Triage Against the Machine: Can AI Reason Deliberatively?

Francesco Veri, Gustavo Umbelino

2025-04-08

Large-Language Models (LLMs) Preview

Table 1: LLMs

	provider	model	type	version
1	anthropic	claude-3-5-haiku-20241022	NA	NA
2	anthropic	claude-3-5-sonnet-20241022	NA	NA
3	anthropic	claude-3-7-sonnet-20250219	NA	NA
4	anthropic	claude-3-haiku-20240307	NA	NA
5	anthropic	claude-3-opus-20240229	NA	NA
6	anthropic	claude-3-sonnet-20240229	NA	NA
7	cohere	command	NA	NA
8	cohere	command-r-08-2024	NA	NA
9	cohere	command-r-plus-08-2024	NA	NA
10	cohere	command-r7b-12-2024	NA	NA
11	deepseek	deepseek-chat	NA	NA
12	deepseek	deepseek-reasoner	reason	NA
13	deepseek	deepseek-v2	NA	NA
14	deepseek	deepseek-v2.5	NA	NA
15	google	gemini-1.5-flash	NA	NA
16	google	gemini-1.5-flash-8b	NA	NA
17	google	gemini-1.5-pro	NA	NA
18	google	gemini-2.0-flash	NA	NA
19	google	gemma	NA	NA
20	google	gemma2:27b	NA	NA
21	google	gemma3:12b	NA	NA
22	meta	llama2:13b	NA	NA
23	meta	llama2:70b	NA	NA
24	meta	llama3.1:405B-turbo	NA	NA
25	meta	llama3.2	NA	NA
26	meta	llama3.3:70b	NA	NA
27	meta	llama3:70b	NA	NA
28	microsoft	phi	NA	NA
29	microsoft	phi2	NA	NA
30	microsoft	phi3	NA	NA
31	microsoft	phi3.5	NA	NA
32	microsoft	phi4	NA	NA
33	mistralai	ministral-3b-latest	NA	NA
34	mistralai	ministral-8b-latest	NA	NA
35	mistralai	mistral-large-latest	reason	NA
36	mistralai	mistral-small-latest	NA	NA

	provider	model	type	version
37	mistralai	open-mistral-7b	NA	NA
38	mistralai	open-mistral-nemo	NA	NA
39	mistralai	open-mixtral-8x22b	SMoE	NA
40	mistralai	open-mixtral-8x7b	SMoE	NA
41	openai	gpt-3.5-turbo	NA	NA
42	openai	gpt-4	NA	NA
43	openai	gpt-4-turbo	NA	NA
44	openai	gpt-4.5-preview	NA	NA
45	openai	gpt-4o	NA	NA
46	openai	gpt-4o-mini	NA	NA
47	openai	o1	reason	NA
48	openai	o1-mini	reason	NA
49	openai	o3-mini	reason	NA
50	qwen	qwen-max	NA	NA
51	qwen	qwen-plus	NA	NA
52	qwen	qwen-turbo	NA	NA
53	qwen	qwen1.5-110b-chat	NA	NA
54	qwen	qwen1.5-72b-chat	NA	NA
55	qwen	qwen2-72b-instruct	NA	NA
56	qwen	qwen2.5-72b-instruct	NA	NA
57	qwen	qwq-plus	reason	NA
58	xai	grok-2-1212	NA	NA

We started the analysis with 58 models, but some models were dropped after data collection. The models and reason for dropping are discussed later on Excluded Models.

Surveys

Table 2: Surveys

	survey	considerations	policies	scale_max	q_method
1	acp	48	5	11	FALSE
2	auscj	45	8	7	FALSE
3	bep	43	7	7	FALSE
4	biobanking_mayo_ubc	38	7	11	FALSE
5	biobanking_wa	49	7	11	FALSE
6	ccps	33	7	11	FALSE
7	ds_aargau	33	7	7	FALSE
8	ds_bellinzona	32	7	7	FALSE
9	energy_futures	45	9	11	FALSE
10	fnqcj	42	5	12	FALSE
11	forestera	45	7	11	FALSE
12	fremantle	36	6	11	TRUE
13	gbr	35	7	7	FALSE
14	swiss_health	24	6	7	FALSE
15	uppsala_speaks	42	7	7	FALSE
16	valsamoggia	36	4	11	TRUE
17	zh_thalwil	31	7	7	FALSE
18	zh_uster	31	7	7	FALSE
19	zh_winterthur	30	6	7	FALSE

	survey	considerations	policies	scale_max	q_method
20	zukunft	20	7	7	FALSE

LLM Data Collection

We collected a total of 30456 valid LLM responses across 20 surveys.

Cost

We spent a total of 313.71 USD. The cost breakdown per API is below.

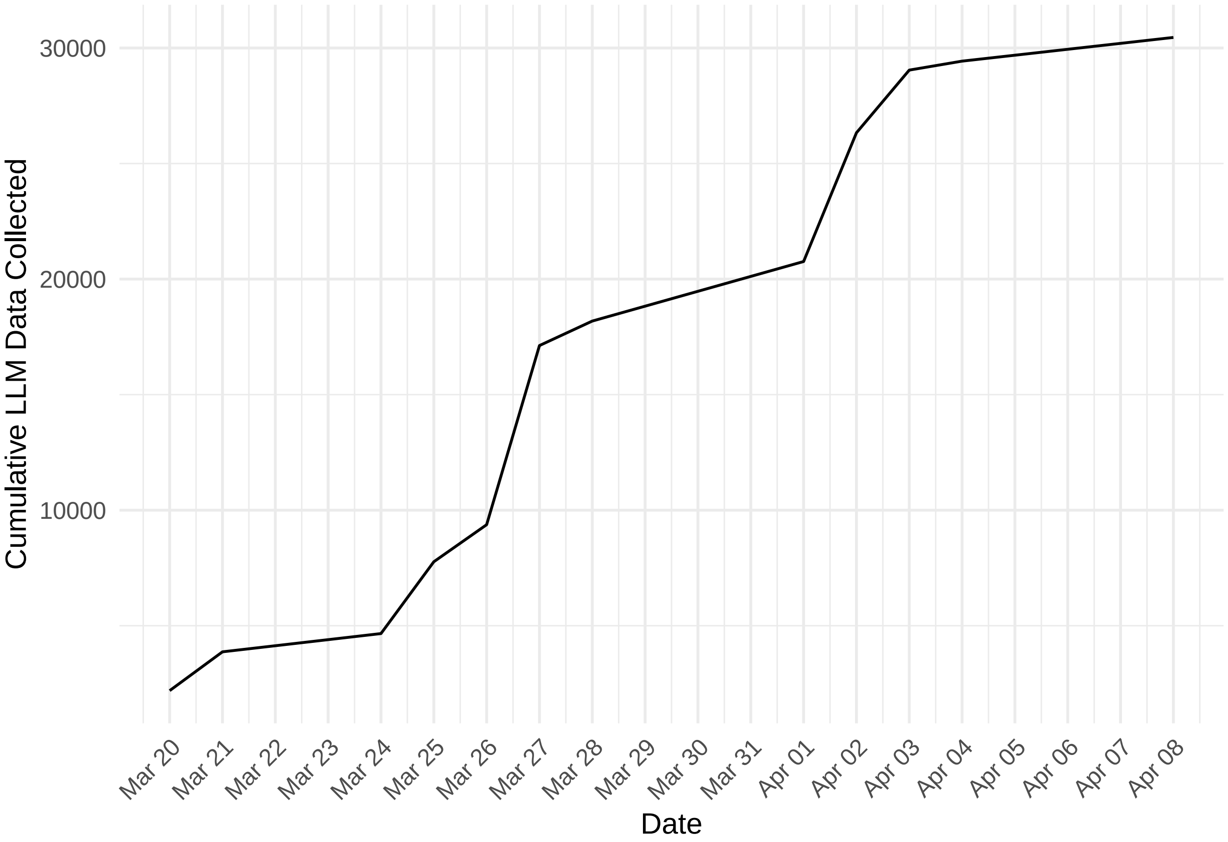
Table 3: Costs by API

api	num_models	credits_paid
OpenAI API	9	155.52
Anthropic API	6	75.00
Mistral AI API	8	20.00
Alibaba Cloud	8	17.49
Together AI	6	13.00
Cohere API	4	12.70
DeepSeek API	2	10.00
xAI API	1	10.00
Google Could	4	NA
ollama	9	NA

Time

It took a total of 150 hours[1] across 19 days to complete data collection. Most of it was done in parallel. The first LLM response was collected on Thursday, Mar 20, 2025 and latest on Tuesday, Apr 08, 2025.

[1] Execution data is mostly accurate. Only a few (3-5) executions failed and, as a result, we have no record of it.



Excluded Models

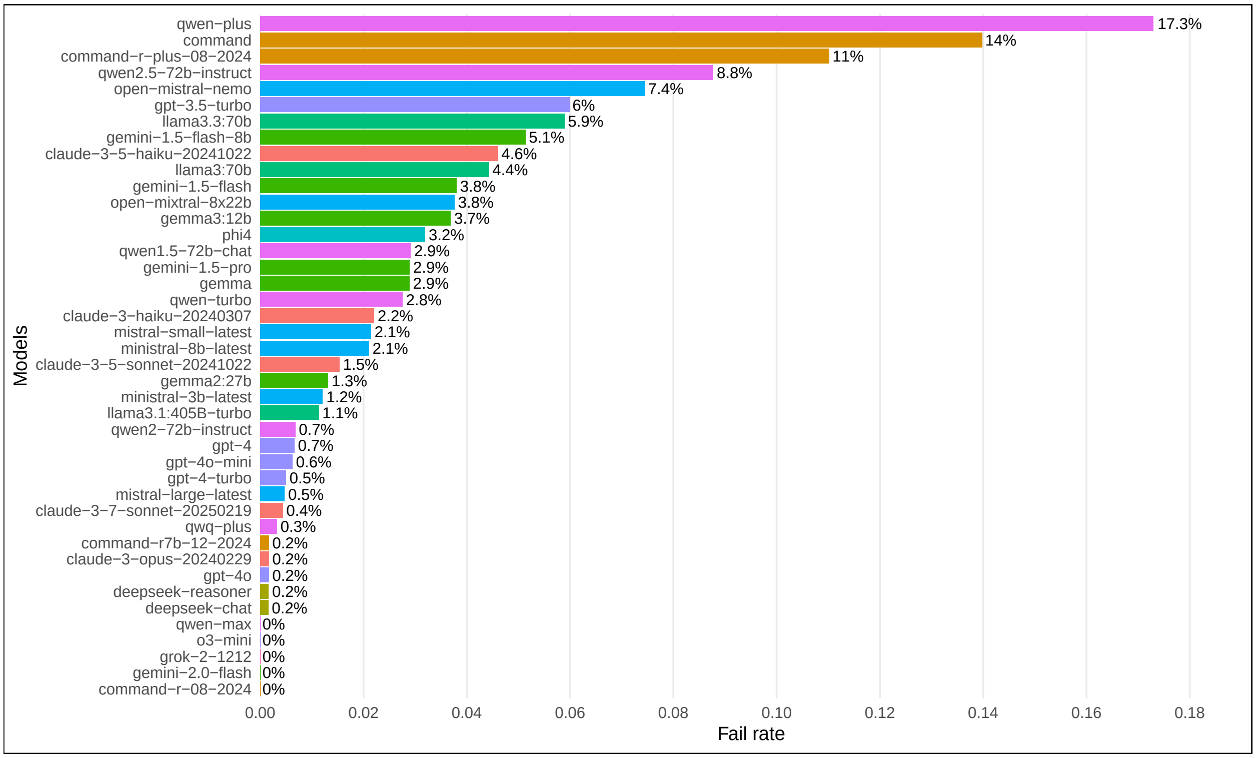
14 out of 60 were excluded from the analysis for the following reasons.

Table 4: Excluded models and reasons

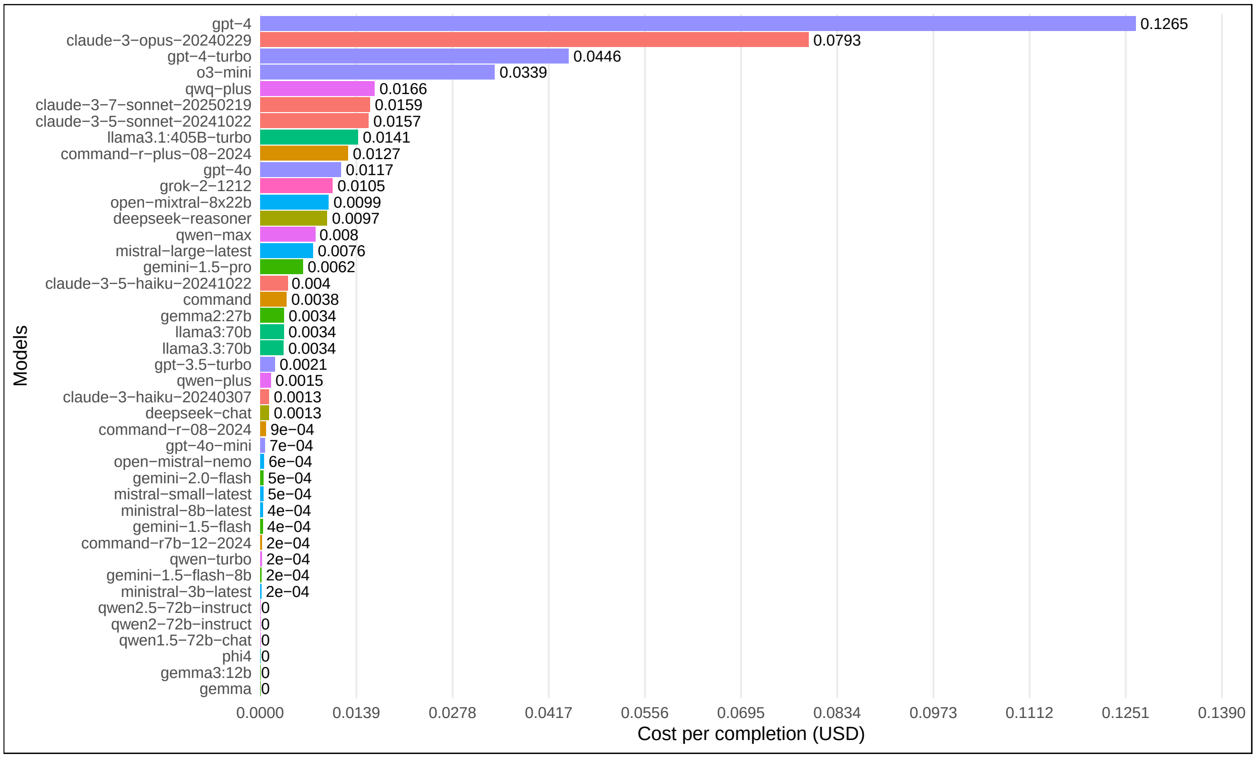
provider	model	reason
anthropic	claude-3-sonnet-20240229	not available in Anthropic API anymore
deepseek	deepseek-v2	high fail rate (85%)
deepseek	deepseek-v2.5	too big to run locally; not available through APIs
meta	llama2:13b	does not respond to prompts correctly
meta	llama2:70b	does not respond to prompts correctly
meta	llama3.2	3% success rate on auscj
microsoft	phi	does not respond to prompts correctly
microsoft	phi2	same model as phi
microsoft	phi3	does not respond to prompts correctly
microsoft	phi3.5	10% success rate for biobanking_wa
mistralai	open-mistral-7b	11% success rate for auscj, uppsala_speaks, and biobanking_wa
mistralai	open-mixtral-8x7b	6% success rate on fremantle only
openai	o1-mini	0% success rate on uppsala_speaks only; responds with “I’m sorry, but I can’t help with that.”
qwen	qwen1.5-110b-chat	has API limit of 10 RPM; too slow

Execution Summary Plots

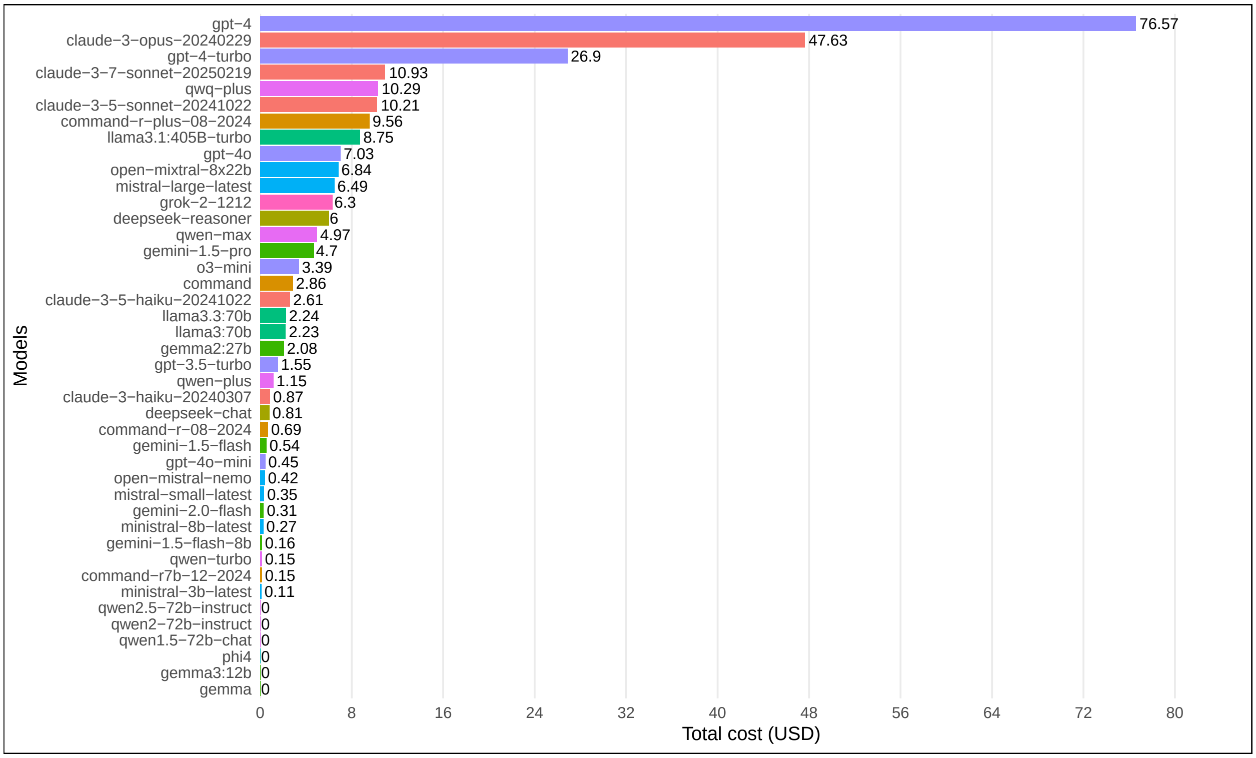
Fail rate



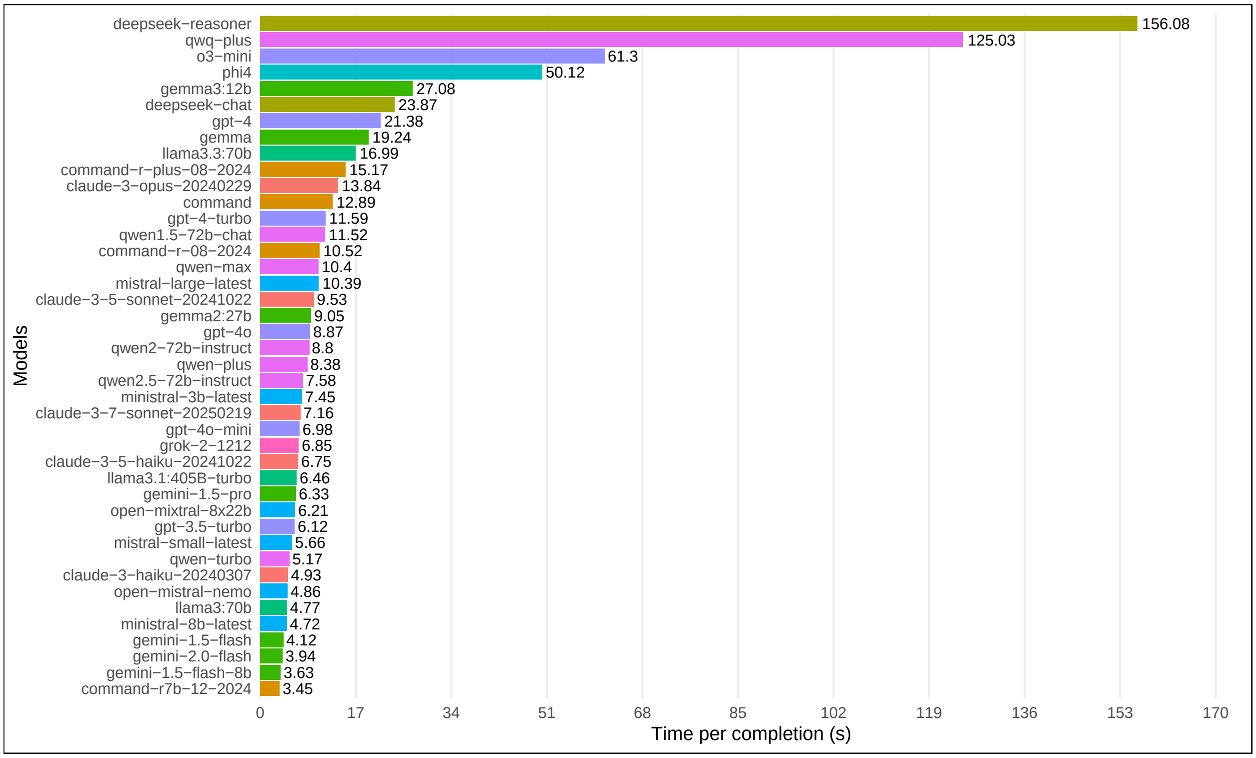
Cost per completion



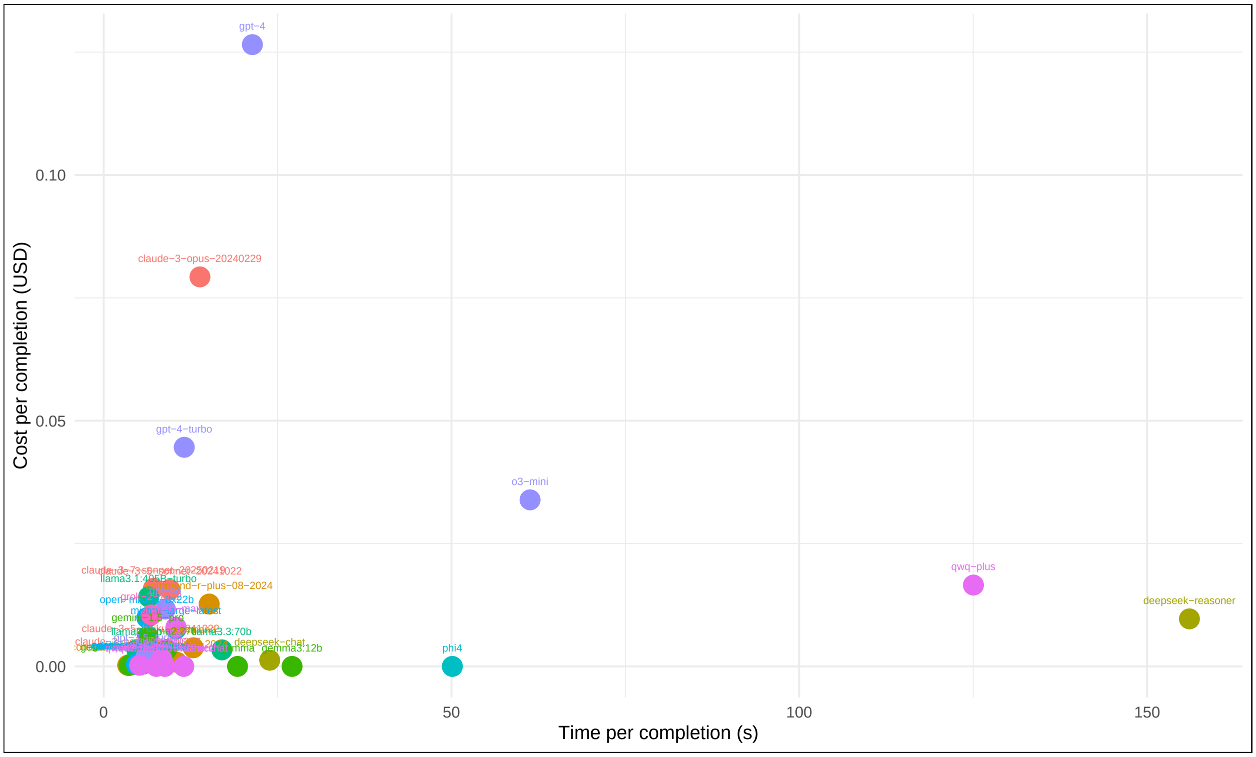
Total cost



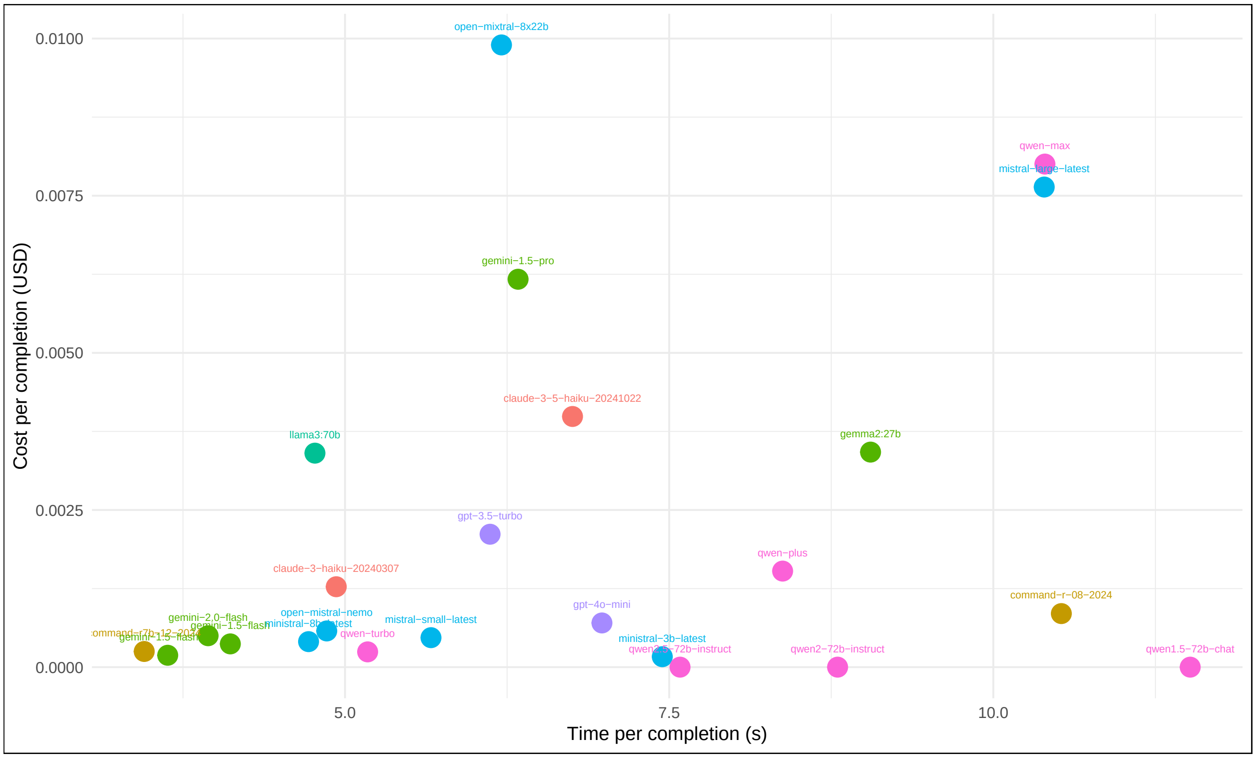
Time per completion



Cost/Time per completion



Zoomed in to cost < 0.01 USD and time < 12 s.



Internal Consistency of Responses

We calculate Cronbach's Alpha from the top 30 iterations.

Check alpha results per model

Table 5: Alpha summary across models, mean across surveys

	provider	model	N	all	considerations	policies
1	qwen	qwen1.5-72b-chat	600	0.70	0.75	0.49
2	google	gemma2:27b	600	0.71	0.75	0.50
3	openai	gpt-4o-mini	600	0.72	0.74	0.45
4	anthropic	claude-3-haiku-20240307	600	0.74	0.82	0.44
5	google	gemini-1.5-flash	600	0.74	0.76	0.52
6	anthropic	claude-3-5-sonnet-20241022	600	0.75	0.81	0.58
7	deepseek	deepseek-reasoner	600	0.75	0.79	0.55
8	openai	gpt-4	600	0.75	0.82	0.52
9	openai	gpt-4-turbo	600	0.75	0.82	0.53
10	google	gemini-1.5-pro	600	0.76	0.78	0.57
11	openai	gpt-4o	600	0.76	0.86	0.50
12	cohere	command	600	0.78	0.78	0.44
13	google	gemma	600	0.78	0.80	0.45
14	meta	llama3.3:70b	600	0.78	0.82	0.52
15	mistralai	mistral-small-latest	600	0.78	0.84	0.52
16	mistralai	open-mistral-nemo	600	0.78	0.80	0.49
17	qwen	qwq-plus	600	0.78	0.79	0.58
18	xai	grok-2-1212	600	0.78	0.89	0.47
19	cohere	command-r-08-2024	600	0.79	0.81	0.50
20	deepseek	deepseek-chat	600	0.79	0.86	0.52
21	google	gemini-1.5-flash-8b	600	0.79	0.84	0.50
22	meta	llama3:70b	600	0.79	0.79	0.52
23	qwen	qwen-turbo	600	0.79	0.83	0.48
24	anthropic	claude-3-7-sonnet-20250219	600	0.80	0.84	0.53
25	qwen	qwen-plus	600	0.80	0.82	0.49
26	qwen	qwen2-72b-instruct	600	0.80	0.86	0.48
27	qwen	qwen2.5-72b-instruct	600	0.80	0.84	0.51
28	anthropic	claude-3-5-haiku-20241022	600	0.81	0.86	0.47
29	google	gemma3:12b	600	0.81	0.81	0.47
30	microsoft	phi4	600	0.81	0.82	0.55
31	mistralai	ministral-8b-latest	600	0.82	0.83	0.51
32	qwen	qwen-max	600	0.82	0.84	0.51
33	anthropic	claude-3-opus-20240229	600	0.83	0.87	0.50
34	mistralai	mistral-large-latest	600	0.83	0.86	0.54
35	google	gemini-2.0-flash	600	0.84	0.84	0.62
36	openai	gpt-3.5-turbo	600	0.84	0.87	0.48
37	meta	llama3.1:405B-turbo	600	0.85	0.88	0.49
38	mistralai	ministral-3b-latest	600	0.85	0.86	0.53
39	cohere	command-r7b-12-2024	600	0.86	0.87	0.46
40	cohere	command-r-plus-08-2024	600	0.87	0.89	0.49
41	mistralai	open-mixtral-8x22b	600	0.87	0.90	0.52
42	openai	o3-mini	100	0.92	0.91	0.80

Aggregation

We then aggregated LLM data into 1 response per model/survey. Based on (Motoki, Pinho Neto, and Rodrigues 2024), we bootstrap considerations 1000 times.

Aggregate considerations and preferences

We aggregated 30456 LLM responses into 959 responses: 1 response per model per survey.

Human Data

Table 6: Number of participants in each case study

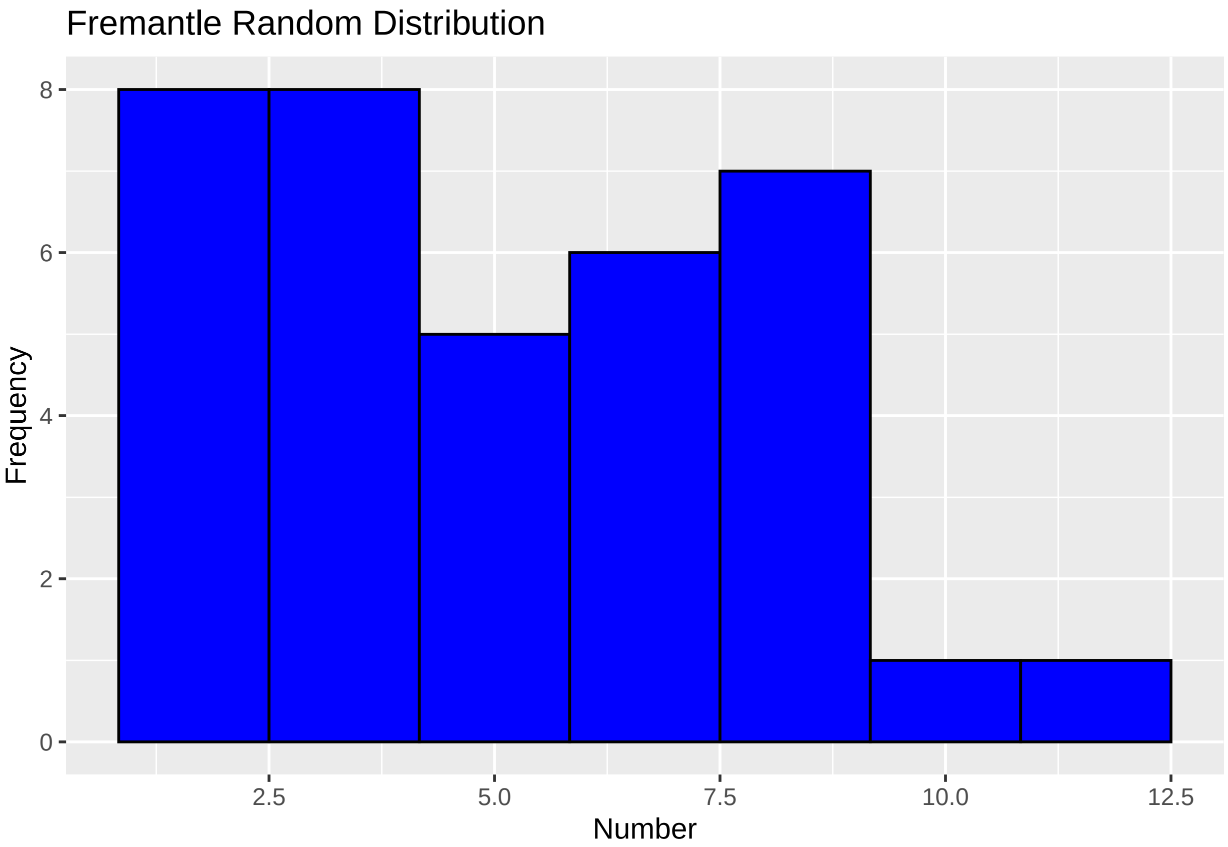
	Case	survey	participants
1	Citizen Parliamentarian	acp	45
2	HGE Control Group	auscj	19
3	HGE Deliberative Group	auscj	23
4	BEP	bep	16
5	Mayo	biobanking_mayo_ubc	17
6	UBC Bio	biobanking_mayo_ubc	17
7	WA Citizens	biobanking_wa	9
8	WA Stakeholder	biobanking_wa	15
9	CCPS ACT Deliberative	ccps	31
10	Aargau	ds_aargau	16
11	Bellinzona	ds_bellinzona	8
12	CSIRO NSW	energy_futures	12
13	CSIRO WA	energy_futures	17
14	FNQCJ	fnqcj	11
15	Forest Lay Citizen	forestera	9
16	Forest Stakeholder	forestera	11
17	Fremantle	fremantle	41
18	GBR	gbr	7
19	Activate	uppsala_speaks	26
20	Standard	uppsala_speaks	22
21	UPSA Control Group	uppsala_speaks	20
22	Valsamoggia	valsamoggia	16
23	Thalwill	zh_thalwil	14
24	USTER	zh_uster	15
25	Winterthur	zh_winterthur	16
26	Zukunft	zukunft	63

We collected 1032 human responses across 26 case studies, including pre-post deliberation responses.

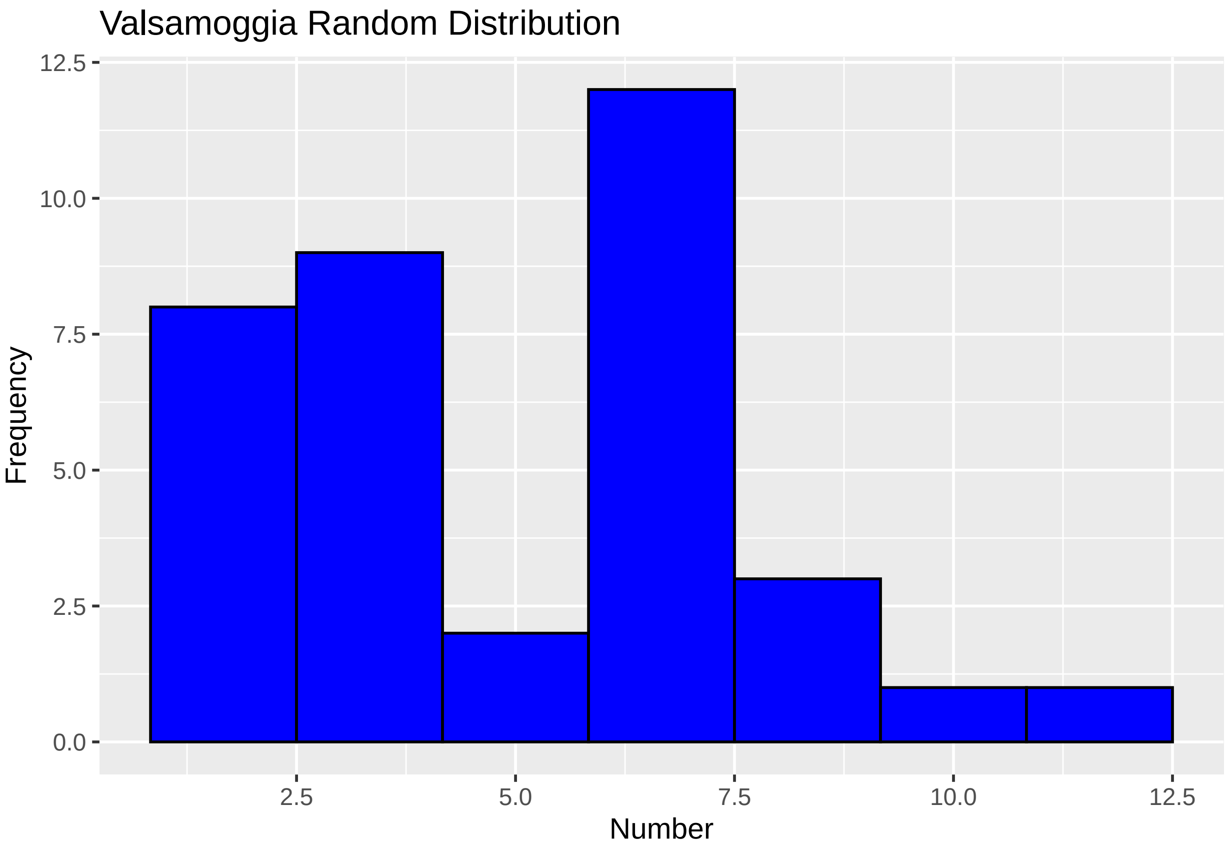
Randomly Generated Data

Then, we generated 20 random reseponses for each survey.

```
## Warning: Removed 14 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
## Warning: Removed 14 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



DRI Analysis

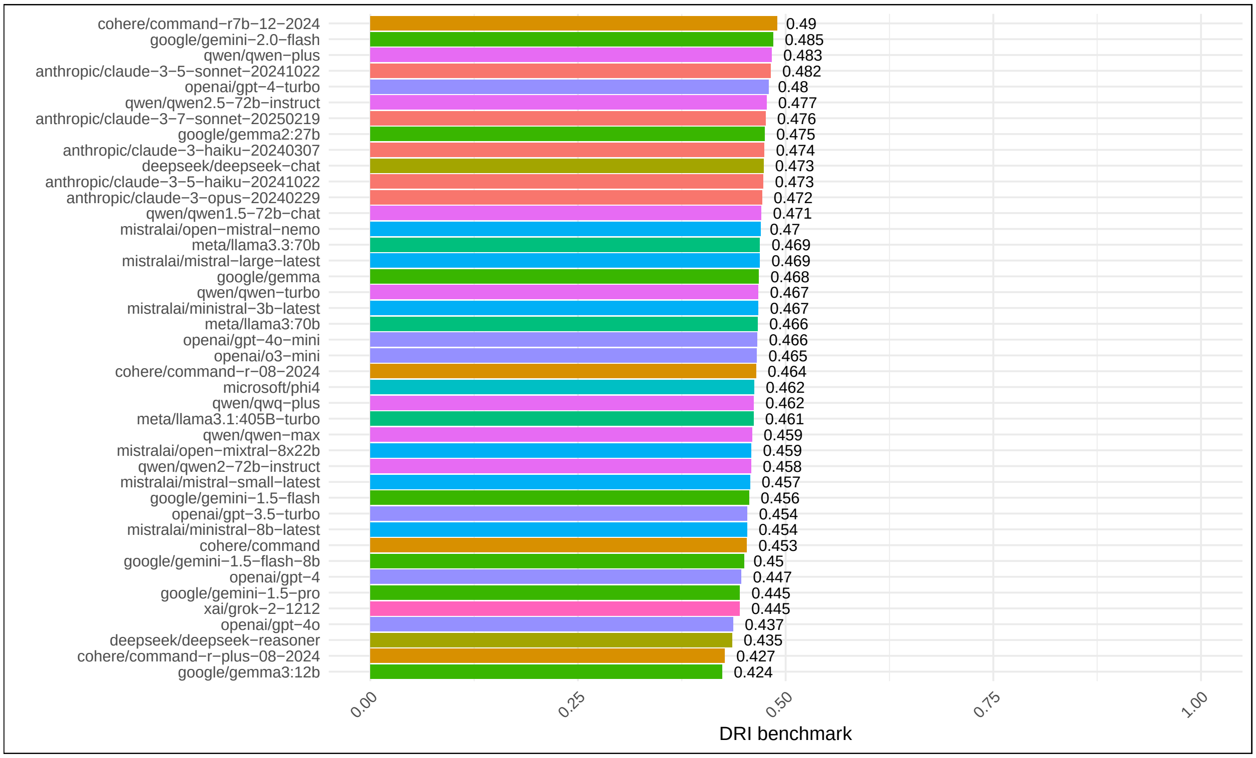
We begin by defining DRI calculation functions.

```
# original DRI formula
dri_calc <- function(data, v1, v2) {
  lambda <- 1 - (sqrt(2) / 2)
  dri <- 2 * (((1 - mean(abs((data[[v1]] - data[[v2]]) / sqrt(2)
  ))) - (lambda)) / (1 - (lambda))) - 1

  return(dri)
}

# updated DRI formula
# FIXME: only accounts for negligible positive correlations, but not negative ones
dri_calc_v2 <- function(data, v1, v2) {
  # Calculate orthogonal distance for each pair
  d <- abs((data[[v1]] - data[[v2]]) / sqrt(2))

  # Define lambda as in the original
  lambda <- 1 - (sqrt(2) / 2)

  # Calculate penalty: 0.5 if both correlations are in [0, 0.2], 1 otherwise
  penalty <- ifelse(data[[v1]] >= 0 & data[[v1]] <= 0.2 & #0.3
                      data[[v2]] >= 0 & data[[v2]] <= 0.2, # 0.3
                    0, 1)
  # Adjusted consistency per pair
  consistency <- (1 - d) * penalty

  # Average consistency across all pairs
  avg_consistency <- mean(consistency)

  # Scale to [-1, 1] as in the original
  dri <- 2 * ((avg_consistency - lambda) / (1 - lambda)) - 1

  return(dri)
}

# updated DRI formula: penalizes both negligible positive and negative correlations in a scalar way.
dri_calc_v3 <- function(data, v1, v2){
  d <- abs((data[[v1]] - data[[v2]]) / sqrt(2))
  lambda <- 1 - (sqrt(2) / 2)

  # Scalar penalty based on strength of signal (|r| and |q|)
  penalty <- ifelse(pmax(abs(data[[v1]]), abs(data[[v2]])) <= 0.2,
                    pmax(abs(data[[v1]]), abs(data[[v2]])) / 0.2,
                    1)

  consistency <- (1 - d) * penalty
  avg_consistency <- mean(consistency)

  dri <- 2 * ((avg_consistency - lambda) / (1 - lambda)) - 1
  return(dri)
}
```

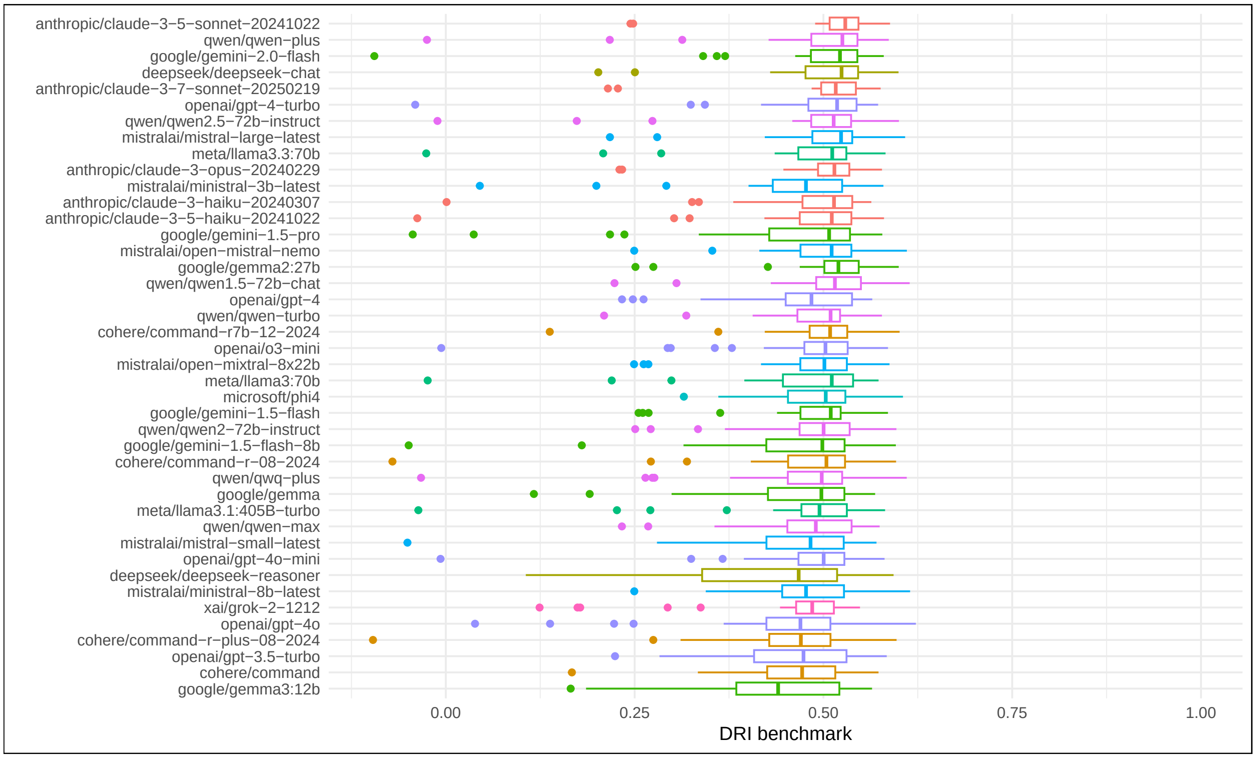
```
## Warning in cor(Q, method = "spearman"): the standard deviation is zero
## Warning in cor(Q, method = "spearman"): the standard deviation is zero
## Warning in cor(Q, method = "spearman"): the standard deviation is zero
## Warning in cor(Q, method = "spearman"): the standard deviation is zero
```

```
## Warning: Missing swiss_health from DRIInd.LLMs!
```

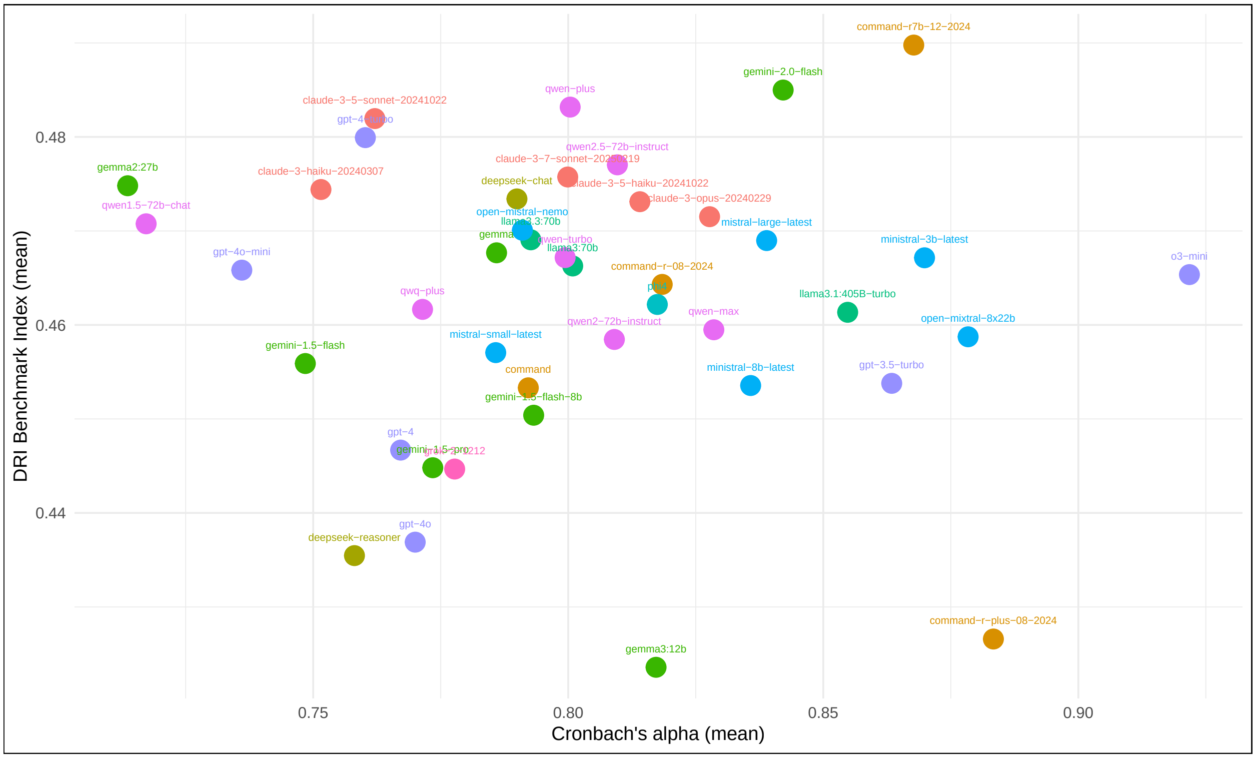
DRI Benchmark



```
## Warning: Removed 18 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
## Warning: Removed 18 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
## `summarise()` has grouped output by 'provider', 'model'. You can override using
## the `.groups` argument.
```



Francesco's DRI Analysis

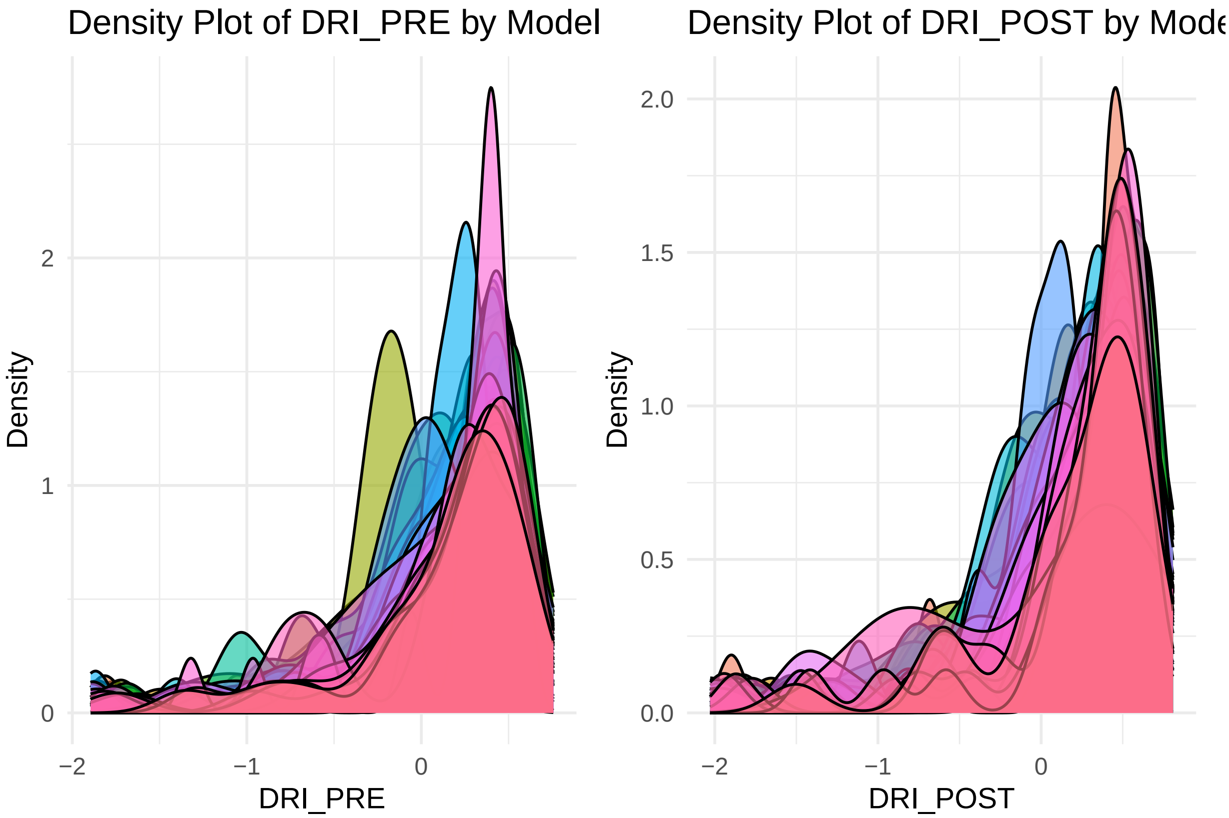
```
##
## Attaching package: 'Metrics'

## The following object is masked from 'package:rlang':
##
##     ll

## `summarise()` has grouped output by 'model'. You can override using the
## `.groups` argument.

## # A tibble: 1,249 x 4
## # Groups:   model [50]
##    model                      case                    DRI_POST DRI_PRE
##    <fct>                      <fct>                      <dbl>   <dbl>
##  1 claude-3-5-haiku-20241022 Aargau                    0.0461  0.272
##  2 claude-3-5-haiku-20241022 Activate                  0.286   0.260
##  3 claude-3-5-haiku-20241022 Bellinzona                0.547   0.281
##  4 claude-3-5-haiku-20241022 BEP                       0.344   0.410
##  5 claude-3-5-haiku-20241022 CCPS ACT Deliberative     0.685   0.538
##  6 claude-3-5-haiku-20241022 Citizen Parliamentarian  -0.0470 -0.0987
##  7 claude-3-5-haiku-20241022 CSIRO NSW                 0.163   0.142
##  8 claude-3-5-haiku-20241022 CSIRO WA                  0.337   0.369
##  9 claude-3-5-haiku-20241022 FNQCJ                     0.186  -0.570
## 10 claude-3-5-haiku-20241022 Forest Lay Citizen        0.690   0.578
## # i 1,239 more rows
## # A tibble: 50 x 3
##    model                      DRI_POST  DRI_PRE
##    <fct>                         <dbl>    <dbl>
##  1 claude-3-5-haiku-20241022  -0.0470  -0.0987
##  2 claude-3-5-sonnet-20241022  0.146    0.147
##  3 claude-3-7-sonnet-20250219  0.0449  -0.0269
##  4 claude-3-haiku-20240307     0.0940   0.139
##  5 claude-3-opus-20240229      0.0445   0.154
##  6 command                    -0.179   -0.169
##  7 command-r-08-2024          -0.235    0.208
##  8 command-r-plus-08-2024     -0.145   -0.190
##  9 command-r7b-12-2024        -0.0953  -0.0716
## 10 deepseek-chat              -0.192   -0.104
## 11 deepseek-reasoner          -0.104   -0.0250
## 12 deepseek-v2                 0.500    0.484
## 13 gemini-1.5-flash           -0.108   -0.485
## 14 gemini-1.5-flash-8b        -0.0492  -0.387
## 15 gemini-1.5-pro              0.145   -0.248
## 16 gemini-2.0-flash            0.249   -0.0930
## 17 gemma                      -0.112    0.0450
## 18 gemma2:27b                 -0.0969   0.101
## 19 gemma3:12b                  0.0444  -0.201
## 20 gpt-3.5-turbo              -0.261   -0.110
## 21 gpt-4                      -0.230   -0.101
## 22 gpt-4-turbo                 0.0512  -0.160
## 23 gpt-4o                     -0.167    0.133
## 24 gpt-4o-mini                 0.129    0.0546
## 25 grok-2-1212                -0.160   -0.0430
## 26 llama3:70b                  0.146    0.00367
## 27 llama3.1:405B-turbo        -0.108   -0.139
## 28 llama3.2                   -0.158   -0.302
## 29 llama3.3:70b                0.0296  -0.332
## 30 ministral-3b-latest        -0.0174  -0.364
## 31 ministral-8b-latest        -0.166    0.0591
## 32 mistral-large-latest        0.0534  -0.307
## 33 mistral-small-latest       -0.253   -0.0824
## 34 o1                         -0.0809   0.0558
## 35 o1-mini                    -0.318   -0.0555
## 36 o3-mini                    -0.0385   0.0531
## 37 open-mistral-7b            -0.00689 -0.204
## 38 open-mistral-nemo           0.00443 -0.209
## 39 open-mixtral-8x22b          0.0207  -0.199
## 40 open-mixtral-8x7b          -0.147   -0.160
## 41 phi3.5                      0.120    0.0457
## 42 phi4                       -0.0270  -0.218
## 43 qwen-max                    0.0497  -0.151
## 44 qwen-plus                   0.306   -0.0312
## 45 qwen-turbo                  0.0372  -0.0637
## 46 qwen1.5-110b-chat           0.446    0.425
## 47 qwen1.5-72b-chat           -0.0607  -0.0526
## # i 3 more rows
```

```
## # A tibble: 10 x 3
##    provider  DRI_POST DRI_PRE
##    <chr>        <dbl>   <dbl>
##  1 anthropic  -0.0470 -0.0987
##  2 cohere     -0.179  -0.169
##  3 deepseek   -0.192  -0.104
##  4 google     -0.108  -0.485
##  5 meta       -0.108  -0.139
##  6 microsoft   0.120   0.0457
##  7 mistralai  -0.0174 -0.364
##  8 openai     -0.261  -0.110
##  9 qwen        0.0497 -0.151
## 10 xai        -0.160  -0.0430
```

```
## # A tibble: 50 x 3
##    model                      CV_DRI_POST CV_DRI_PRE
##    <fct>                            <dbl>      <dbl>
##  1 claude-3-5-haiku-20241022         178.       283.
##  2 claude-3-5-sonnet-20241022        192.       274.
##  3 claude-3-7-sonnet-20250219        219.       305.
##  4 claude-3-haiku-20240307           172.       191.
##  5 claude-3-opus-20240229            224.       338.
##  6 command                           172.       227.
##  7 command-r-08-2024                 217.       296.
##  8 command-r-plus-08-2024            411.       706.
##  9 command-r7b-12-2024               105.       112.
## 10 deepseek-chat                     245.       288.
## 11 deepseek-reasoner                 496.      3972.
## 12 deepseek-v2                       206.     45270.
## 13 gemini-1.5-flash                  332.       443.
## 14 gemini-1.5-flash-8b               311.       426.
## 15 gemini-1.5-pro                    405.       678.
## 16 gemini-2.0-flash                  173.       174.
## 17 gemma                             205.       143.
## 18 gemma2:27b                        195.       226.
## 19 gemma3:12b                        466.       397.
## 20 gpt-3.5-turbo                     174.       291.
## 21 gpt-4                             397.       454.
## 22 gpt-4-turbo                       161.       261.
## 23 gpt-4o                            505.       669.
## 24 gpt-4o-mini                       166.       350.
## 25 grok-2-1212                       277.       375.
## 26 llama3:70b                        210.       375.
## 27 llama3.1:405B-turbo               236.       295.
## 28 llama3.2                          428.       269.
## 29 llama3.3:70b                      203.       353.
## 30 ministral-3b-latest               229.       332.
## 31 ministral-8b-latest               339.       333.
## 32 mistral-large-latest              230.       336.
## 33 mistral-small-latest              286.       431.
## 34 o1                               6442.       595.
## 35 o1-mini                           208.       356.
## 36 o3-mini                           205.       213.
## 37 open-mistral-7b                   169.       203.
## 38 open-mistral-nemo                 232.       354.
## 39 open-mixtral-8x22b                324.       330.
## 40 open-mixtral-8x7b                 167.       218.
## 41 phi3.5                            750.       985.
## 42 phi4                              256.       292.
## 43 qwen-max                          295.       373.
## 44 qwen-plus                         163.       281.
## 45 qwen-turbo                        244.       339.
## 46 qwen1.5-110b-chat            26871137.       378.
## # i 4 more rows
```

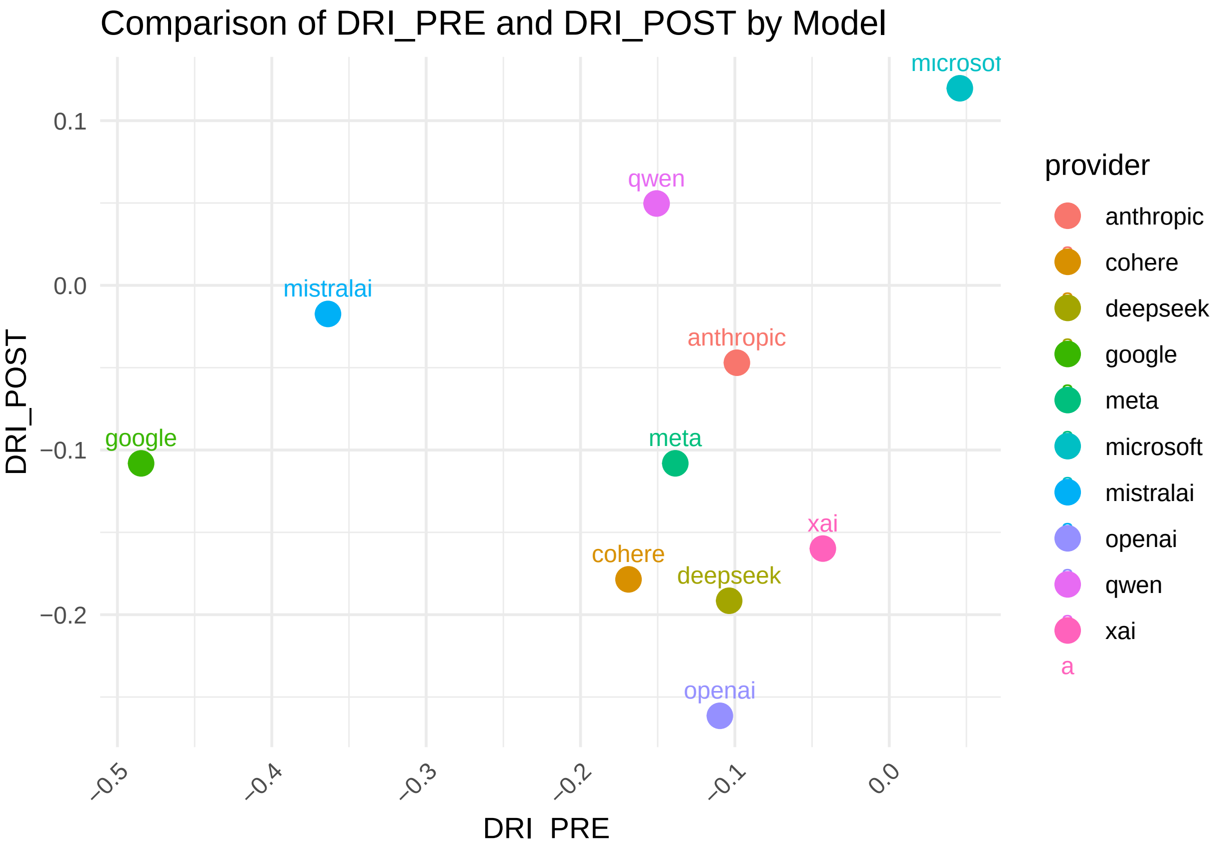
```
## # A tibble: 50 x 3
##    model                      CV_DRI_POST CV_DRI_PRE
##    <fct>                            <dbl>      <dbl>
##  1 claude-3-5-haiku-20241022   0.264        0.163
##  2 claude-3-5-sonnet-20241022  0.298        0.200
##  3 claude-3-7-sonnet-20250219  0.269        0.184
##  4 claude-3-haiku-20240307     0.261        0.226
##  5 claude-3-opus-20240229      0.254        0.164
##  6 command                     0.197        0.146
##  7 command-r-08-2024           0.228        0.157
##  8 command-r-plus-08-2024     NA           NA
##  9 command-r7b-12-2024         0.323        0.284
## 10 deepseek-chat               0.247        0.194
## 11 deepseek-reasoner           0.0966      -0.0126
## 12 deepseek-v2                 0.151       -0.000736
## 13 gemini-1.5-flash            0.172        0.123
## 14 gemini-1.5-flash-8b         0.167        0.118
## 15 gemini-1.5-pro              0.146        0.0846
## 16 gemini-2.0-flash            0.295        0.273
## 17 gemma                       0.211        0.228
## 18 gemma2:27b                  0.289        0.228
## 19 gemma3:12b                 NA           NA
## 20 gpt-3.5-turbo               0.185        0.104
## 21 gpt-4                       0.145        0.119
## 22 gpt-4-turbo                 0.286        0.170
## 23 gpt-4o                      0.0992       0.0729
## 24 gpt-4o-mini                 0.258        0.114
## 25 grok-2-1212                 0.152        0.118
## 26 llama3:70b                  0.233        0.123
## 27 llama3.1:405B-turbo         0.215        0.161
## 28 llama3.2                    0.0811       0.132
## 29 llama3.3:70b                0.245        0.144
## 30 ministral-3b-latest         0.190        0.128
## 31 ministral-8b-latest         0.152        0.136
## 32 mistral-large-latest        0.256        0.168
## 33 mistral-small-latest        0.176        0.121
## 34 o1                          0.00598      0.0556
## 35 o1-mini                     0.216        0.125
## 36 o3-mini                     0.234        0.206
## 37 open-mistral-7b             0.228        0.198
## 38 open-mistral-nemo           0.236        0.151
## 39 open-mixtral-8x22b          0.179        0.170
## 40 open-mixtral-8x7b           0.281        0.202
## 41 phi3.5                     -0.0730      -0.0508
## 42 phi4                        0.209        0.171
## 43 qwen-max                   0.186        0.142
## 44 qwen-plus                  0.304        0.176
## 45 qwen-turbo                 0.220        0.148
## 46 qwen1.5-110b-chat          0.00000237   0.129
## # i 4 more rows
```

```
## # A tibble: 50 x 3
##    model                      CV_DRI_POST CV_DRI_PRE
##    <fct>                            <dbl>      <dbl>
##  1 claude-3-5-haiku-20241022       0.221      0.213
##  2 claude-3-5-sonnet-20241022      0.329      0.300
##  3 claude-3-7-sonnet-20250219      0.345      0.316
##  4 claude-3-haiku-20240307         0.200      0.187
##  5 claude-3-opus-20240229          0.323      0.309
##  6 command                         0.116      0.109
##  7 command-r-08-2024               0.245      0.216
##  8 command-r-plus-08-2024         NA         NA
##  9 command-r7b-12-2024             0.115      0.101
## 10 deepseek-chat                   0.367      0.312
## 11 deepseek-reasoner               0.230      0.252
## 12 deepseek-v2                     0.0972     0.111
## 13 gemini-1.5-flash                0.325      0.296
## 14 gemini-1.5-flash-8b             0.269      0.254
## 15 gemini-1.5-pro                  0.348      0.329
## 16 gemini-2.0-flash                0.260      0.225
## 17 gemma                           0.187      0.106
## 18 gemma2:27b                      0.318      0.265
## 19 gemma3:12b                     NA         NA
## 20 gpt-3.5-turbo                   0.104      0.0924
## 21 gpt-4                           0.333      0.290
## 22 gpt-4-turbo                     0.213      0.197
## 23 gpt-4o                          0.251      0.238
## 24 gpt-4o-mini                     0.184      0.158
## 25 grok-2-1212                     0.178      0.196
## 26 llama3:70b                      0.240      0.211
## 27 llama3.1:405B-turbo             0.259      0.226
## 28 llama3.2                        0.121      0.126
## 29 llama3.3:70b                    0.247      0.260
## 30 ministral-3b-latest             0.188      0.182
## 31 ministral-8b-latest             0.265      0.205
## 32 mistral-large-latest            0.348      0.318
## 33 mistral-small-latest            0.253      0.273
## 34 o1                              0.148      0.109
## 35 o1-mini                         0.202      0.198
## 36 o3-mini                         0.229      0.193
## 37 open-mistral-7b                 0.148      0.162
## 38 open-mistral-nemo               0.302      0.285
## 39 open-mixtral-8x22b              0.337      0.314
## 40 open-mixtral-8x7b               0.219      0.195
## 41 phi3.5                          0.300      0.251
## 42 phi4                            0.285      0.251
## 43 qwen-max                        0.301      0.280
## 44 qwen-plus                       0.245      0.243
## 45 qwen-turbo                      0.288      0.253
## 46 qwen1.5-110b-chat                0.404      0.238
## # i 4 more rows

##
##  Pearson's product-moment correlation
##
## data:  DATA_LLM$DRIPostV2 and as.numeric(DATA_LLM$human_only_DRIPost_meanV2)
## t = 1.6112, df = 1245, p-value = 0.1074
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.009921605  0.100873185
## sample estimates:
##        cor
## 0.04561607

## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
## Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```

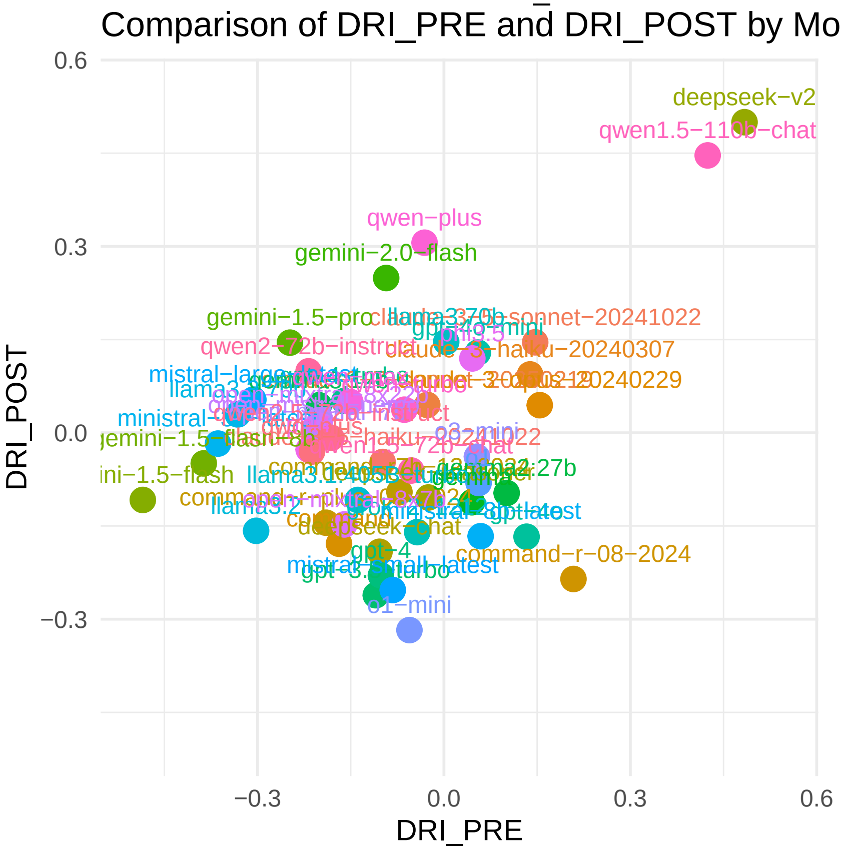


Density Plot of DRI_PRE by Model

Density Plot of DRI_POST by Mode

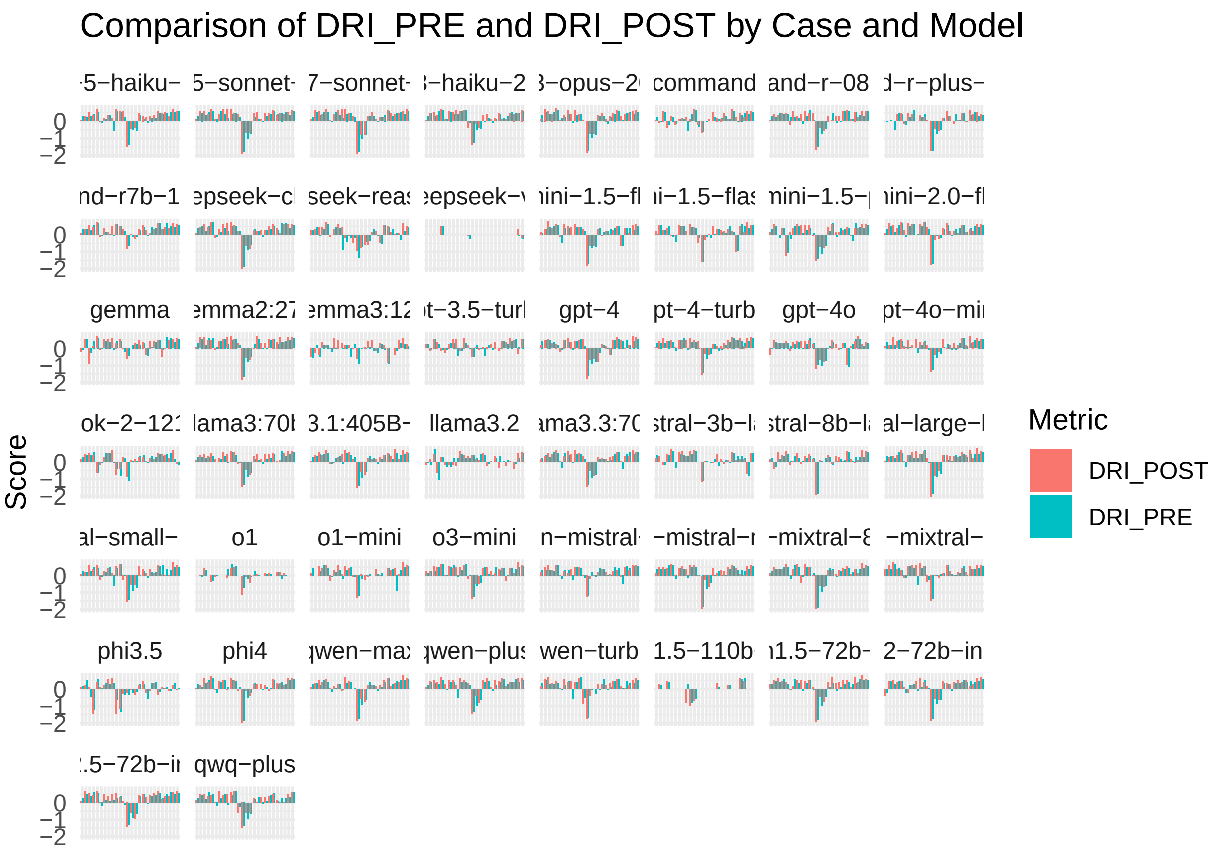


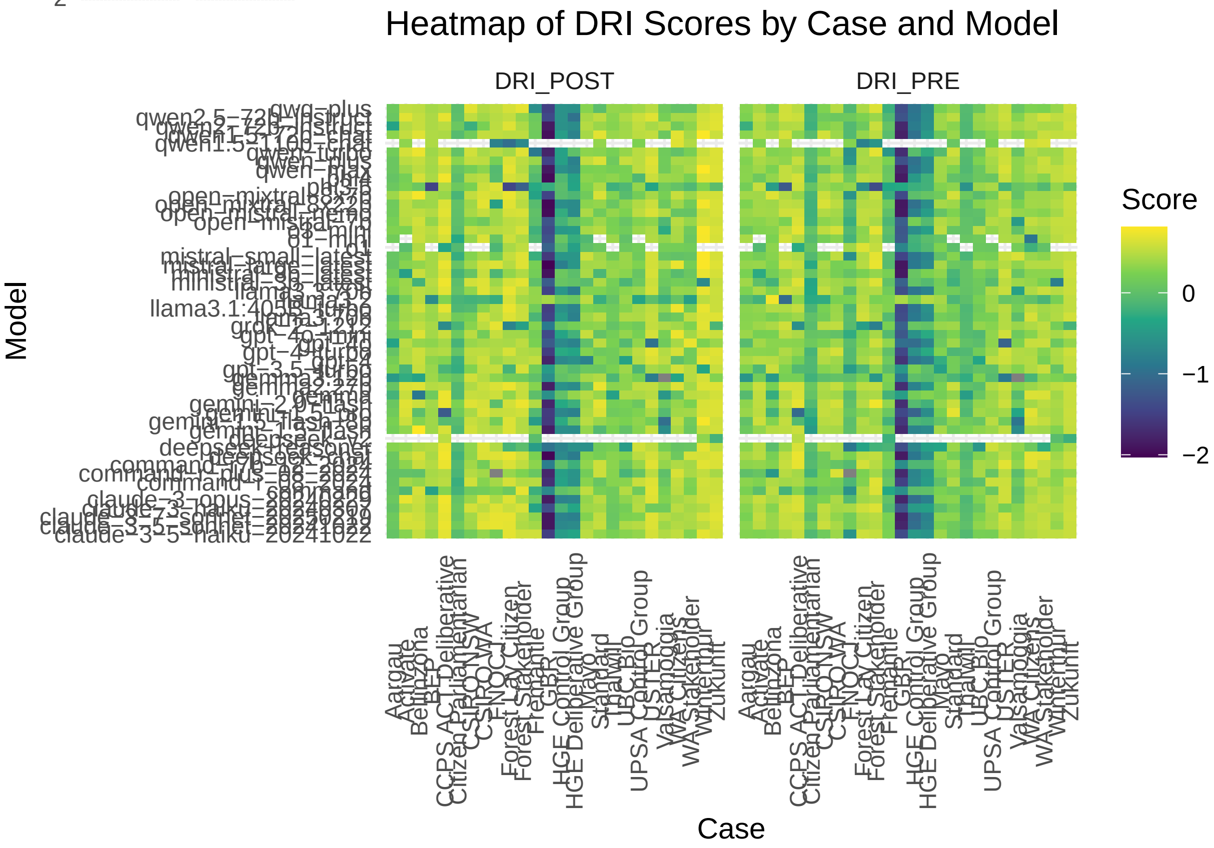
Comparison of DRI_PRE and DRI_POST by Model

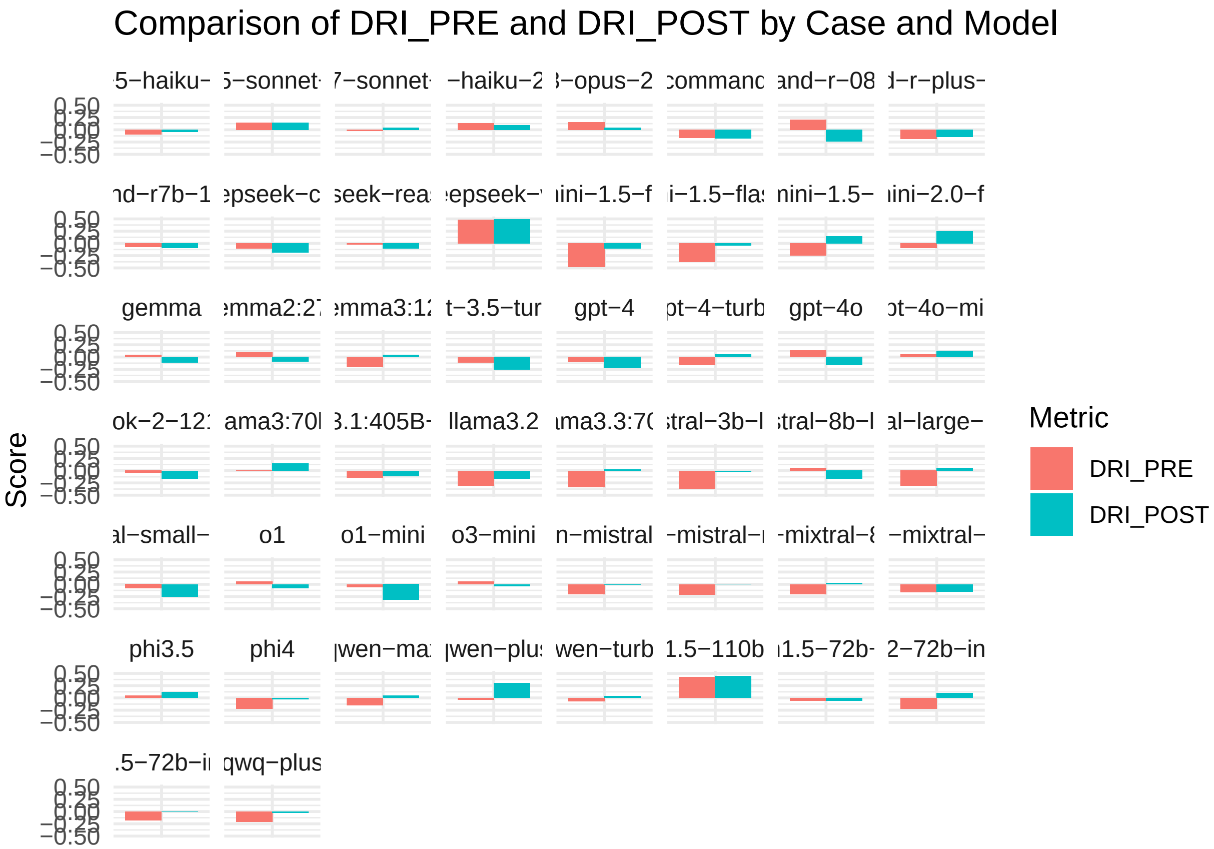


Comparison of DRI_PRE and DRI_POST by Model

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

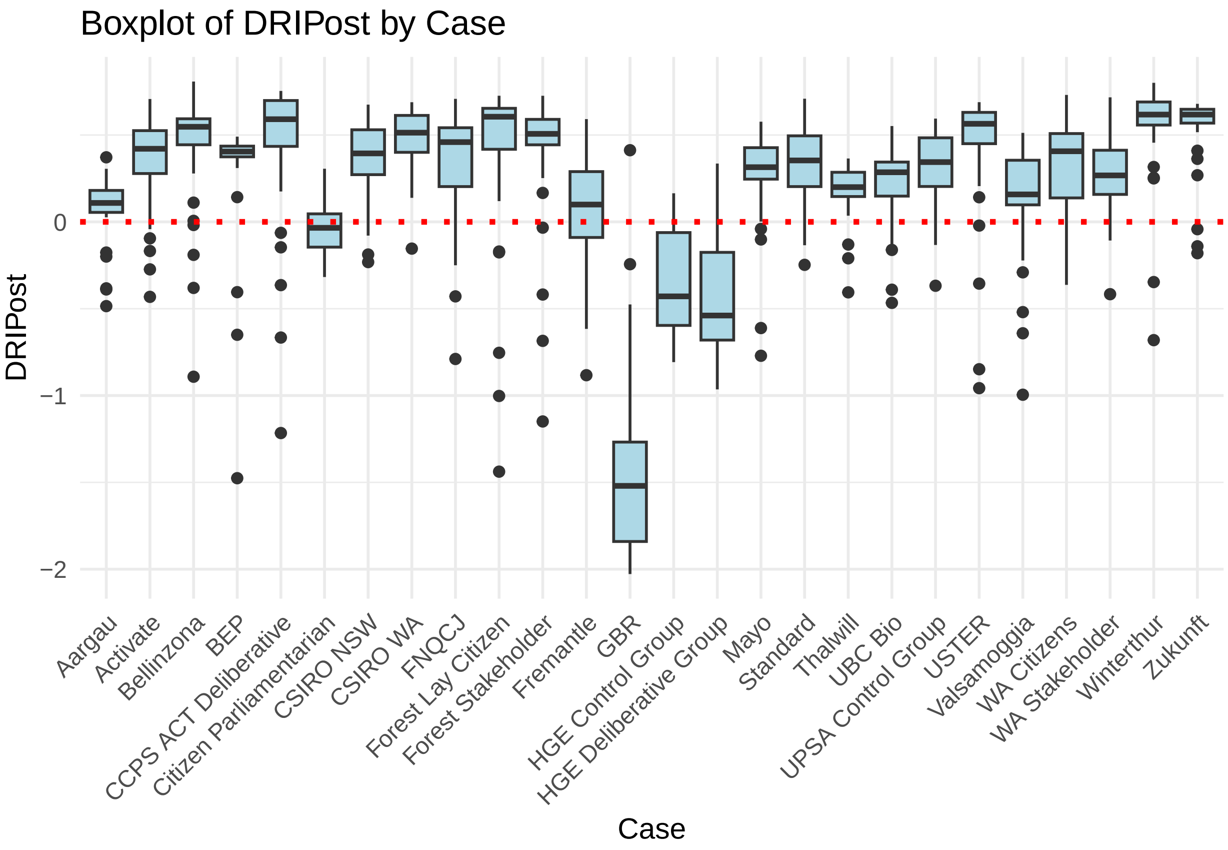

Comparison of DRI_PRE and DRI_POST by Case and Model
-5−haiku−
5−sonnet-
7−sonnet-
-haiku−2
-opus−2
command
and−r−08
d−r−plus−
nd−r7b−1
epseek−c
seek−rea
eepseek−
nini−1.5−fl
i−1.5−fla
mini−1.5−
nini−2.0−fl
gemma
emma2:27
emma3:12
t−3.5−tur
gpt−4
pt−4−turb
gpt−4o
pt−4o−mi
ok−2−121
lama3:70
3.1:405B−
llama3.2
ama3.3:70
stral−3b−l
stral−8b−l
al−large−l
al−small−l
o1
o1−mini
o3−mini
n−mistral
−mistral−
−mixtral−8
−mixtral−
phi3.5
phi4
qwen−ma
qwen−plus
wen−turb
1.5−110b
1.5−72b−
2−72b−in
.5−72b−i
qwq−plus
Score
0
−1
−2
Metric
DRI_POST
DRI_PRE


Heatmap of DRI Scores by Case and Model
DRI_POST
DRI_PRE
Model
qwq−plus
qwen2.5−72b−instruct
qwen2−72b−instruct
qwen1.5−110b−chat
qwen−turbo
qwen−plus
qwen−max
phi4
phi3.5
open−mixtral−8x22b
open−mixtral−8x7b
open−mistral−7b
o3−mini
o1−mini
o1
mistral−small−latest
mistral−large−latest
ministral−8b−latest
llama3.3:70b
llama3.2
llama3.1:405B
llama3:70b
grok−2−1212
gpt−4o
gpt−4−turbo
gpt−4
gpt−3.5−turbo
gemma3:12b
gemma2:27b
gemma
gemini−2.0−flash
gemini−1.5−flash
deepseek−reasoner
deepseek−chat
command−r7b
command−r−plus
command−r−08−2024
command
claude−3−opus−20240229
claude−3−haiku−20240307
claude−3−7−sonnet−20250219
claude−3−5−sonnet−20241022
claude−3−5−haiku−20241022
Case
Aargau
Activate
Bellinzona
BEP
CCPS ACT Deliberative
Citizen Parliamentarian
CSIRO NSW
CSIRO WA
FNQCJ
Forest Lay Citizen
Forest Stakeholder
Fremantle
GBR
HGE Control Group
HGE Deliberative Group
Mayo
Standard
Uster
UBC Bio
UPSA Control Group
UPSA ER
Valsamoggia
WA Citizens
WA Stakeholder
Winterthur
Zukunft
Score
0
−1
−2

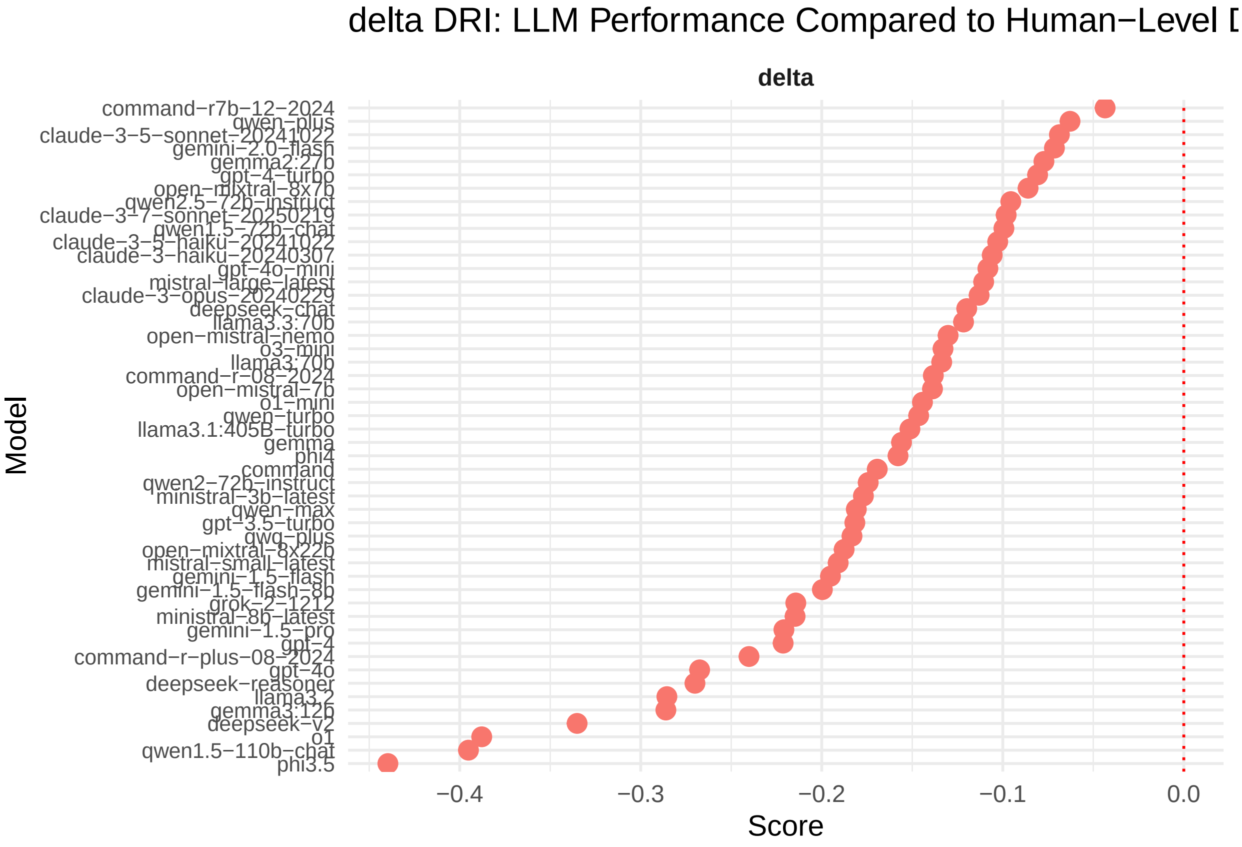


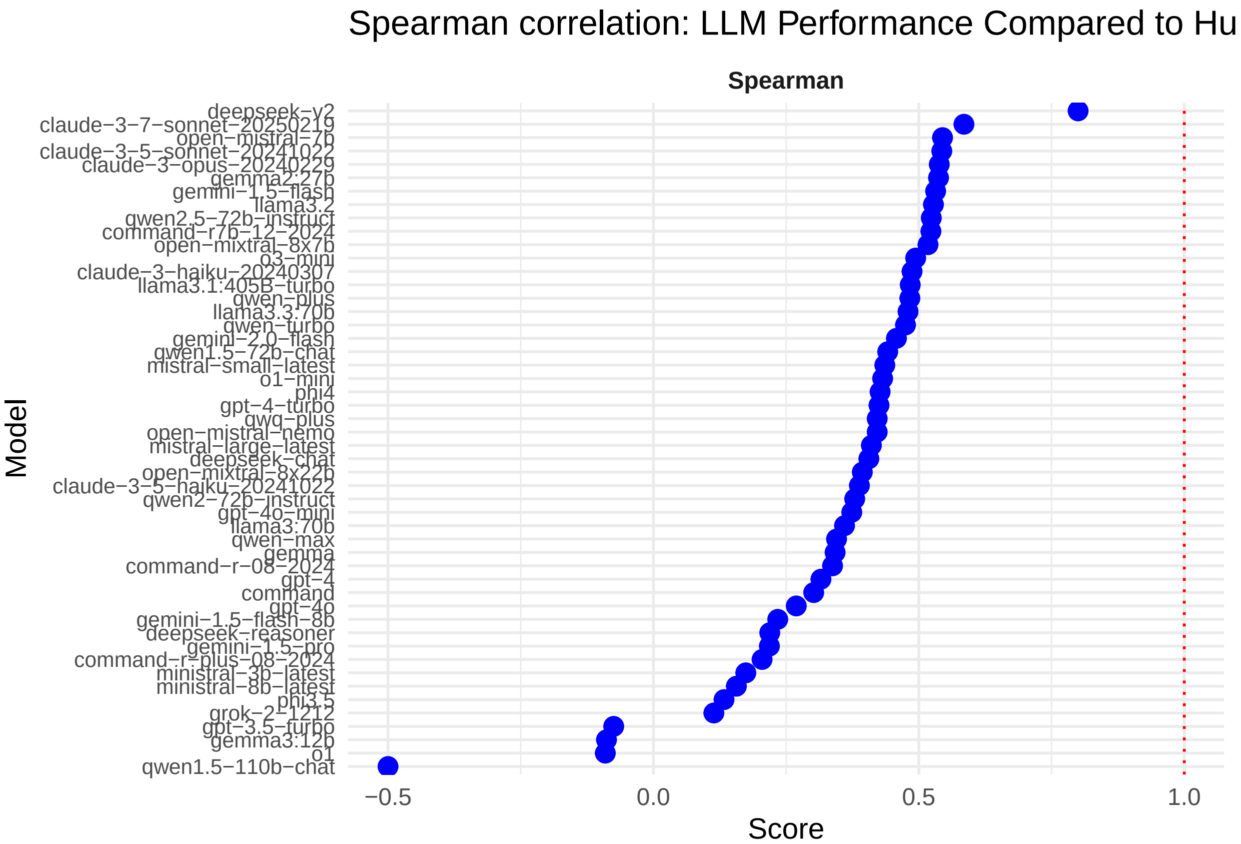
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

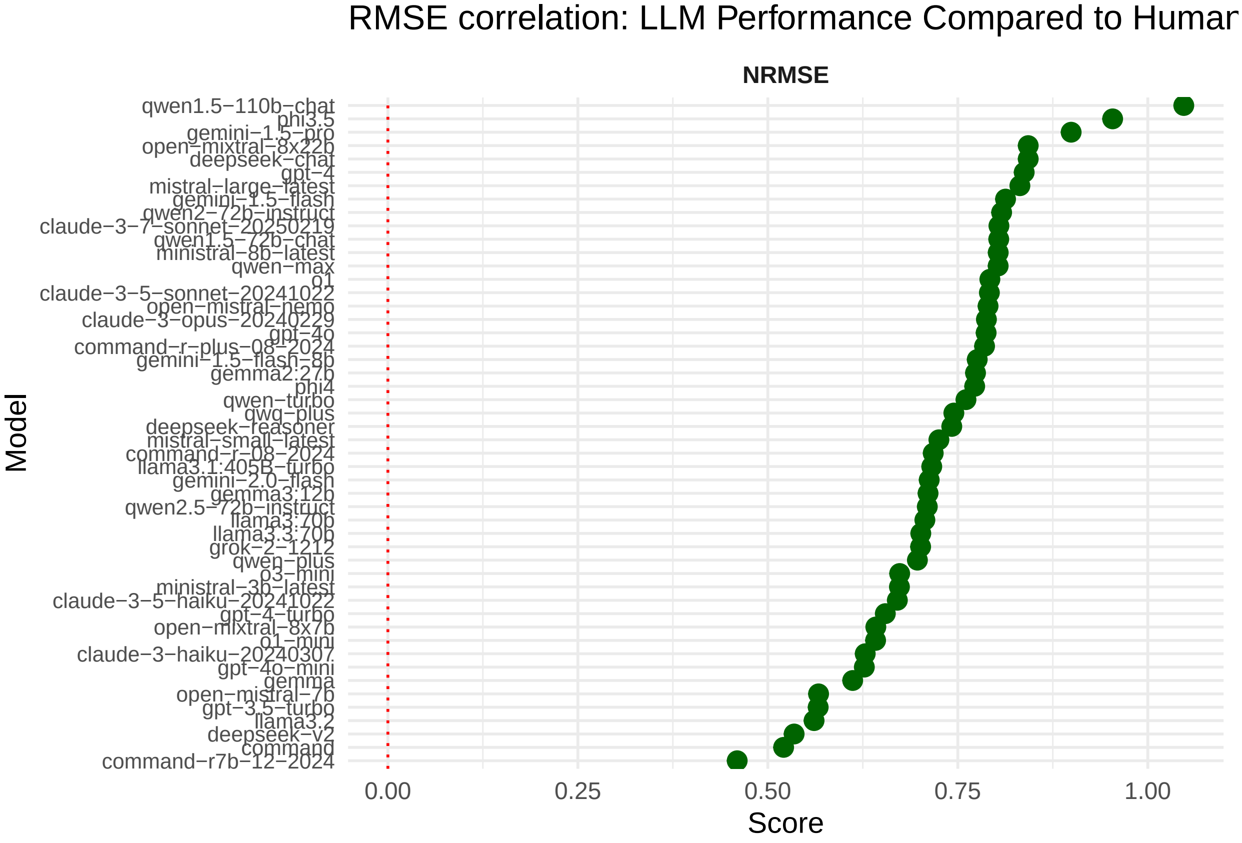
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
## # A tibble: 50 x 8
##    model                        MAE  RMSE  MAPE  NMAE NRMSE Spearman   delta
##    <fct>                      <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>
##  1 command-r7b-12-2024        0.198 0.342  65.9 0.266 0.460   0.523  -0.0435
##  2 command                    0.277 0.387  90.0 0.373 0.521   0.302  -0.169
##  3 deepseek-v2                0.335 0.398  76.8 0.451 0.535   0.8    -0.335
##  4 llama3.2                   0.335 0.417 116.  0.450 0.561   0.528  -0.286
##  5 gpt-3.5-turbo              0.314 0.421 104.  0.422 0.566  -0.0749 -0.182
##  6 open-mistral-7b            0.222 0.421  68.2 0.298 0.567   0.545  -0.139
##  7 gemma                      0.278 0.455  89.9 0.374 0.612   0.342  -0.156
##  8 gpt-4o-mini                0.250 0.466  82.7 0.336 0.627   0.374  -0.108
##  9 claude-3-haiku-20240307    0.265 0.467  86.6 0.356 0.628   0.487  -0.106
## 10 o1-mini                    0.277 0.477 112.  0.372 0.642   0.432  -0.144
## 11 open-mixtral-8x7b          0.271 0.478  91.0 0.365 0.642   0.517  -0.0860
## 12 gpt-4-turbo                0.258 0.487  79.3 0.346 0.655   0.425  -0.0808
## 13 claude-3-5-haiku-20241022  0.270 0.499  75.5 0.363 0.670   0.388  -0.103
## 14 ministral-3b-latest        0.312 0.501  86.6 0.420 0.673   0.174  -0.177
## 15 o3-mini                    0.282 0.501  85.0 0.379 0.673   0.494  -0.133
## 16 qwen-plus                  0.291 0.518 112.  0.392 0.697   0.483  -0.0629
## 17 grok-2-1212                0.303 0.521  79.2 0.408 0.701   0.114  -0.214
## 18 llama3.3:70b               0.273 0.522  81.6 0.367 0.701   0.480  -0.122
## 19 llama3:70b                 0.308 0.526 103.  0.414 0.707   0.36   -0.134
## 20 qwen2.5-72b-instruct       0.268 0.528  84.8 0.360 0.710   0.523  -0.0956
## 21 gemma3:12b                 0.369 0.529 101.  0.497 0.711  -0.0885 -0.286
## 22 gemini-2.0-flash           0.280 0.530  99.4 0.377 0.712   0.458  -0.0715
## 23 llama3.1:405B-turbo        0.278 0.532  81.5 0.373 0.716   0.484  -0.151
## 24 command-r-08-2024          0.289 0.534  99.4 0.389 0.718   0.337  -0.138
## 25 mistral-small-latest       0.318 0.539 107.  0.428 0.725   0.436  -0.191
## 26 deepseek-reasoner          0.397 0.552 126.  0.534 0.742   0.219  -0.270
## 27 qwq-plus                   0.315 0.554  94.7 0.424 0.745   0.422  -0.183
## 28 qwen-turbo                 0.287 0.566  88.6 0.385 0.761   0.475  -0.146
## 29 phi4                       0.284 0.574  81.0 0.382 0.772   0.427  -0.158
## 30 gemma2:27b                 0.300 0.575  92.1 0.403 0.773   0.537  -0.0773
## 31 gemini-1.5-flash-8b        0.338 0.577  97.2 0.454 0.776   0.234  -0.200
## 32 command-r-plus-08-2024     0.338 0.584 105.  0.454 0.785   0.205  -0.240
## 33 gpt-4o                     0.361 0.585 120.  0.486 0.787   0.269  -0.267
## 34 claude-3-opus-20240229     0.277 0.586  84.2 0.373 0.788   0.538  -0.113
## 35 open-mistral-nemo          0.301 0.587  91.1 0.404 0.790   0.422  -0.130
## 36 claude-3-5-sonnet-20241022 0.292 0.589  96.4 0.393 0.792   0.543  -0.0687
## 37 o1                         0.401 0.589  91.8 0.539 0.792  -0.0907 -0.388
## 38 qwen-max                   0.323 0.597  95.0 0.434 0.803   0.345  -0.181
## 39 ministral-8b-latest        0.332 0.597  97.8 0.446 0.803   0.156  -0.215
## 40 qwen1.5-72b-chat           0.312 0.598  95.6 0.419 0.804   0.441  -0.0994
## 41 claude-3-7-sonnet-20250219 0.279 0.598  80.7 0.375 0.804   0.585  -0.0981
## 42 qwen2-72b-instruct         0.332 0.601 115.  0.447 0.808   0.379  -0.174
## 43 gemini-1.5-flash           0.313 0.604  94.7 0.421 0.813   0.532  -0.195
## 44 mistral-large-latest       0.318 0.618  95.8 0.427 0.832   0.411  -0.111
## 45 gpt-4                      0.372 0.623 122.  0.500 0.837   0.316  -0.221
## 46 deepseek-chat              0.328 0.627 107.  0.441 0.843   0.406  -0.120
## # i 4 more rows
```

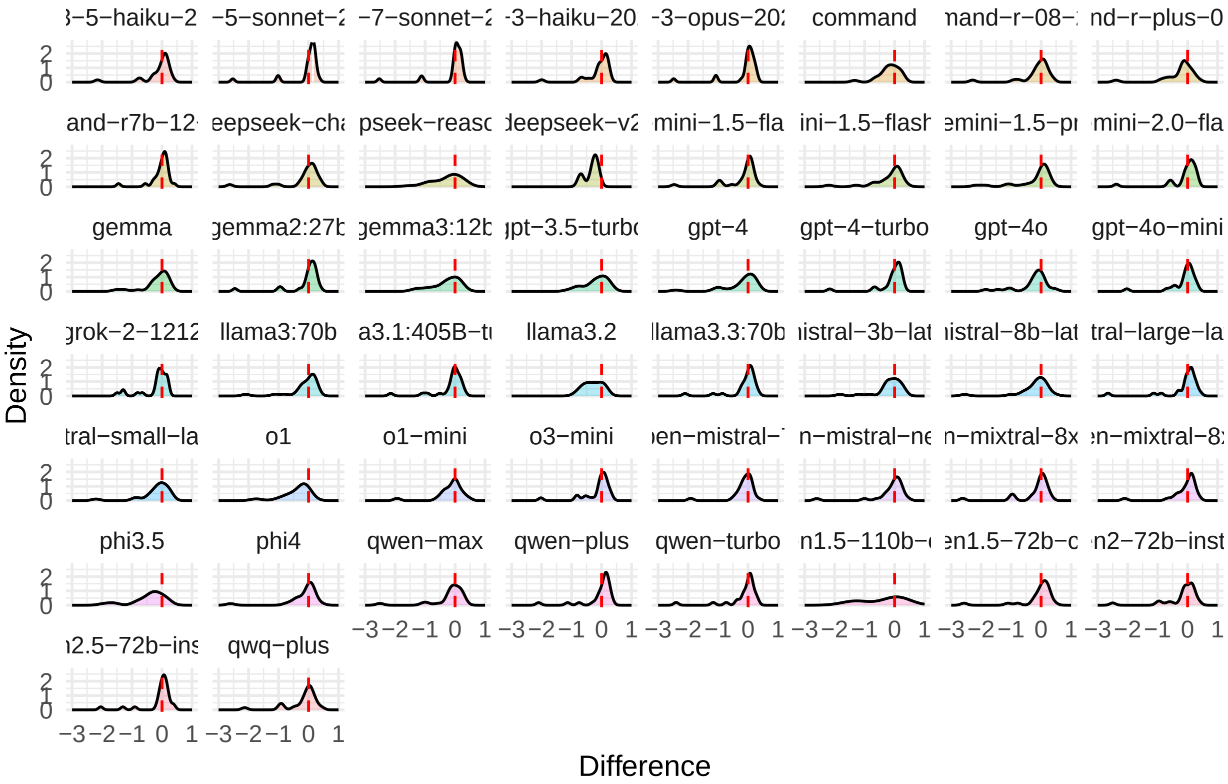






```
## [1] -0.006868248  0.185752032  0.084988051  0.134099925  0.084585033
## [6] -0.138415676

## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```

Distribution of Differences (DRIPost − DRI_POST_MEAN_H) per Model



```
## [1] 0.4189944

## # A tibble: 50 x 8
##    model                       MAE  RMSE  MAPE  NMAE NRMSE Spearman   delta
##    <fct>                     <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>
##  1 command-r7b-12-2024       0.201 0.353  49.5 0.481 0.843  0.446   -0.0619
##  2 deepseek-v2               0.335 0.398  76.8 0.800 0.949  0.8     -0.335
##  3 command                   0.293 0.407  73.3 0.698 0.971  0.199   -0.193
##  4 llama3.2                  0.353 0.436 102.  0.842 1.04   0.520   -0.309
##  5 gpt-3.5-turbo             0.325 0.439  75.9 0.776 1.05  -0.204   -0.216
##  6 open-mistral-7b           0.232 0.443  57.4 0.553 1.06   0.536   -0.176
##  7 gemma                     0.284 0.475  71.0 0.679 1.13   0.208   -0.146
##  8 gpt-4o-mini               0.262 0.491  62.1 0.625 1.17   0.279   -0.141
##  9 claude-3-haiku-20240307   0.281 0.494  71.4 0.670 1.18   0.405   -0.130
## 10 o1-mini                   0.280 0.497  70.0 0.668 1.19   0.411   -0.177
## 11 open-mixtral-8x7b         0.276 0.499  68.2 0.658 1.19   0.493   -0.119
## 12 gpt-4-turbo               0.272 0.514  66.2 0.650 1.23   0.312   -0.103
## 13 claude-3-5-haiku-20241022 0.287 0.525  68.0 0.685 1.25   0.276   -0.126
## 14 o3-mini                   0.298 0.527  77.3 0.711 1.26   0.468   -0.171
## 15 ministral-3b-latest       0.334 0.528  78.9 0.797 1.26   0.137   -0.219
## 16 qwen-plus                 0.302 0.544  77.4 0.722 1.30   0.404   -0.0944
## 17 gemma3:12b                0.386 0.547  84.0 0.922 1.31  -0.331   -0.299
## 18 llama3.3:70b              0.292 0.552  72.0 0.697 1.32   0.394   -0.154
## 19 grok-2-1212               0.328 0.553  68.2 0.784 1.32   0.00593 -0.246
## 20 llama3:70b                 0.327 0.555  84.4 0.780 1.32   0.267   -0.169
## 21 qwen2.5-72b-instruct       0.282 0.555  73.8 0.673 1.32   0.454   -0.125
## 22 gemini-2.0-flash           0.289 0.556  67.0 0.690 1.33   0.422   -0.104
## 23 command-r-08-2024          0.297 0.560  71.7 0.710 1.34   0.208   -0.160
## 24 llama3.1:405B-turbo        0.296 0.563  71.3 0.707 1.34   0.409   -0.175
## 25 mistral-small-latest       0.336 0.569  83.9 0.802 1.36   0.371   -0.215
## 26 qwq-plus                   0.336 0.582  86.6 0.802 1.39   0.406   -0.226
## 27 deepseek-reasoner          0.427 0.582 119.  1.02  1.39   0.211   -0.322
## 28 qwen-turbo                 0.305 0.598  77.3 0.729 1.43   0.440   -0.184
## 29 phi4                       0.298 0.604  69.5 0.712 1.44   0.430   -0.194
## 30 gemma2:27b                 0.316 0.605  78.7 0.754 1.44   0.481   -0.0992
## 31 o1                         0.423 0.607  91.2 1.01  1.45  -0.171   -0.410
## 32 gemini-1.5-flash-8b        0.361 0.608  89.7 0.862 1.45   0.218   -0.246
## 33 gpt-4o                     0.366 0.608  91.4 0.872 1.45   0.142   -0.288
## 34 command-r-plus-08-2024     0.356 0.616  86.7 0.850 1.47   0.120   -0.281
## 35 open-mistral-nemo          0.318 0.618  78.7 0.759 1.47   0.369   -0.169
## 36 claude-3-opus-20240229     0.298 0.620  74.5 0.711 1.48   0.472   -0.141
## 37 claude-3-5-sonnet-20241022 0.311 0.623  78.9 0.743 1.49   0.467   -0.0930
## 38 qwen2-72b-instruct         0.330 0.623  81.8 0.788 1.49   0.291   -0.197
## 39 ministral-8b-latest        0.349 0.627  76.7 0.832 1.50   0.123   -0.258
## 40 qwen1.5-72b-chat           0.325 0.627  81.1 0.775 1.50   0.374   -0.139
## 41 qwen-max                   0.346 0.632  83.7 0.826 1.51   0.272   -0.220
## 42 claude-3-7-sonnet-20250219 0.303 0.634  73.4 0.722 1.51   0.526   -0.123
## 43 gemini-1.5-flash           0.334 0.638  83.6 0.797 1.52   0.517   -0.235
## 44 mistral-large-latest       0.333 0.651  77.1 0.794 1.55   0.311   -0.144
## 45 gpt-4                      0.399 0.659 104.  0.952 1.57   0.201   -0.255
## 46 deepseek-chat              0.346 0.660  86.4 0.825 1.58   0.319   -0.145
## # i 4 more rows
```

```
## # A tibble: 50 x 12
##    model           MAE  RMSE  MAPE  NMAE NRMSE Spearman   delta normalized_NRMSE
##    <fct>         <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>            <dbl>
##  1 deepseek-v2   0.335 0.398  76.8 0.451 0.535   0.8    -0.335             0.873
##  2 llama3.2      0.335 0.417 116.  0.450 0.561   0.528  -0.286             0.828
##  3 open-mistral~ 0.222 0.421  68.2 0.298 0.567   0.545  -0.139             0.818
##  4 command       0.277 0.387  90.0 0.373 0.521   0.302  -0.169             0.896
##  5 command-r7b-~ 0.198 0.342  65.9 0.266 0.460   0.523  -0.0435            1
##  6 gemma         0.278 0.455  89.9 0.374 0.612   0.342  -0.156             0.741
##  7 o1-mini       0.277 0.477 112.  0.372 0.642   0.432  -0.144             0.690
##  8 phi3.5        0.492 0.709 148.  0.662 0.954   0.133  -0.440             0.159
##  9 deepseek-rea~ 0.397 0.552 126.  0.534 0.742   0.219  -0.270             0.519
## 10 mistral-smal~ 0.318 0.539 107.  0.428 0.725   0.436  -0.191             0.548
## 11 claude-3-hai~ 0.265 0.467  86.6 0.356 0.628   0.487  -0.106             0.713
## 12 o3-mini       0.282 0.501  85.0 0.379 0.673   0.494  -0.133             0.636
## 13 o1            0.401 0.589  91.8 0.539 0.792  -0.0907 -0.388             0.434
## 14 gpt-4o        0.361 0.585 120.  0.486 0.787   0.269  -0.267             0.443
## 15 llama3.1:405~ 0.278 0.532  81.5 0.373 0.716   0.484  -0.151             0.564
## 16 open-mixtral~ 0.271 0.478  91.0 0.365 0.642   0.517  -0.0860            0.690
## 17 qwq-plus      0.315 0.554  94.7 0.424 0.745   0.422  -0.183             0.515
## 18 gemini-1.5-f~ 0.313 0.604  94.7 0.421 0.813   0.532  -0.195             0.399
## 19 gpt-4o-mini   0.250 0.466  82.7 0.336 0.627   0.374  -0.108             0.715
## 20 llama3.3:70b  0.273 0.522  81.6 0.367 0.701   0.480  -0.122             0.589
## 21 gemma3:12b    0.369 0.529 101.  0.497 0.711  -0.0885 -0.286             0.573
## 22 qwen-turbo    0.287 0.566  88.6 0.385 0.761   0.475  -0.146             0.488
## 23 gpt-3.5-turbo 0.314 0.421 104.  0.422 0.566  -0.0749 -0.182             0.819
## 24 qwen2.5-72b-~ 0.268 0.528  84.8 0.360 0.710   0.523  -0.0956            0.574
## 25 grok-2-1212   0.303 0.521  79.2 0.408 0.701   0.114  -0.214             0.589
## 26 ministral-3b~ 0.312 0.501  86.6 0.420 0.673   0.174  -0.177             0.637
## 27 command-r-pl~ 0.338 0.584 105.  0.454 0.785   0.205  -0.240             0.446
## 28 claude-3-5-h~ 0.270 0.499  75.5 0.363 0.670   0.388  -0.103             0.641
## 29 gpt-4-turbo   0.258 0.487  79.3 0.346 0.655   0.425  -0.0808            0.668
## 30 phi4          0.284 0.574  81.0 0.382 0.772   0.427  -0.158             0.468
## 31 llama3:70b    0.308 0.526 103.  0.414 0.707   0.36   -0.134             0.580
## 32 command-r-08~ 0.289 0.534  99.4 0.389 0.718   0.337  -0.138             0.561
## 33 gpt-4         0.372 0.623 122.  0.500 0.837   0.316  -0.221             0.357
## 34 gemini-1.5-f~ 0.338 0.577  97.2 0.454 0.776   0.234  -0.200             0.463
## 35 claude-3-opu~ 0.277 0.586  84.2 0.373 0.788   0.538  -0.113             0.442
## 36 qwen2-72b-in~ 0.332 0.601 115.  0.447 0.808   0.379  -0.174             0.408
## 37 qwen-max      0.323 0.597  95.0 0.434 0.803   0.345  -0.181             0.416
## 38 qwen-plus     0.291 0.518 112.  0.392 0.697   0.483  -0.0629            0.596
## 39 open-mixtral~ 0.320 0.627  87.0 0.430 0.843   0.394  -0.188             0.348
## 40 claude-3-7-s~ 0.279 0.598  80.7 0.375 0.804   0.585  -0.0981            0.414
## 41 gemini-2.0-f~ 0.280 0.530  99.4 0.377 0.712   0.458  -0.0715            0.570
## 42 open-mistral~ 0.301 0.587  91.1 0.404 0.790   0.422  -0.130             0.438
## 43 ministral-8b~ 0.332 0.597  97.8 0.446 0.803   0.156  -0.215             0.416
## 44 gemma2:27b    0.300 0.575  92.1 0.403 0.773   0.537  -0.0773            0.466
## 45 claude-3-5-s~ 0.292 0.589  96.4 0.393 0.792   0.543  -0.0687            0.435
## 46 qwen1.5-72b-~ 0.312 0.598  95.6 0.419 0.804   0.441  -0.0994            0.414
## 47 gemini-1.5-p~ 0.375 0.669 110.  0.504 0.899   0.218  -0.221             0.252
## 48 deepseek-chat 0.328 0.627 107.  0.441 0.843   0.406  -0.120             0.348
## 49 mistral-larg~ 0.318 0.618  95.8 0.427 0.832   0.411  -0.111             0.367
## # i 1 more row
## # i 3 more variables: normalized_Spearman <dbl>, normalized_Delta <dbl>,
## #   aggregate_index <dbl>
```

```
## # A tibble: 50 x 4
##    model                          W   p_value effect_size_r
##    <fct>                      <dbl>     <dbl>         <dbl>
##  1 phi3.5                       130 0.0000393         25.5
##  2 o1                            39 0.0000639          9.46
##  3 llama3.2                     174 0.00114           34.1
##  4 gemma3:12b                   171 0.00162           33.5
##  5 command                      223 0.0177            43.7
##  6 command-r-plus-08-2024       221 0.0253            43.3
##  7 grok-2-1212                  231 0.0254            45.3
##  8 deepseek-reasoner            234 0.0289            45.9
##  9 gpt-3.5-turbo                234 0.0289            45.9
## 10 gpt-4o                       235 0.0301            46.1
## 11 ministral-8b-latest          242 0.0402            47.5
## 12 deepseek-v2                    2 0.0571             1
## 13 ministral-3b-latest          256 0.0684            50.2
## 14 open-mistral-7b              258 0.0734            50.6
## 15 gemini-1.5-flash-8b          265 0.0931            52.0
## 16 mistral-small-latest         270 0.109             53.0
## 17 qwen-max                     278 0.139             54.5
## 18 qwen1.5-110b-chat             29 0.170              9.67
## 19 phi4                         292 0.204             57.3
## 20 gpt-4                        294 0.215             57.7
## 21 gpt-4o-mini                  294 0.215             57.7
## 22 o1-mini                      230 0.229             48.0
## 23 qwq-plus                     298 0.236             58.4
## 24 open-mixtral-8x22b           299 0.242             58.6
## 25 gemini-1.5-flash             300 0.248             58.8
## 26 llama3.1:405B-turbo          300 0.248             58.8
## 27 gemini-1.5-pro               302 0.260             59.2
## 28 gemma                        310 0.309             60.8
## 29 qwen-turbo                   312 0.322             61.2
## 30 qwen2-72b-instruct           312 0.322             61.2
## 31 command-r-08-2024            315 0.341             61.8
## 32 o3-mini                      318 0.362             62.4
## 33 llama3:70b                   322 0.389             63.1
## 34 claude-3-haiku-20240307      326 0.417             63.9
## 35 llama3.3:70b                 327 0.424             64.1
## 36 claude-3-5-haiku-20241022    330 0.446             64.7
## 37 open-mistral-nemo            335 0.482             65.7
## 38 open-mixtral-8x7b            345 0.554             67.7
## 39 command-r7b-12-2024          347 0.568             68.1
## 40 gpt-4-turbo                  347 0.568             68.1
## 41 mistral-large-latest         350 0.590             68.6
## 42 deepseek-chat                354 0.618             69.4
## 43 gemini-2.0-flash             355 0.625             69.6
## 44 claude-3-opus-20240229       358 0.645             70.2
## 45 qwen2.5-72b-instruct         368 0.710             72.2
## 46 claude-3-7-sonnet-20250219   374 0.746             73.3
## 47 qwen1.5-72b-chat             376 0.758             73.7
## # i 3 more rows
```

References

Motoki, Fabio, Valdemar Pinho Neto, and Victor Rodrigues. 2024. “More Human Than Human: Measuring ChatGPT Political Bias.” *Public Choice* 198(1): 3–23. doi:10.1007/s11127-023-01097-2.