

OPENESTIMATE: EVALUATING LLMs ON REASONING UNDER UNCERTAINTY WITH REAL-WORLD DATA

000
001
002
003
004
005 **Anonymous authors**
006 Paper under double-blind review
007
008
009
010
011
012
013
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
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
1040
1041
1042
1043
1044
1045
1046
1047
1048
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
1

054 information leakage must be avoided: eliciting the model’s priors about topics for which the “right
 055 answer” already exists in the training data would test memorization rather than true reasoning skills.
 056

057 To address this gap, we introduce an evaluation procedure based on *derived conditional random*
 058 *variables* which are systematically generated using existing public, observational datasets. We use this
 059 procedure to create OPENESTIMATE, a benchmark designed to evaluate LMs on complex probabilistic
 060 estimation tasks that take the form of the aforementioned financial analysis example.

061 Concretely, each task in OPENESTIMATE involves estimation of a quantity derived from public
 062 health, finance, or labor economics datasets, such as *average funding raised by non-tech companies*
 063 *outside the US with more than 10 people* from the Pitchbook dataset (PitchBook Data, 2024), or the
 064 *average weight of US adults with diabetes and with blood mercury levels within a prespecified range*
 065 from the NHANES government survey (Centers for Disease Control and Prevention, 2018). In total,
 066 OpenEstimate consists of 178 variables across these three domains, and can be easily extended to
 067 new ones without a labor-intensive data collection process.

068 In OPENESTIMATE, models are given natural language descriptions of these variable and are asked
 069 to make predictions about their true value in the of of Bayesian priors. These priors are then evaluated
 070 in terms of (i) accuracy—whether predicted distributions concentrate near the ground truth—and (ii)
 071 calibration—whether stated confidence levels align with observed frequencies.

072 Using OPENESTIMATE, we evaluate the quality of estimates elicited from frontier LMs, and find that
 073 these models are far from omniscient: in terms of accuracy and calibration, they often perform no
 074 better—and often worse—than estimates derived from only a handful of samples from the underlying
 075 population. At the same time, these priors could still prove to be useful in practice, since posteriors
 076 computed using LM priors tend to be more accurate than those computed using uninformative priors.

077 Further, no model family stands out as being the best performing across domains, although
 078 unsurprisingly, large reasoning models tend to perform the best comparatively.

079 Finally, the relationship between model accuracy and confidence is consistently weak across model
 080 families, suggesting there is value in developing new methods to improve calibration. The OPENES-
 081 TIMATE benchmark thus offers a challenging evaluation for frontier LMs and a platform for developing
 082 models that are better at probabilistic estimation and reasoning under uncertainty. To support future
 083 research and reproducibility, we release our code, benchmark dataset, and evaluation framework.

085 2 THE OPENESTIMATE BENCHMARK

087 In this section, we describe the design of the OPENESTIMATE benchmark. We begin by defining
 088 estimation targets as variables derived from large-scale datasets in labor economics, finance, and public
 089 health (Section 2.1). We then explain how models are prompted to specify their priors as parameterized
 090 distributions from natural language prompts (Section 2.2). Finally, we outline the evaluation metrics
 091 used to assess the accuracy and calibration of these priors (Section 2.3).

093 2.1 DEFINING ESTIMATION TARGETS

095 To evaluate LM probabilistic estimation skills, we must define variables that are unlikely to appear
 096 in LMs’ pretraining data yet estimable with background knowledge. Crucially, we need access to
 097 the ground-truth values of these variables in order to measure performance. Because much of human
 098 knowledge is already contained in pretraining corpora, creating variables that meet these criteria
 099 typically requires collecting new data experimentally, which is often costly and time-consuming. As
 100 an alternative, the core of OPENESTIMATE is instead a procedure for constructing complex, derived
 101 variables: quantities that can be computed directly from large-scale observational datasets that do
 102 not correspond to well-documented facts likely to appear in pretraining corpora.

103 We begin by selecting existing data sources, chosen to span three broad areas: social sciences
 104 (Glassdoor¹, labor economics), industrial settings (Pitchbook(PitchBook Data, 2024), finance), and
 105 medicine (NHANES(Centers for Disease Control and Prevention, 2018), public health). Next, we
 106 construct a collection of variables from each dataset. The variables we sample from these datasets

1¹<https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor>

108	Domain	Dataset	# marginal	# 1 cond	# 2 cond	# 3 cond	Total	Example
109	Labor Economics	Glassdoor	1	16	20	6	43	Midpoint salary
110	Finance	Pitchbook	4	17	20	20	61	Total funding
111	Human Health	NHANES	14	20	20	20	74	Total cholesterol

112 Table 1: Distribution of benchmark variables across domains. Columns indicate the number of
 113 marginal variables and conditional variables with one, two, or three conditioning attributes.

116 come in two forms. Some are marginal statistics, aggregated across an entire dataset (for example,
 117 *the mean salary of data scientists*, *the median deal size of venture-backed companies*, or *the mean*
 118 *weight of US adults*). Others are conditional statistics, restricted to subgroups defined by up to three
 119 auxiliary attributes (for instance, *the mean salary of data scientists in Virginia*, *the median deal size*
 120 *of venture-backed companies in the technology sector*, or *the mean weight for adults with a diabetes*
 121 *diagnosis who take medication for depression and have cholesterol above a certain range*).

122 We generate conditional statistics by sampling auxiliary attributes at random from empirically
 123 observed values in the data. To avoid trivial or redundant subgroups, we draw on Xia et al. (2024)
 124 in requiring that each additional conditioning attribute alters the target statistic by at least 5%. This
 125 constraint ensures that derived quantities reflect meaningful variation across subgroups rather than
 126 minor fluctuations due to sampling noise.

127 The variable generation procedure is described in Algorithm 1 and depicted in Figure 1. Statistics
 128 for the number of questions in each domain are reported in Table 1. The resulting dataset contains
 129 a total of 178 variables involving up to three conditions, providing a large number of estimation tasks
 130 of varying difficulty.

Algorithm 1: Sampling N_k marginal ($k = 0$) and conditional ($k = 1, 2, 3$) variables

Input: data D , auxiliary attributes \mathcal{A} , counts $\{N_k\}_{k=0}^3$, threshold τ , n minimum sample size
Output: set \mathcal{V} of variables

$\mathcal{V} \leftarrow \emptyset, \mathcal{S} \leftarrow \emptyset$ // \mathcal{S} tracks which attributes have already been used
for $k \in \{0, 1, 2, 3\}$ **do**

while number of variables in \mathcal{V} with k attributes $< N_k$ **do**
 sample k distinct attributes $\mathbf{a}_k \subset \mathcal{A}$ // \mathbf{a}_k is a set of k attributes
 $D' \leftarrow$ filter D by \mathbf{a}_k // keep rows matching attributes in \mathbf{a}_k
if $|D'| < n$ **then**
continue // skip if filtered sample is too small
 $\mu^* \leftarrow \text{mean}[d_v : d \in D']$ // estimate mean on D'
 $se^* \leftarrow \text{SE}(\mu^*; D')$ // estimate standard error on D'
 $\mu_0 \leftarrow \text{mean}[d_v : d \in D]$ // unconditional mean on full D
if $|\mu^* - \mu_0| > \tau$ **and** $|\mu^* - \mu_0| > se^*$ **and** $\mathbf{a}_k \notin \mathcal{S}$ **then**
add $(\mathbf{a}_k, \mu^*, se^*)$ **to** \mathcal{V} // store valid variables
add \mathbf{a}_k **to** \mathcal{S} // store attributes to avoid reuse

return \mathcal{V}

149 While some variables of this kind may overlap with information already present in pretraining corpora
 150 (e.g., widely reported statistics such as overall diabetes prevalence in the United States), many others
 151 are far less likely to have been explicitly documented. In particular, conditional variants of these
 152 quantities—such as *the mean weight of adults with diabetes who are over 40, have elevated cholesterol,*
 153 *and take medication for depression*, or *the median deal size for companies in a specific sector with*
 154 *a given number of employees*—represent fine-grained combinations of attributes that are almost never
 155 reported in textual sources. By systematically varying the conditioning attributes, we generate a large
 156 set of estimation targets that remain grounded in real-world observational data yet are empirically
 157 difficult for LMs to predict.

158
 159 **2.2 SPECIFYING ESTIMATES AS BAYESIAN PRIORS**

160 How should we elicit LM estimates about the likely values of these variables? One simple approach
 161 would be to prompt LMs to produce *point estimates*, then evaluate the accuracy of these point estimates

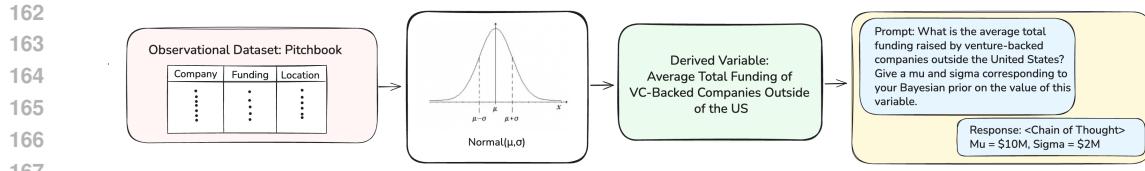


Figure 1: Variable generation and prior elicitation pipeline. We construct derived variables from large-scale observational datasets (e.g., PitchBook), specify them as statistical targets (e.g., Gaussian means), and prompt language models to provide Bayesian priors in the form of distributional parameters.

by reporting the distance (e.g. squared error) between these estimates and the ground-truth value in the data. However, as previously discussed, evaluation of point estimates leaves out much of what is necessary for such predictions to be useful in the real world: with such estimates, it is not possible to distinguish predictions that are right by chance from those that are right as a result of an accurate reasoning procedure; or conversely between predictions that are wrong but confident and predictions that are wrong but highly uncertain. Thus, rather than measuring predictions in the form of point estimates, OPEN-ESTIMATE requires predictions to be specified as probability distributions on the variable of interest.

Models are provided with a brief natural language description of the variable of interest and instructed to select and parameterize the functional form of the target distribution accurately. (Some of our experiments investigate other strategies for eliciting parameters.) For all experiments in this paper, models specified the target distributions as a Gaussian, Beta, or log-normal distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2), \quad X \sim \text{Beta}(\alpha, \beta), \quad \text{or} \quad X \sim \text{LogNormal}(\mu, \sigma^2),$$

depending on whether the target variable is continuous or a proportion. We hypothesize that these three forms are chosen by LMs because they arise frequently in our domains of interest—Gaussians for continuous, symmetric quantities like wages; Betas for proportions like disease rates; and log-normals for right-skewed quantities like startup valuations. (The benchmark itself is agnostic to the choice of parameterization.)

We refer to these distributions as priors to emphasize the fact that they’re not derived directly from examples of the distribution in question from the dataset, and that they can be combined with such samples to produce real posteriors.

2.3 EVALUATION METRICS

Given a prediction from the LM in the form of a probability distribution, how should we evaluate its quality? We focus on two complementary dimensions of performance:

- **Accuracy:** The degree to which the model assigns high probability density to regions close to the empirical ground-truth value.
- **Calibration:** The consistency between the model’s stated uncertainty and empirical frequencies. A model is well-calibrated if events assigned probability p occur with long-run frequency p , such that nominal coverage levels of prediction intervals match their realized coverage.

2.3.1 ACCURACY

To assess accuracy, we ask the question: does the model place the mean of its distribution close to the ground-truth statistic?

To quantify this, we first compute the **mean absolute error (MAE)** between the mode of the predicted distribution, $\hat{p}_i(\mu)$, and the empirical ground-truth value μ_i^* estimated from the full dataset for each of the n variables in the dataset:

$$\text{MAE}_{\text{LLM}} = \frac{1}{n} \sum_{i=1}^n |\mu_i^* - \text{mean}(\hat{p}_i)| .$$

To interpret these errors across variables with different units, we report LM predictions relative to a statistical baseline derived from small empirical samples. Starting from naïve flat priors ($\alpha = 1, \beta = 1$

216 for Beta distributions; $\mu = 0, \sigma^2 = 10^5$ for Gaussians), we draw a random sample \tilde{D} of size $|\tilde{D}| = 5$
 217 from the relevant sub-population (D' in Algorithm 1, corresponding to a sample of e.g. 5 patients
 218 or 5 job postings), from which we can compute a posterior $\tilde{p}_i(\mu | \tilde{D})$.
 219

220 We then compute the statistical baseline MAE as the expected error across such samples:

$$221 \quad \text{MAE}_{\text{baseline}} = \mathbb{E}_{\tilde{D}} |\mu_i^* - \text{mean}(\tilde{p}_i(\cdot | \tilde{D}))| .$$

223 We summarize performance using the error ratio, defined as the LM's MAE relative to this baseline:

$$225 \quad \text{Error Ratio} = \frac{\text{MAE}_{\text{LLM}}}{\text{MAE}_{\text{baseline}}} .$$

227 An error ratio below one indicates that the LM's prediction is more accurate than a small, noisy sample
 228 from the population whose properties are being estimated.
 229

230 We also consider the **win rate** of the LLM prior to the statistical baseline, which is the percentage
 231 of the time that the model's estimate is closer to the ground truth than the statistical baseline:
 232

$$233 \quad \text{Win Rate (LLM prior > baseline)} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\text{MAE}_{\text{LLM}, i} < \text{MAE}_{\text{baseline}, i}\} .$$

235 In addition to the $N = 5$ baseline used for computing MAEs, we report win rates against baselines
 236 with varying numbers of samples.
 237

238 Finally, we evaluate the usefulness of these priors *in combination* with data by computing an **LLM**
 239 **posterior**:

$$240 \quad \hat{p}(\mu | \tilde{D}) \propto \hat{p}(\mu) p(\tilde{D} | \mu) \quad (1)$$

241 (as in the statistical baselines, but replacing the naïve prior with \hat{p}). As with priors, we evaluate the
 242 win rate of these posteriors relative to statistical baselines.

243 Together, these two dimensions provide a more complete picture of accuracy: the error ratio tests
 244 the average error of models relative to the statistical baselines whereas the win rate determines how
 245 consistently the LLMs are outperforming these same baselines.
 246

247 2.3.2 CALIBRATION

249 A model is well-calibrated if the probabilities it assigns correspond to empirical frequencies: events
 250 predicted to occur with probability p should occur about p of the time. In our setting, this means that
 251 the ground-truth value should fall into each predicted quantile with the correct long-run frequency.

252 To measure this, we partition each model's predictive distribution into quartiles and record how often
 253 the ground-truth values fall into each bin. For a perfectly calibrated model, each quartile should contain
 254 the ground truth 25% of the time. Deviations from this ideal reflect miscalibration.

255 Let Q_{ij} be the j -th quartile bin of \hat{p}_i . We define $\hat{q}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\mu_i^* \in Q_{ij}\}$. Formally, we compute
 256 the **quartile expected calibration error (ECE)** as:

$$258 \quad \text{ECE} = \sum_{j=1}^4 |\hat{q}_j - 0.25| .$$

261 Lower values indicate better calibration, with $\text{ECE} = 0$ corresponding to perfect calibration (at
 262 quartile granularity).

263 As a complementary metric, we compute the **continuous ranked probability score (CRPS)**, which pen-
 264 nalizes both miscalibrated predictions and overly dispersed distributions. CRPS measures the distance
 265 between the predicted cumulative distribution function F and the ground truth y without binning:
 266

$$267 \quad \text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{I}(x \geq y))^2 dx$$

268 where $\mathbb{I}(x \geq y)$ is the indicator function. Lower values indicate better predictive performance.
 269

As with MAE, we compare LM performance to a statistical baseline computed from small samples, where $\tilde{p}_i(\cdot \mid \tilde{D})$ is the posterior distribution obtained from a sample \tilde{D} of size $|\tilde{D}|$:

$$\text{CRPS}_{\text{baseline}} = \mathbb{E}_{\tilde{D}} \left[\frac{1}{n} \sum_{i=1}^n \text{CRPS}(\tilde{p}_i(\cdot \mid \tilde{D}), \mu_i^*) \right]$$

We then report the CRPS ratio:

$$\text{CRPS Ratio} = \frac{\text{CRPS}_{\text{LLM}}}{\text{CRPS}_{\text{baseline}}}$$

3 EVALUATION

Our evaluation is divided into two parts. In Section 3.1, we evaluate the zero-shot performance of current language models under standard inference settings, using a consistent elicitation protocol without fine-tuning or prompt engineering. In Section 3.2, we take a deeper look at the best-performing models by analyzing how changes to the system prompt, temperature, and elicitation strategy affect prediction quality.

3.1 ZERO-SHOT EVALUATION

In this section, we focus on zero-shot performance under standard inference settings. We do not apply fine-tuning, retrieval augmentation, or prompt engineering beyond directly asking the model to parameterize the distribution of a variable. To contextualize the LMs’ performance, we compare to four statistical baselines that use $N \in [5, 10, 20, 30]$ examples that are computed using the procedure described in Section 2.3.1.

We evaluate six state-of-the-art language models, including three reasoning models²: Meta Llama 3.1 8B, Meta Llama 3.1 70B (Grattafiori et al., 2024), OpenAI GPT-4 (Achiam et al., 2023), OpenAI o3-mini (OpenAI, 2025a), OpenAI o4-mini (OpenAI, 2025b), and Qwen3-235B-A22B (Yang et al., 2025). We exclude Llama 3.1 8B after it fails to correctly interpret units. We evaluate each model at a medium temperature or reasoning effort—corresponding to 0.5 for GPT-4, “medium” for o3-mini and o4-mini, 0.5 for Llama 3.1 70B Instruct Turbo, and 0.6 for Qwen3-235B-A22B. We use a standard system prompt and prior elicitation prompt which are described in full in Appendix A.1.

Domain	Sample Size	% Prior Better	% Posterior Better
Glassdoor	5	37.0%	71.4%
	10	21.7%	69.0%
	20	13.0%	68.1%
	30	8.7%	70.5%
Pitchbook	5	50.8%	69.6%
	10	50.8%	76.5%
	20	49.2%	80.1%
	30	50.8%	81.6%
NHANES	5	74.3%	70.4%
	10	59.5%	65.1%
	20	47.3%	56.6%
	30	37.8%	50.4%

Table 2: Win rate of the LLM prior relative to an N -sample statistical baseline, and win rate of an LLM posterior (LLM prior + N samples) relative to a statistical baseline (uninformative prior + N samples).

Accuracy. We compare the win rates of LLM priors against statistical baselines computed using $N \in \{5, 10, 20, 30\}$ data points sampled from the true distribution. We fix the model family to o4-mini for this comparison. We also evaluate LLM posteriors, which are formed by updating the LLM prior

²Here, reasoning models are defined as models that have undergone a dedicated training step that involves reinforcement learning for chain-of-thought.

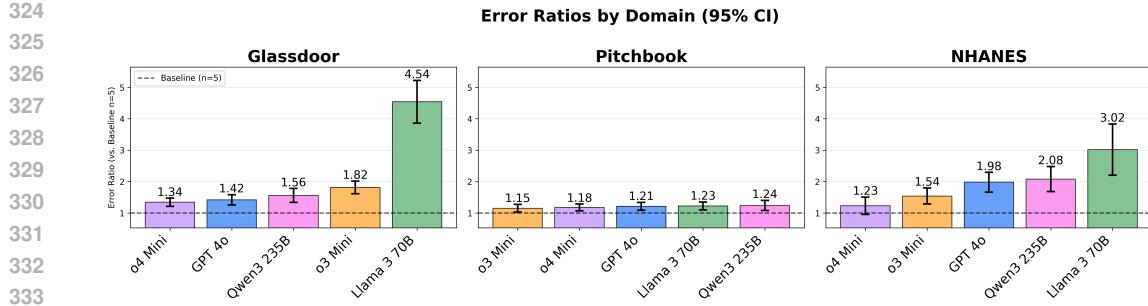


Figure 2: MAE error ratio of LLM prior to a naive statistical baseline computed using a uninformative prior and five examples from the true distribution. Most models are no better than five examples; some are significantly worse. There isn't a statistically significant gap in performance between most model families.

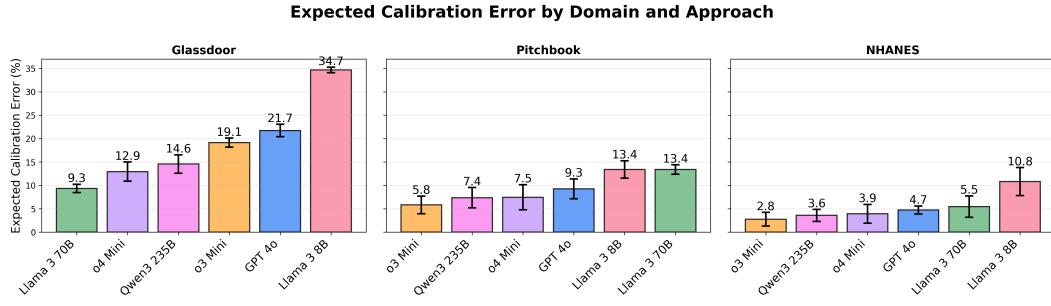


Figure 3: Expected calibration error (in percentage points) across domains and model families. The best model varies by domain, with reasoning models performing the best in Pitchbook and NHANES but not in Glassdoor. Again, most model families are not statistically different from each other in performance.

with the same N examples used to compute each baseline, and compare their win rates against the corresponding statistical baselines. The LLM prior vs. statistical baseline win rate addresses the question: “how many data samples is the LLM prior equivalent to?” The LLM posterior vs. statistical baseline win rate indicates whether incorporating LLM priors yields better posteriors than starting from an uninformative prior.

Results are shown in Table 2. We find that in general, the standalone LLM priors outperform the five-sample baseline in 40-70% of cases, with win rates rapidly dropping off with larger numbers of samples. However, even though these priors are often inaccurate in isolation, they can be effectively combined with data, outperforming or matching baselines with naive priors.

Next, we compare the accuracy of different model families across domains, as defined by MAE relative to the five-sample statistical baseline. The results are shown in Figure 2. We find relatively little variation between most models (with the exception of Llama-70B), and that again, most models have average errors that are no better than five examples; some are significantly worse. This suggests that while the LM priors are often consistently better than the statistical baseline, they are worse in terms of average absolute error. On the whole, these results suggest that OPENESTIMATE is challenging for frontier models. **Calibration.** Next, we assess model calibration.³ First, we consider the overall expected calibration error (ECE) (as defined in Section 2.3.2) of each model family. Results are shown in Figure 3. Larger models and reasoning models tend to outperform smaller, non-reasoning models, but again, no single model family consistently outperforms the rest; specific rankings are domain dependent. The gap between model families is less than 10% across domains with the exception of Llama-3-8b in the Glassdoor domain.

³We exclude the statistical baselines from Figure 4 in this analysis because the baselines derive their posteriors from the same dataset used to compute the ground-truth values. Therefore, larger sample sizes produce extremely tight distributions centered on the ground-truth mean, which leads the ground truth to almost always fall in the middle quartiles (e.g., second or third).

Model	Glassdoor	NHANES	Pitchbook
GPT-4o	3.31	1.86	1.10
Llama-3-70B	4.56	2.76	1.13
Llama-3-8B	10.56	19.17	2.74
Qwen3-235B	2.50	1.65	1.04
o3-mini	3.17	1.35	0.99
o4-mini	2.42	1.17	1.01

Table 3: CRPS Ratio by Model Family Across Domains (vs. 5-Sample Baseline)

Table 3 presents CRPS ratios comparing each model family to the 5-sample baseline and reveals more nuanced differences than ECE. Reasoning models (o3-mini and o4-mini) achieve the best overall performance. Performance varies considerably by domain: in Pitchbook, all models perform comparably to the baseline, while in NHANES, smaller models struggle significantly: Llama-3-8B performs 20 times worse than the baseline. Overall, model size and reasoning capabilities appear most critical in the NHANES domain, while even smaller models achieve reasonable performance in Pitchbook.

Next, we analyze the specific patterns of over- and under-estimation by model family. The results are shown in (Figure 4). All model families exhibit a tendency towards systematic overestimation. In Pitchbook, overestimation is compounded by high rates of underestimation as well, with both tails overweighted.

Next, we examine the cumulative distribution of ground-truth values relative to the predicted priors (Figure 5) to understand how tightly models concentrate their uncertainty. We find the best models cover 80% of the ground truth values within two to three standard deviations of the mean. However, performance is domain-dependent: in Glassdoor and NHANES, the best models cover over 80% of ground-truth values within two standard deviations, while in Pitchbook, three standard deviations are required. This suggests that even the strongest models vary substantially in how they express uncertainty across domains.

Finally, we analyze whether model-reported uncertainty is a reliable guide to predictive accuracy (Figure 6) by comparing the standard deviation ratio to the error ratio. Ideally, models are low error and well-calibrated. In the Glassdoor domain, models appear reasonably well-calibrated relative to the five-sample statistical baseline, but are consistently less accurate than this baseline. In contrast, models in Pitchbook are consistently more confident and less accurate than this baseline. Results in NHANES fall in between these extremes: models generally achieve lower error than in Glassdoor, but their uncertainty estimates are less well-calibrated, with several models exhibiting either under- or over-dispersion. Taken together, these results indicate that the relationship between uncertainty and accuracy is once again strongly domain-dependent.

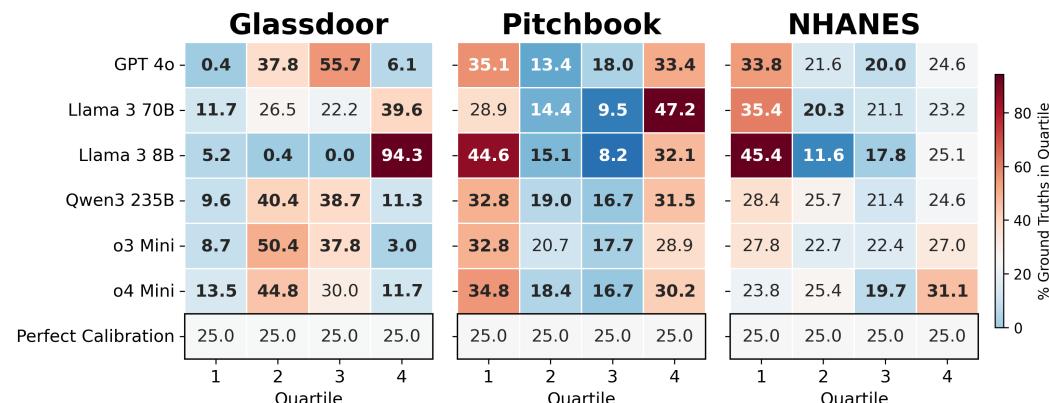


Figure 4: Heatmap describing the deviations from perfect calibration of each approach. Bolded values are statistically significant according to a per-quartile binomial test ($p < 0.05$). All approaches systematically overestimated across domains (Quartile 1 is greater than 25%). In some instances, there was high rates of both over and under-estimation (Quartile 1 and 4 are greater than 25%).

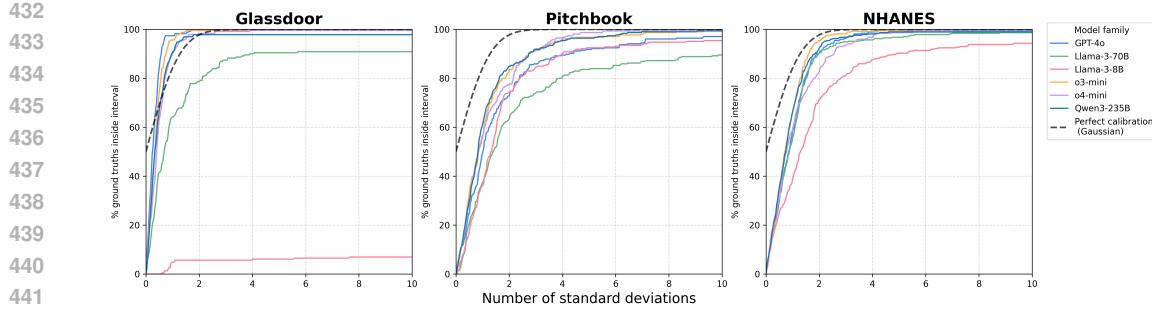


Figure 5: Cumulative distribution function displaying the percentage of ground truth values that fall within $n\sigma$ standard deviations away from the mean of the prior, where σ is the standard deviation of the prior. The dashed line represents perfect calibration for a Gaussian. The best performing models have 80% of the ground truths within 1-2.5 standard deviations from the prior mean. There is overconfidence in Pitchbook and NHANES but underconfidence in Glassdoor.

We also assess whether predictive uncertainty aligns with accuracy by examining the rank correlation between the two for each model family. A stronger correlation between predictive uncertainty and accuracy would indicate that uncertainty is a good indicator of accuracy. However, the reality is mixed: uncertainty is a good indicator of accuracy in NHANES but not necessarily in Pitchbook or Glassdoor.

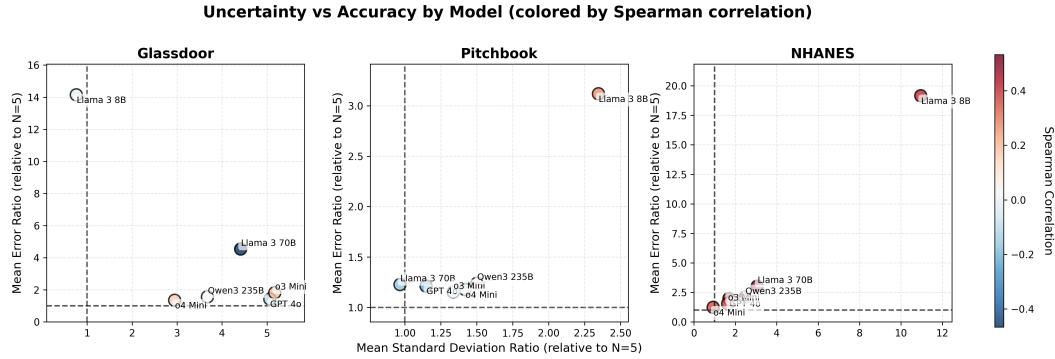


Figure 6: Relationship between uncertainty and accuracy across domains. Each point shows a model’s error ratio versus its standard deviation ratio relative to the $N = 5$ baseline. Colors indicate the Spearman correlation between predictive uncertainty and accuracy within a single model’s predictions, addressing the question of whether a given model tends to be comparatively more confident when it’s more accurate. These correlations differ more so by domain than by model.

3.2 ABLATIONS

We investigate how inference-time settings influence the quality of elicited priors, focusing on three factors: (i) temperature or reasoning effort, (ii) system prompt, and (iii) elicitation protocol. To isolate their effects, we evaluate both a reasoning model (OpenAI o4-mini) and a non-reasoning model (OpenAI gpt-4o). The full set of results is shown in Appendix A.2. None of the settings tested has a consequential impact on performance, indicating that more sophisticated approaches to improving accuracy and calibration are needed.

4 RELATED WORK

Our work intersects with three major lines of language model research: evaluating probabilistic reasoning as a mathematical skill, structuring probabilistic reasoning for better estimation, and applications to forecasting. **Evaluating probabilistic reasoning.** One line of research examines how well LMs perform at problem-solving tasks involving structured probabilistic models. For example, Paruchuri

486 et al. (2024) evaluate models’ probabilistic reasoning given simple idealized distributions; Nafar et al.
 487 (2025) tests models’ ability to provide probabilistic estimates given a Bayesian network; and Jin et al.
 488 (2023) examine the models’ causal reasoning given probabilities. Collectively, these studies frame prob-
 489abilistic reasoning as a mathematical exercise with clearly defined inputs and well-specified outputs. By
 490 contrast, our benchmark targets real-world estimation problems, where the relevant information must be
 491 inferred rather than provided and the ground truth itself may be ambiguous or unavailable. **Structuring**
 492 **probabilistic reasoning.** Another line of work proposes structures for LM-based probabilistic reason-
 493 ing to improve performance. Using “guesstimation” questions similar to ours, Xia et al. (2024) prompt
 494 LMs to propose relevant random variables and moment constraints, and then fits a log-linear distribution
 495 that satisfies these constraints. Feng et al. (2024) take a similar approach, and evaluate a multi-step
 496 process in which LMs brainstorm relevant factors, make coarse probabilistic assessments, and construct
 497 an approximate Bayesian network for inference. Huynh et al. (2025) use LLMs to generate synthetic
 498 counterfactual outcomes by sampling pseudo-observations, constructing empirical distributions. These
 499 approaches extend beyond single-variable reasoning by introducing latent structure and explicit in-
 500 termediate steps. However, the focus for Xia et al. (2024) and Feng et al. (2024) is answering discrete
 501 multiple-choice questions, while Huynh et al. (2025) focuses on augmenting small datasets for down-
 502 stream causal inference tasks rather than directly evaluating the quality of LLM-generated distributions.
 503

504 Like our approach, Selby et al. (2025) elicit parametric Bayesian priors from LLMs. However, they
 505 evaluate priors by comparing them to human expert elicitation in existing psychology studies or to
 506 historical observational data in specific settings (e.g., precipitation and temperature in particular cities
 507 in December). By contrast, we specifically construct *derived variables*—complex aggregations and
 508 cross-sections of tabular data—across diverse domains; we directly evaluate accuracy and calibration
 509 relative to estimated ground truth; and we systematically evaluate how model family and inference-time
 510 settings impact results.

511 **Language model-based forecasting.** Recent studies have also evaluated LMs’ forecasting capabilities
 512 (Karger et al., 2024; Halawi et al., 2024; Ye et al., 2024; Chang et al., 2024; Schoenegger et al., 2025).
 513 These works also test whether models can synthesize heterogeneous evidence into well-calibrated
 514 estimates, but they focus on making predictions about real-world future events. In contrast to our
 515 benchmark, the outcomes of forecasting questions are, by design, highly likely to appear in LMs’
 516 training data after they resolve; they thus perpetually become “stale” and must be replaced with
 517 new questions, as noted by Karger et al. (2024). By focusing on questions that require reasoning
 518 about fine-grained cross of tabular datasets, rather than future events, OPENESTIMATE questions are
 519 designed to remain challenging over time.

519 5 LIMITATIONS AND FUTURE WORK

520 While OPENESTIMATE provides a first step toward evaluating uncertainty in open-domain estimation,
 521 several limitations remain that point to directions for future work. Ground truth values in OPENES-
 522 TIMATE were estimated from finite samples, and therefore might exhibit estimation error. Moreover,
 523 while OPENESTIMATE was constructed to reduce systematic information leakage, leakage still can
 524 occur to varying degrees. In terms of scope, the current benchmark is limited to variables derived
 525 from three datasets across three domains; expanding to new domains would lead to a more thorough
 526 evaluation of priors. In terms of evaluation, we focus our attention on zero-shot methods without
 527 retrieval or fine-tuning; studying training-time interventions for uncertainty awareness and domain
 528 adaptation would be a complementary next step in future work.

529 6 CONCLUSION

530 We introduced OPENESTIMATE, a benchmark and evaluation framework for assessing language
 531 models on open-ended probabilistic estimation with real-world tabular data. The benchmark (i) defines
 532 a realistic task where models must express beliefs as full probability distributions, (ii) elicits priors
 533 through several protocols, and (iii) evaluates performance along accuracy and calibration against
 534 statistical baselines that use only a handful of true samples. By focusing on cross-sectional quantities
 535 from domains such as public health, labor economics, and finance, OPENESTIMATE probes reasoning
 536 under uncertainty while limiting direct lookup and information leakage.

540 REFERENCES
541

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

545 Centers for Disease Control and Prevention. National health and nutrition examination
546 survey (nhanes), 2017–2018. U.S. Government, National Center for Health
547 Statistics, 2018. URL <https://www.cdc.gov/nchs/nhanes>. Data set, accessed via
548 <https://www.cdc.gov/nchs/nhanes>; includes questionnaires, datasets, and documentation.

549 He Chang, Chenchen Ye, Zhulin Tao, Jie Wu, Zhengmao Yang, Yunshan Ma, Xianglin Huang, and
550 Tat-Seng Chua. A comprehensive evaluation of large language models on temporal event forecasting.
551 *arXiv preprint arXiv:2407.11638*, 2024.

552 Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. Bird: A trustworthy bayesian inference framework
553 for large language models. *arXiv preprint arXiv:2404.12494*, 2024.

554 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
555 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd
556 of models. *arXiv preprint arXiv:2407.21783*, 2024.

557 Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level
558 forecasting with language models. *Advances in Neural Information Processing Systems*, 37:
559 50426–50468, 2024.

560 Nicolas Huynh, Julianna Piskorz, Jeroen Berrevoets, Max Ruiz Luyten, and Mihaela
561 van der Schaar. Improving treatment effect estimation with LLM-based data augmentation.
562 In *1st ICML Workshop on Foundation Models for Structured Data*, 2025. URL
563 <https://openreview.net/forum?id=2Q3gFNbpAr>.

564 Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin,
565 Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing
566 causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:
567 31038–31065, 2023.

568 Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and
569 Philip E Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv
570 preprint arXiv:2409.19839*, 2024.

571 Aliakbar Nafar, Kristen Brent Venable, Zijun Cui, and Parisa Kordjamshidi. Extracting probabilistic
572 knowledge from large language models for bayesian network parameterization. *arXiv preprint
573 arXiv:2505.15918*, 2025.

574 OpenAI. Openai o3-mini system card. System card, OpenAI, January 2025a. Published January 31,
575 2025.

576 OpenAI. Openai o3 and o4-mini system card. System card, OpenAI, April 2025b. Published April
577 16, 2025.

578 Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu,
579 and Daniel McDuff. What are the odds? language models are capable of probabilistic reasoning.
580 *arXiv preprint arXiv:2406.12830*, 2024.

581 PitchBook Data. Pitchbook database. Accessed via Wharton Research Data Services (WRDS), 2024.
582 URL <https://wrds-www.wharton.upenn.edu/>. Accessed December 29, 2024.

583 Philipp Schoenegger, Peter S Park, Ezra Karger, Sean Trott, and Philip E Tetlock. Ai-augmented
584 predictions: Llm assistants improve human forecasting accuracy. *ACM Transactions on Interactive
585 Intelligent Systems*, 15(1):1–25, 2025.

586 David Selby, Yuichiro Iwashita, Kai Spriestersbach, Mohammad Saad, Dennis Bappert, Archana
587 Warrier, Sumantrak Mukherjee, Koichi Kise, and Sebastian Vollmer. Had enough of experts?
588 quantitative knowledge retrieval from large language models. *Stat*, 14(2):e70054, 2025.

594 Lionel Wong, Katherine M. Collins, Lance Ying, Adrian Weller Cedegao E. Zhang, Tobias
595 Gersternberg, Alexander K. Lew Timothy O'Donnell, Joshua B. Tenenbaum Jacob D. Andreas,
596 and Tyler Brooke-Wilson. Modeling open-world cognition as on-demand synthesis of probabilistic
597 models. *arXiv:2507.12547*, 2025.

598 Shepard Xia, Brian Lu, and Jason Eisner. Let's think var-by-var: Large language models enable ad
599 hoc probabilistic reasoning. *arXiv preprint arXiv:2412.02081*, 2024.

600 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,
601 Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

602 Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang.
603 Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

648 **A APPENDIX A**
649650 **A.1 ZERO-SHOT ESTIMATION**
651652 We tested Llama 3 8B but excluded it from our analysis because it incorrectly followed instructions
653 pertaining to units and had an average error that was orders of magnitude larger than the other models
654 due to this mistake.655 **System Prompt.**
656

657 Glassdoor

658 You are a helpful assistant that can answer questions about the labor market.
659

660 Pitchbook

662 You are a helpful assistant.
663

664 NHANES

666 You are a helpful assistant that can answer questions about human health.
667668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701

702 A.2 ABLATIONS
703704 **Elicitation Protocol.**
705706
707 Direct

708 You are a statistical
709 expert tasked with constructing a prior distribution for a variable. Your goal
710 is to choose the most appropriate distribution type and estimate its parameters.
711
712 Your estimates should reflect uncertainty about
713 the population-level parameter, not the variation across individual observations.
714
715 Here is the variable you need to model:
716
717 {{variable}}
718
719 {{units_description}}
720
721 Available Distribution Types: Normal (Gaussian), Lognormal, Beta
722
723 Instructions:
724
725 1. Reasoning: First, provide detailed reasoning
726 explaining how you arrived at your specific parameter values. Address: What range
727 do you expect the population parameter to fall in and why? How certain/uncertain
728 are you about this parameter? How do your chosen parameter values translate
729 to meaningful quantities in the original scale (e.g., median, mean, quantiles,
730 credible intervals)? Why is this distribution type appropriate for this variable?
731
732 2. Output: After your reasoning, provide
733 your answer using EXACTLY these XML tags based on which distribution you choose:
734
735 If you choose Normal:
736 <distribution_type>Normal</distribution_type>
737 <mu>value</mu>
738 <sigma>value</sigma>
739
740 If you choose Lognormal:
741 <distribution_type>Lognormal</distribution_type>
742 <mu>value</mu>
743 <sigma>value</sigma>
744
745 CRITICAL: mu and sigma are parameters in LOG-SPACE, not real-space!
746
747 Key relationships to real-space values:
748 - MEDIAN (real-space) = $\exp(\mu)$
749 - MEAN (real-space) = $\exp(\mu + \sigma^2/2)$
750 - MODE (real-space) = $\exp(\mu - \sigma^2)$
751
752 How to set mu: First decide what you think
753 the MEDIAN value should be (in the original units), then set $\mu = \ln(\text{median})$.
754 Examples: If median should be 30 dollars, then $\mu = \ln(30) = 3.4$ approximately.
755 If median should be 100 employees, then $\mu = \ln(100) = 4.6$ approximately.
756 If median should be 1000 dollars, then $\mu = \ln(1000) = 6.9$ approximately.
757
758 How to set sigma:
759 sigma controls the spread in log-space (typical values: 0.2 to 1.0). $\sigma = 0.3$
760 gives roughly a 95 percent credible interval of $[\exp(\mu-0.6), \exp(\mu+0.6)]$. $\sigma = 0.5$
761 gives roughly a 95 percent credible interval of $[\exp(\mu-1.0), \exp(\mu+1.0)]$.
762
763 Common mistake to avoid:
764

756 WRONG: Setting $\mu = 30$ when you
 757 mean the value is 30 dollars (This gives median = $\exp(30) = 10$ trillion dollars!)
 758
 759 CORRECT: Setting $\mu = \ln(30) = 3.4$ approximately
 760 when you mean the value is 30 dollars (This gives median = $\exp(3.4) = 30$ dollars)
 761
 762 Always verify: Calculate $\exp(\mu)$. Does this match
 763 your expected median? Calculate $\exp(\mu + \sigma^2/2)$. Does this match your expected
 764 mean? If these are wildly different from what you expect, you have made an error!
 765
 766 If you choose Beta:
 767 <distribution_type>Beta</distribution_type>
 768 <alpha>value</alpha>
 769 <beta>value</beta>
 770
 771 Critical Unit Check: Pay close attention to units. If the variable says in millions
 772 USD, you need to work in millions! For example, I think the typical company
 773 has raised about 3.5 million dollars. In millions, this is: 3.5, NOT 3500000!
 774
 775 Now, please analyze the variable
 776 and provide your reasoning followed by your distribution choice and parameters.

Quantile

777 You are a statistical expert tasked with constructing a prior distribution
 778 for a variable. Your goal is to choose the most appropriate distribution type and
 779 express your uncertainty about the parameters true value using quantile estimates.

780
 781 Your estimates should reflect uncertainty about
 782 the population-level parameter, not the variation across individual observations.

783
 784 Here is the variable you need to model:

785 {{variable}}

786 {{units_description}}

787 Available Distribution Types:

788 Normal (Gaussian):

789 For variables that can be positive or negative, symmetric around the mean

790 Lognormal:

791 For strictly positive variables, often right-skewed (e.g., prices, sizes, counts)

792 Beta: For variables bounded between 0 and 1 (e.g., proportions, probabilities)

793 Instructions:

794 1. Consider the context of the variable,
 795 including its meaning and any relevant information that informs your beliefs.

796 2. Choose the
 797 most appropriate distribution type based on: The natural bounds of the variable
 798 (can it be negative? is it bounded between 0 and 1?). The expected shape of
 799 uncertainty (symmetric vs. skewed?). The nature of the quantity being estimated.

800 3. Estimate the following percentiles of the parameters true
 801 value: 5th percentile (only a 5 percent chance the true value is below this). 25th
 802 percentile. 50th percentile (median, your best estimate of the true value). 75th
 803 percentile. 95th percentile (only a 5 percent chance the true value is above this).

810
 811 4. Begin your analysis by showing your thought process inside
 812 <parameter_estimation_process> tags. Include the following elements: Explicitly
 813 state the type of parameter being estimated (e.g., population mean, proportion).
 814 Explain why you chose a particular distribution type. List any known facts or data
 815 points about the variable. Consider and list possible data sources or methods for
 816 estimating this parameter. Brainstorm factors that might influence the parameters
 817 value. Note potential biases or limitations in the available information.
 818 State any assumptions you are making. Consider how the parameter might have
 819 changed over time or across different subgroups. Provide your quantile estimates
 820 with a brief explanation for each. Include relevant facts or context about
 821 the variable. Justify your choices. Emphasize population parameter uncertainty
 822 (not individual variability). Reflect on what your estimate spread indicates
 823 about your certainty. Consider any plausible edge cases or alternative scenarios.
 824
 825 5. After your analysis, provide your final answer in the following format:
 826
 827 <distribution_type>[Normal, Lognormal, or Beta]</distribution_type>
 828 <q5>[5th percentile value]</q5>
 829 <q25>[25th percentile value]</q25>
 830 <q50>[50th percentile (median) value]</q50>
 831 <q75>[75th percentile value]</q75>
 832 <q95>[95th percentile value]</q95>
 833
 834 <justification>
 835 [Brief summary of your reasoning, including why you chose this distribution type]
 836 </justification>
 837
 838 <confidence_level>
 839 [Description of how certain or uncertain you are, and why]
 840 </confidence_level>
 841
 842 Examples:
 843
 844 1. Normal Distribution Example:
 845 Variable: Average temperature in a city during summer
 846 Units: Degrees Celsius
 847
 848 <distribution_type>Normal</distribution_type>
 849 <q5>22</q5>
 850 <q25>24</q25>
 851 <q50>26</q50>
 852 <q75>28</q75>
 853 <q95>30</q95>
 854
 855 <justification>
 856 Normal distribution is appropriate because temperature can
 857 be positive or negative and uncertainty about the mean is approximately symmetric.
 858 Based on historical climate data and considering year-to-year variation.
 859 The spread reflects uncertainty in long-term averages due to climate variability.
 860 </justification>
 861
 862 <confidence_level>
 863 Moderately confident. Climate data is well-documented,
 864 but climate change introduces some uncertainty about current averages.
 865 </confidence_level>
 866
 867 2. Lognormal Distribution Example:
 868 Variable: Average home price in a metropolitan area
 869 Units: Thousands of USD
 870
 871 <distribution_type>Lognormal</distribution_type>
 872 <q5>280</q5>
 873 <q25>350</q25>

```

864 <q50>420</q50>
865 <q75>520</q75>
866 <q95>680</q95>
867
868 <justification>
869 Lognormal distribution is appropriate
870     because home prices are strictly positive and typically right-skewed. Based on
871     recent market data and regional economic indicators. The asymmetric spread (wider
872     on the high end) reflects the possibility of higher prices in desirable areas.
873 </justification>
874
875 <confidence_level>
876 Somewhat uncertain. Housing markets are volatile and
877     influenced by many factors including interest rates and local economic conditions.
878 </confidence_level>
879
880 3. Beta Distribution Example:
881 Variable: Proportion of customers who complete a purchase after adding items to cart
882 Units: Proportion (0 to 1)
883
884 <distribution_type>Beta</distribution_type>
885 <q5>0.55</q5>
886 <q25>0.62</q25>
887 <q50>0.68</q50>
888 <q75>0.74</q75>
889 <q95>0.80</q95>
890
891 <justification>
892 Beta distribution is appropriate
893     because this is a proportion bounded between 0 and 1. Based on industry benchmarks
894     for e-commerce conversion rates and typical cart abandonment patterns. The spread
895     accounts for variation across different product categories and customer segments.
896 </justification>
897
898 <confidence_level>
899 Moderately confident. Conversion rates are well-studied
900     in e-commerce, but can vary significantly by industry and website design.
901 </confidence_level>
902
903 Critical Unit Check: Pay close attention to units. If the variable says in millions
904     USD, you need to work in millions. For example, I think the typical company
905     has raised about 3.5 million dollars. In millions, this is 3.5, not 3500000.
906
907 Remember to tailor
908     your analysis to the specific variable and units provided, focusing on uncertainty
909     about the population-level parameter rather than individual variability.
910
911

```

904 Mean-Variance

905 You are a statistical expert tasked with constructing
906 a prior distribution for a variable. Your goal is to choose the most appropriate
907 distribution type and estimate its parameters using mean and standard deviation.

908 Your estimates should reflect uncertainty about
909 the population-level parameter, not the variation across individual observations.

910

911 Here is the variable you need to model:

912 {{variable}}

913 {{units_description}}

914 Available Distribution Types:

```

918
919
920 Normal (Gaussian):
921     For variables that can be positive or negative, symmetric around the mean
922
923 Lognormal:
924     For strictly positive variables, often right-skewed (e.g., prices, sizes, counts)
925
926 Beta: For variables bounded between 0 and 1 (e.g., proportions, probabilities)
927
928 Instructions:
929
930     1. Consider the context of the variable, including what it
931         represents and any relevant information or assumptions that inform your beliefs.
932
933     2. Choose the
934         most appropriate distribution type based on: The natural bounds of the variable
935         (can it be negative? is it bounded between 0 and 1?). The expected shape of
936         uncertainty (symmetric vs. skewed?). The nature of the quantity being estimated.
937
938     3. Estimate the following quantities:
939         Best guess (mean): your estimate of the most likely value of the population-level
940         parameter. Standard deviation: a numerical expression of your uncertainty
941         about the true value, not the variability across individual observations.
942
943     4. Begin your analysis by showing your thought process
944         inside <parameter_estimation_process> tags. Include the following elements:
945
946 Clearly state
947     the type of parameter being estimated (e.g., population mean, true proportion).
948     Explain why you chose a particular distribution type. List any known facts, data
949     points, or previous estimates about the variable. Consider possible data sources,
950     analogous populations, or related studies that inform your belief. Identify
951     key factors that might influence the value of the parameter. Note any limitations,
952     uncertainties, or assumptions in your reasoning. Reflect on how the parameter
953     might differ across subgroups or change over time. Provide your best guess
954     (mean) and your estimate of the standard deviation. Justify your choices with
955     reference to the context, data, and assumptions. Emphasize that your uncertainty
956     pertains to the population parameter, not individual variation. Reflect on what
957     the magnitude of your standard deviation implies about your confidence. Consider
958     plausible edge cases or outliers that helped you calibrate your uncertainty.
959
960     5. After your analysis, provide your final answer in the following format:
961
962         <distribution_type>[Normal, Lognormal, or Beta]</distribution_type>
963         <mean>[Best guess for the true value]</mean>
964         <std_dev>[Standard deviation representing your uncertainty]</std_dev>
965
966         <justification>
967             [Brief summary of your reasoning, including
968                 why you chose this distribution type and what informed your parameter estimates]
969             </justification>
970
971         <confidence_level>
972             [Explanation of how confident or uncertain you are, and why]
973             </confidence_level>
974
975 Examples:
976
977     1. Normal Distribution Example:
978         Variable: Average height of adult males in a country
979         Units: Centimeters
980
981         <distribution_type>Normal</distribution_type>

```

```

972
973 <mean>175</mean>
974 <std_dev>2.5</std_dev>
975
976 <justification>
977 Normal distribution is appropriate because
978 height can theoretically take any value and is approximately symmetric around
979 the mean. Based on global averages, previous studies in similar populations,
980 and considering factors like nutrition and genetics. The standard deviation
981 reflects uncertainty due to potential sampling biases and regional variations.
982 </justification>
983
984 <confidence_level>
985 Moderately confident. While height is
986 well-studied, variations between regions and over time introduce some uncertainty.
987 </confidence_level>
988
989 2. Lognormal Distribution Example:
990 Variable: Average annual revenue of small businesses in a region
991 Units: Thousands of USD
992
993 <distribution_type>Lognormal</distribution_type>
994 <mean>250</mean>
995 <std_dev>150</std_dev>
996
997 <justification>
998 Lognormal distribution
999 is appropriate because revenue is strictly positive and typically right-skewed,
1000 with some businesses earning significantly more than the median. Based on industry
1001 reports and regional economic data. The standard deviation reflects substantial
1002 uncertainty due to variation across industries and economic conditions.
1003 </justification>
1004
1005 <confidence_level>
1006 Somewhat uncertain. Business revenue varies widely by industry
1007 and economic conditions, and available data may not be fully representative.
1008 </confidence_level>
1009
1010 3. Beta Distribution Example:
1011 Variable: Proportion of people who prefer tea over coffee in a city
1012 Units: Proportion (0 to 1)
1013
1014 <distribution_type>Beta</distribution_type>
1015 <mean>0.6</mean>
1016 <std_dev>0.05</std_dev>
1017
1018 <justification>
1019 Beta distribution is appropriate because this is a proportion bounded
1020 between 0 and 1. Estimated based on local cultural preferences, limited survey
1021 data, and comparison with similar cities. The standard deviation accounts for
1022 potential biases in available data and variations across different demographics.
1023 </justification>
1024
1025 <confidence_level>
1026 Somewhat uncertain. Beverage preferences can vary significantly based on
1027 factors like age, cultural background, and local trends, which are not fully known.
1028 </confidence_level>
1029
1030 Critical Unit Check: Pay close attention to units. If the variable says in millions
1031 USD, you need to work in millions. For example, I think the typical company
1032 has raised about 3.5 million dollars. In millions, this is 3.5, not 3500000.
1033
1034 Remember: you are
1035 modeling beliefs about the parameter, not the spread of raw data. Your standard
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
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
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134

```

1026
1027
1028

deviation should reflect how much uncertainty you have about the single true value that governs the population, not the spread of outcomes across individuals.

1029
1030
1031

Provide your analysis and final answer based on the given variable and units description. Your final output should consist only of the formatted answer and should not duplicate or rehash any of the work you did in the thinking block.

1032

Additional Results.

1033

1034

1035

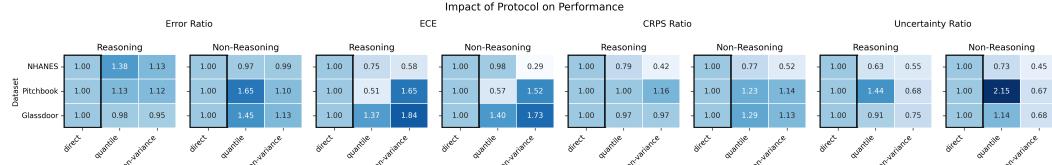
1040
1041
1042

Figure 7: Effect of elicitation protocol (direct, quantile, mean–variance) on error ratio, expected calibration error (ECE), CRPS ratio, and uncertainty (standard deviation) across reasoning and non-reasoning models, relative to direct elicitation.

1043

1044

1045

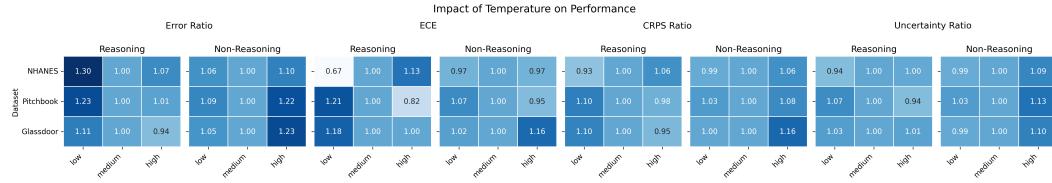
1051
1052
1053

Figure 8: We examine the impact of changing temperature or reasoning effort on accuracy, calibration, and certainty.

1054

1055

1056

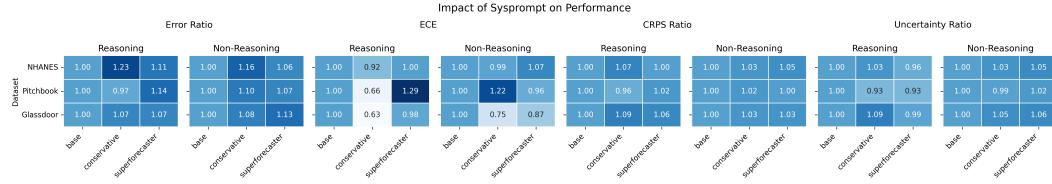
1062
1063

Figure 9: We examine the impact of changing the system prompt or reasoning effort on accuracy, calibration, and certainty.

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079