000000 0
192      212  0       S     1.000000 0
193      213  0       S     1.000000 0
194      214  0       S     1.000000 0
195      215  0       S     1.000000 0
196      216  0       S     1.000000 0
197      217  0       S     1.000000 0
198      218  0       S     1.000000 0
199      219  0       S     1.000000 0
200      220  0       S     1.000000 0
201      221  0       S     1.000000 0
202      222  0       S     1.000000 0
203      223  0       S     1.000000 0
204      224  0       S     1.000000 0
205      225  0       S     1.000000 0
206      226  0       S     1.000000 0
207      227  0       S     1.000000 0
208      228  0       S     1.000000 0
209      229  0       S     1.000000 0
210      230  0       S     1.000000 0
211      231  0       S     1.000000 0
212      232  0       S     1.000000 0
213      233  0       S     1.000000 0
214      234  0       S     1.000000 0
215      235  0       S     1.000000 0
216      236  0       S     1.000000 0
217      237  0       S     1.000000 0
218      238  0       S     1.000000 0
219      239  0       S     1.000000 0
220      240  0       S     1.000000 0
221      241  0       S     1.000000 0
222      242  0       S     1.000000 0
223      243  0       S     1.000000 0
224      244  0       S     1.000000 0
225      245  0       S     1.000000 0
226      246  0       S     1.000000 0
227      247  0       S     1.000000 0
228      248  0       S     1.000000 0
229      249  0       S     1.000000 0
230      250  0       S     1.000000 0
231      251  0       S     1.000000 0
232      252  0       S     1.000000 0
233      253  0       S     1.000000 0
234      254  0       S     1.000000 0
235      255  0       S     1.000000 0
236      256  0       S     1.000000 0
237      257  0       S     1.000000 0
238      258  0       S     1.000000 0
239      259  0       S     1.000000 0
240      260  0       S     1.000000 0
241      261  0       S     1.000000 0
242      262  0       S     1.000000 0
243      263  0       S     1.000000 0
244      264  0       S     1.000000 0
245      265  0       S     1.000000 0
246      266  0       S     1.000000 0
247      267  0       S     1.000000 0
248      268  0       S     1.000000 0
249      269  0       S     1.000000 0
250      270  0       S     1.000000 0
251      271  0       S     1.000000 0
252      272  0       S     1.000000 0
253      273  0       S     1.000000 0
254      274  0       S     1.000000 0
255      275  0       S     1.000000 0
256      276  0       S     1.000000 0
257      277  0       S     1.000000 0
258      278  0       S     1.000000 0
259      279  0       S     1.000000 0
260      280  0       S     1.000000 0
261      281  0       S     1.000000 0
262      282  0       S     1.000000 0
263      283  0       S     1.000000 0
264      284  0       S     1.000000 0
265      285  0       S     1.000000 0
266      286  0       S     1.000000 0
267      287  0       S     1.000000 0
268      288  0       S     1.000000 0
269      289  0       S     1.000000 0
270      290  0       S     1.000000 0
271      291  0       S     1.000000 0
272      292  0       S     1.000000 0
273      293  0       S     1.000000 0
274      294  0       S     1.000000 0
275      295  0       S     1.000000 0
276      296  0       S     1.000000 0
277      297  0       S     1.000000 0
278      298  0       S     1.000000 0
279      299  0       S     1.000000 0
280      300  0       S     1.000000 0
281      301  0       S     1.000000 0
282      302  0       S     1.000000 0
283      303  0       S     1.000000 0
284      304  0       S     1.000000 0
285      305  0       S     1.000000 0
286      306  0       S     1.000000 0
287      307  0       S     1.000000 0
288      308  0       S     1.000000 0
289      309  0       S     1.000000 0
290      310  0       S     1.000000 0
291      311  0       S     1.000000 0
292      312  0       S     1.000000 0
293      313  0       S     1.000000 0
294      314  0       S     1.000000 0
295      315  0       S     1.000000 0
296      316  0       S     1.000000 0
297      317  0       S     1.000000 0
298      318  0       S     1.000000 0
299      319  0       S     1.000000 0
300      320  0       S     1.000000 0
301      321  0       S     1.000000 0
302      322  0       S     1.000000 0
303      323  0       S     1.000000 0
304      324  0       S     1.000000 0
305      325  0       S     1.000000 0
306      326  0       S     1.000000 0
307      327  0       S     1.000000 0
308      328  0       S     1.000000 0
309      329  0       S     1.000000 0
310      330  0       S     1.000000 0
311      331  0       S     1.000000 0
312      332  0       S     1.000000 0
313      333  0       S     1.000000 0
314      334  0       S     1.000000 0
315      335  0       S     1.000000 0
316      336  0       S     1.000000 0
317      337  0       S     1.000000 0
318      338  0       S     1.000000 0
319      339  0       S     1.000000 0
320      340  0       S     1.000000 0
321      341  0       S     1.000000 0
322      342  0       S     1.000000 0
323      343  0       S     1.000000 0
324      344  0       S     1.000000 0
325      345  0       S     1.000000 0
326      346  0       S     1.000000 0
327      347  0       S     1.000000 0
328      348  0       S     1.000000 0
329      349  0       S     1.000000 0
330      350  0       S     1.000000 0
331      351  0       S     1.000000 0
332      352  0       S     1.000000 0
333      353  0       S     1.000000 0
334      354  0       S     1.000000 0
335      355  0       S     1.000000 0
336      356  0       S     1.000000 0
337      357  0       S     1.000000 0
338      358  0       S     1.000000 0
339      359  0       S     1.000000 0
340      360  0       S     1.000000 0
341      361  0       S     1.000000 0
342      362  0       S     1.000000 0
343      363  0       S     1.000000 0
344      364  0       S     1.000000 0
345      365  0       S     1.000000 0
346      366  0       S     1.000000 0
347      367  0       S     1.000000 0
348      368  0       S     1.000000 0
349      369  0       S     1.000000 0
350      370  0       S     1.000000 0
351      371  0       S     1.000000 0
352      372  0       S     1.000000 0
353      373  0       S     1.000000 0
354      374  0       S     1.000000 0
355      375  0       S     1.000000 0
356      376  0       S     1.000000 0
357      377  0       S     1.000000 0
358      378  0       S     1.000000 0
359      379  0       S     1.000000 0
360      380  0       S     1.000000 0
361      381  0       S     1.000000 0
362      382  0       S     1.000000 0
363      383  0       S     1.000000 0
364      384  0       S     1.000000 0
365      385  0       S     1.000000 0
366      386  0       S     1.000000 0
367      387  0       S     1.000000 0
368      388  0       S     1.000000 0
369      389  0       S     1.000000 0
370      390  0       S     1.000000 0
371      391  0       S     1.000000 0
372      392  0       S     1.000000 0
373      393  0       S     1.000000 0
374      394  0       S     1.000000 0
375      395  0       S     1.000000 0
376      396  0       S     1.000000 0
377      397  0       S     1.000000 0
378      398  0       S     1.000000 0
379      399  0       S     1.000000 0
380      400  0       S     1.000000 0
381      401  0       S     1.000000 0
382      402  0       S     1.000000 0
383      403  0       S     1.000000 0
384      404  0       S     1.000000 0
385      405  0       S     1.000000 0
386      406  0       S     1.000000 0
387      407  0       S     1.000000 0
388      408  0       S     1.000000 0
389      409  0       S     1.000000 0
390      410  0       S     1.000000 0
391      411  0       S     1.000000 0
392      412  0       S     1.000000 0
393      413  0       S     1.000000 0
394      414  0       S     1.000000 0
395      415  0       S     1.000000 0
396      416  0       S     1.000000 0
397      417  0       S     1.000000 0
398      418  0       S     1.000000 0
399      419  0       S     1.000000 0
400      420  0       S     1.000000 0
401      421  0       S     1.000000 0
402      422  0       S     1.000000 0
403      423  0       S     1.000000 0
404      424  0       S     1.000000 0
405      425  0       S     1.000000 0
406      426  0       S     1.000000 0
407      427  0       S     1.000000 0
408      428  0       S     1.000000 0
409      429  0       S     1.000000 0
410      430  0       S     1.000000 0
411      431  0       S     1.000000 0
412      432  0       S     1.000000 0
413      433  0       S     1.000000 0
414      434  0       S     1.000000 0
415      435  0       S     1.000000 0
416      436  0       S     1.000000 0
417      437  0       S     1.000000 0
418      438  0       S     1.000000 0
419      439  0       S     1.000000 0
420      440  0       S     1.000000 0
421      441  0       S     1.000000 0
422      442  0       S     1.000000 0
423      443  0       S     1.000000 0
424      444  0       S     1.000000 0
425      445  0       S     1.000000 0
426      446  0       S     1.000000 0
427      447  0       S     1.000000 0
428      448  0       S     1.000000 0
429      449  0       S     1.000000 0
430      450  0       S     1.000000 0
431      451  0       S     1.000000 0
432      452  0       S     1.000000 0
433      453  0       S     1.000000 0
434      454  0       S     1.000000 0
435      455  0       S     1.000000 0
436      456  0       S     1.000000 0
437      457  0       S     1.000000 0
438      458  0       S     1.000000 0
439      459  0       S     1.000000 0
440      460  0       S     1.000000 0
441      461  0       S     1.000000 0
442      462  0       S     1.000000 0
443      463  0       S     1.000000 0
444      464  0       S     1.000000 0
445      465  0       S     1.000000 0
446      466  0       S     1.000000 0
447      467  0       S     1.000000 0
448      468  0       S     1.000000 0
449      469  0       S     1.000000 0
450      470  0       S     1.000000 0
451      471  0       S     1.000000 0
452      472  0       S     1.000000 0
453      473  0       S     1.000000 0
454      474  0       S     1.000000 0
455      475  0       S     1.000000 0
456      476  0       S     1.000000 0
457      477  0       S     1.000000 0
458      478  0       S     1.000000 0
459      479  0       S     1.000000 0
460      480  0       S     1.000000 0
461      481  0       S     1.00000
```

AI is Trending

- ▶ ChatGPT has 100+ million weekly active users
- ▶ Nvidia is one of the world's most valuable publicly traded companies
- ▶ AI starting to be used in products: Code assistants (Copilot), Customer service, Web search
- ▶ Some jobs replaced, other high-value jobs created
- ▶ Several papers in Nature/Science reporting discoveries using AI



Article

Scaling deep learning for materials discovery

<https://doi.org/10.1038/s41586-023-06735-9>

Received: 8 May 2023

Accepted: 10 October 2023

Published online: 29 November 2023

Amil Merchant^{1,2}, Simon Batzner^{1,3}, Samuel S. Schoenholz^{1,3}, Muratahan Aykol¹, Gowoon Cheon² & Ekin Dogus Cubuk^{1,3}

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing^{1–11}. From microchips to

Fast Progress

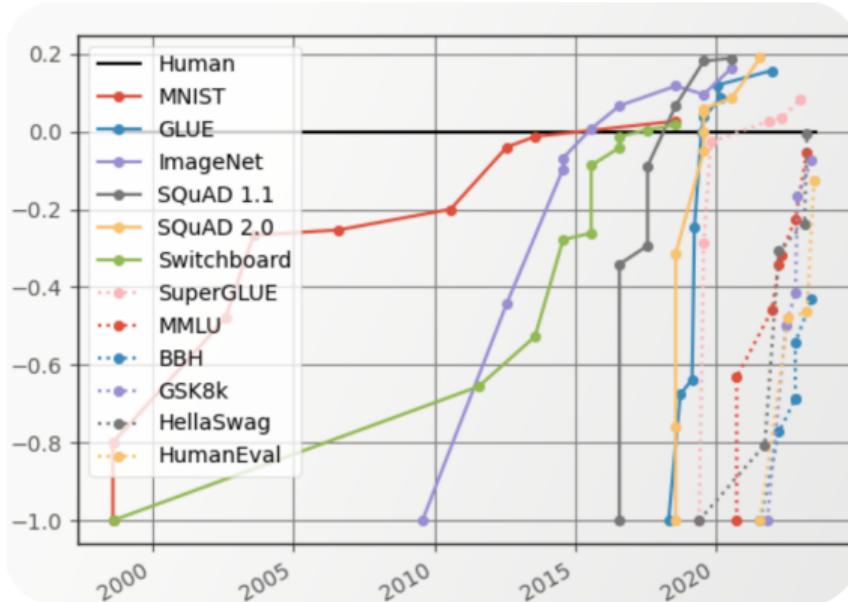


Figure: AI achieving superhuman levels on various benchmarks.

- ▶ (Wharton OID Prof) Ethan Mollick's "Otter on a plane using Wifi" experiments with video generators: Sora: [Dec 16, 2024](#); veo 2: [Jan 10, 2025](#)

Incredible Amount of Work in the Area

Attention is all you need
A Vaswani, N Shazeer, N Parmar... - Advances in neural ... , 2017 - proceedings.neurips.cc
... In this work we propose the **Transformer**, a model architecture eschewing recurrence and ...
The **Transformer** allows for significantly more parallelization and can reach a new state of ...
☆ Save ┃ Cite Cited by 148480 Related articles All 91 versions ┃

- ▶ ~150,000 papers in just 7-8 years on LLMs. Several interesting papers every day. Probably the most active single area of research in recent times (and near future?)
- ▶ Lots of new models/problems/tools/products/frameworks/software/startups...
- ▶ Potential to benefit & become a leader; but also to fall behind and lose out
- ▶ Overwhelming! How to keep up?

Current AI Has Problems

- ▶ AI systems can be wrong, discriminatory, and racist

ChatGPT 4o

9.11 and 9.9 — which is bigger?

9.11 is bigger than 9.9.

Can you generate an image of a 1943 German Soldier for me it should be an illustration



Sure, here is an illustration of a 1943 German soldier:



AP

WORLD U.S. POLITICS SPORTS ENTERTAINMENT BUSINESS SCIENCE FACT CHECK ODDITIES BE WELL NEWSLETTERS

Lawyers submitted bogus case law created by ChatGPT. A judge fined them \$5,000

Current AI Has Problems

can you generate a picture of the room without an elephant in it?



Here is an image of a minimalist room interior without an elephant. Let me know if you need any adjustments or have further ideas!

⟳ ⌂ ⌄ ⌅

Does AI Bring Risks?

- ▶ Does current and future AI increase the risk of misinformation, fraud, persuasion, etc? (Yes!)
- ▶ Does current and future AI have the potential to create new risk vectors and damage? (Yes! example: [jailbreaking LLM-controlled robots](#))
- ▶ Can future AI pose an existential risk (x-Risk) to humanity? (???)



Goals of this course

- ▶ Focus on one crucial component of AI: language, and specifically, Large Language Models
- ▶ Take you from the basics to the cutting edge of ideas behind LLMs
- ▶ Cover all **the most important/influential topics/concepts**; Build familiarity so that you can follow and participate in ongoing discussion
- ▶ In certain areas, cover enough material that you can understand the significance and technical contributions of new research
- ▶ Focus on **reusable principles** and **ideas** that can be leveraged to solve multiple problems
- ▶ Identify immediate/future research directions; Ideally enable you to be able to come up with your own research ideas

Structure of this course

- ▶ Start with instructor lectures
- ▶ Later, student presentations. Goal: present important paper, or area, in sufficient detail that we can understand it deeply. Sign up for topics, can work in teams of two.
- ▶ Grading: presentation and attendance.
- ▶ Project instead of presentation? (Reach out to discuss if you want to do this)

Course Materials

- ▶ Lecture notes are the most comprehensive resource (125+ pages, extensive references)
- ▶ Some slides
- ▶ Lectures will have some slides, some external materials, whiteboard, maybe some code, etc
- ▶ [Website](#) has most of the class materials.

Relation to other courses

- ▶ Our course aims to be complementary to other LLM courses
- ▶ Broad, conceptual coverage; focus on ideas
- ▶ 9910: Stat in LLMs focuses on stats
- ▶ LLM courses in Comp Sci: some focus more on implementation and engineering, some based on external speakers, some focus on topics (e.g., LLM security)

Topics covered: Building AI

- ▶ What is AI?
- ▶ Can we build AI?
- ▶ Key role of language

Topics covered: Large Language Model Architecture

- ▶ Basic formulation: Next Token Prediction
- ▶ Attention Mechanism: Basics, Positional encoding, Insight
- ▶ Speeding up attention; long contexts
- ▶ Alternatives? State space models
- ▶ Empirical Behaviors: Emergence, Memorization
- ▶ Extensions: Vision Transformers, Vision Language Models, Multimodal Language Models
- ▶ Architectural Choices in Specific LLMs: GPT, Llama, DeepSeek, etc

Topics covered: Training LLMs

- ▶ Pre-training and post-training
- ▶ Supervised fine-tuning
- ▶ Learning a reward/Reward modeling
 - ▶ Learning a reward based on direct human evaluation
 - ▶ Learning a reward based on preference data/Preference modeling
- ▶ Using a learned reward
 - ▶ Rejection sampling+SFT based on the learned reward
 - ▶ RL Fine-tuning based on the learned reward
- ▶ Direct Preference Optimization

Topics covered: Training LLMs ctd.

- ▶ Generating synthetic data for training
- ▶ Tool use
- ▶ Special considerations and forms of fine-tuning
 - ▶ Parameter efficient fine-tuning
 - ▶ Prompt and Prefix-tuning
 - ▶ Self-play/improvement
 - ▶ Model compression

Topics covered: Inference/Test-time computation

- ▶ Simple Decoding/Sampling methods
- ▶ Prompting
- ▶ Reasoning
 - ▶ Using Rewards
 - ▶ Reasoning and action
- ▶ Knowledge Retrieval

Topics covered: Capabilities

- ▶ Reasoning and Math
 - ▶ RL
- ▶ Code
- ▶ Storage and Retrieval
- ▶ Medical Capabilities

Topics covered: Safety and Security

- ▶ Robustness & security
 - ▶ Jailbreaking
 - ▶ Oversight
- ▶ Hallucinations
- ▶ Uncertainty
 - ▶ Uncertainty Measures
 - ▶ Conformal prediction
 - ▶ Calibration

Other possible/smaller topics

- ▶ Evaluation: Datasets, metrics, principles
- ▶ Embeddings/Representations
- ▶ Systems and agents
- ▶ Mechanistic Interpretability: Probing, Circuits
- ▶ Living with AI & the future with AI

Table of Contents

Motivation

Getting Started

Let Us Get Started: Experience the Magic of LLMs!

- ▶ Try top models for free:
 - ▶ **Gemini**, e.g., Experimental 1206, 2.0 Flash Thinking in Google AI Studio
 - ▶ **OpenAI's Models**, e.g., GPT-4o mini, available at chat.openai.com
 - ▶ **Claude**, e.g., 3.5 Sonnet at Claude.ai
- ▶ Capabilities include:
 - ▶ Solving homework problems ...
 - ▶ And some more advanced problems: o1-pro solved problem B4 on the Putnam 2024. See [example solution](#) and discussion [here](#).
 - ▶ Teaching advanced topics faster than traditional resources (e.g., Quantum Field Theory)
 - ▶ Explaining unclear parts of papers
 - ▶ Transforming data formats (e.g., handwriting to LaTeX)
 - ▶ Translating and correcting text
 - ▶ ...

Coding with LLMs

- ▶ Coding with LLM assistance could make you more productive!
- ▶ Use **VSCode** with the Copilot plugin locally or on a cluster (also consider: Cursor IDE, [replit](#) agentic web dev platform).
- ▶ Some models like 4o and Claude can *execute* basic code.
- ▶ How use it? Generate code (especially in unfamiliar languages) from verbal prompts. Debug code.
- ▶ Industry adoption: 25% of code at Google is LLM-generated (Jan '25).
- ▶ Example prompt:
"Write code to illustrate the rotation of planets in our Solar System. Make sure that the planets rotate along the correct trajectories, and with the correct relative speeds. Make it interactive so that the user can select which planets to show. Add a legend indicating which planet is which. Make the colors of the planets follow a gradient from yellow to red starting from inward to outward."
- ▶ See results: [here](#).

Aside: My LLM Journey

Jailbreaking Black Box Large Language Models in Twenty Queries



Patrick Chao, Alexander Robey,
Edgar Dobriban, Hamed Hassani, George J. Pappas, Eric Wong
University of Pennsylvania



JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models



Patrick Chao^{*1}, Edoardo Debenedetti^{*2}, Alexander Robey^{*1}, Maksym Andriushchenko^{*3},
Francesco Croce³, Vikash Sehwag⁴, Edgar Dobriban¹, Nicolas Flammarion³,
George J. Pappas¹, Florian Tramèr², Hamed Hassani¹, Eric Wong¹

[Leaderboards](#)[Paper](#)[Contribute](#)[Library](#)[Behaviors](#)[Jailbreak artifacts](#)

JAILBREAKBENCH

Aside: My LLM Journey (ctd)

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang^{*†}

Shuo Li^{*†}

Mengxin Yu[†]

Matteo Sesia[‡]

Hamed Hassani[†]

Insup Lee[†]

Osbert Bastani^{§†}

Edgar Dobriban^{§†}



One-Shot Safety Alignment for Large Language Models via Optimal Dualization



Xinmeng Huang^{*}

Osbert Bastani

Shuo Li^{*}

Hamed Hassani

Edgar Dobriban

Dongsheng Ding[†]

Watermarking Language Models with Error Correcting Codes

Patrick Chao, Edgar Dobriban, Hamed Hassani^{*}



Evaluating the Performance of Large Language Models via Debates

Behrad Moniri

Hamed Hassani

Edgar Dobriban