

# Triage Against the Machine: Can AI Reason Deliberatively?

Francesco Veri, Gustavo Umbelino

2025-04-10

## Large-Language Models (LLMs) Preview

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.  
## Use 'xfun::attr2()' instead.  
## See help("Deprecated")  
  
## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.  
## Use 'xfun::attr2()' instead.  
## See help("Deprecated")
```

Table 1: LLMs

	provider	model	type	version
1	anthropic	claude-3-5-haiku-20241022	NA	2
2	anthropic	claude-3-5-sonnet-20241022	reason	2
3	anthropic	claude-3-7-sonnet-20250219	reason	3
4	anthropic	claude-3-haiku-20240307	NA	1
5	anthropic	claude-3-opus-20240229	NA	1
6	anthropic	claude-3-sonnet-20240229	NA	1
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

	provider	model	type	version
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
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

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

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

	survey	considerations	policies	scale_max	q_method
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
20	zukunft	20	7	7	FALSE

## LLM Data Collection

We collected a total of 30703 valid LLM responses across 20 surveys.

### Cost

We spent a total of 383.71 USD. The cost breakdown per API is below.

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

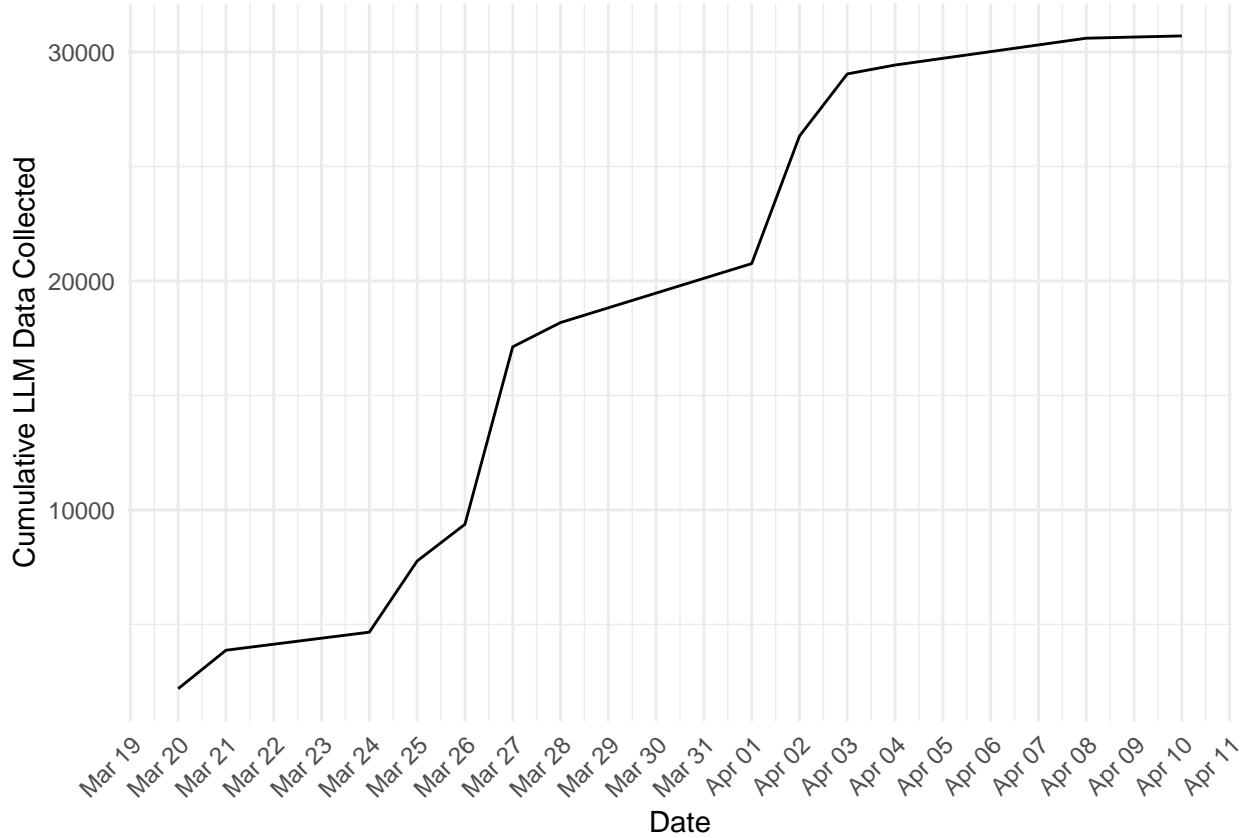
Table 3: Costs by API

api	num_models	credits_paid
OpenAI API	9	225.52
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 Cloud	4	NA
ollama	9	NA

### Time

It took a total of 153 hours<sup>1</sup> across 21 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 Thursday, Apr 10, 2025.

<sup>1</sup>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.

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

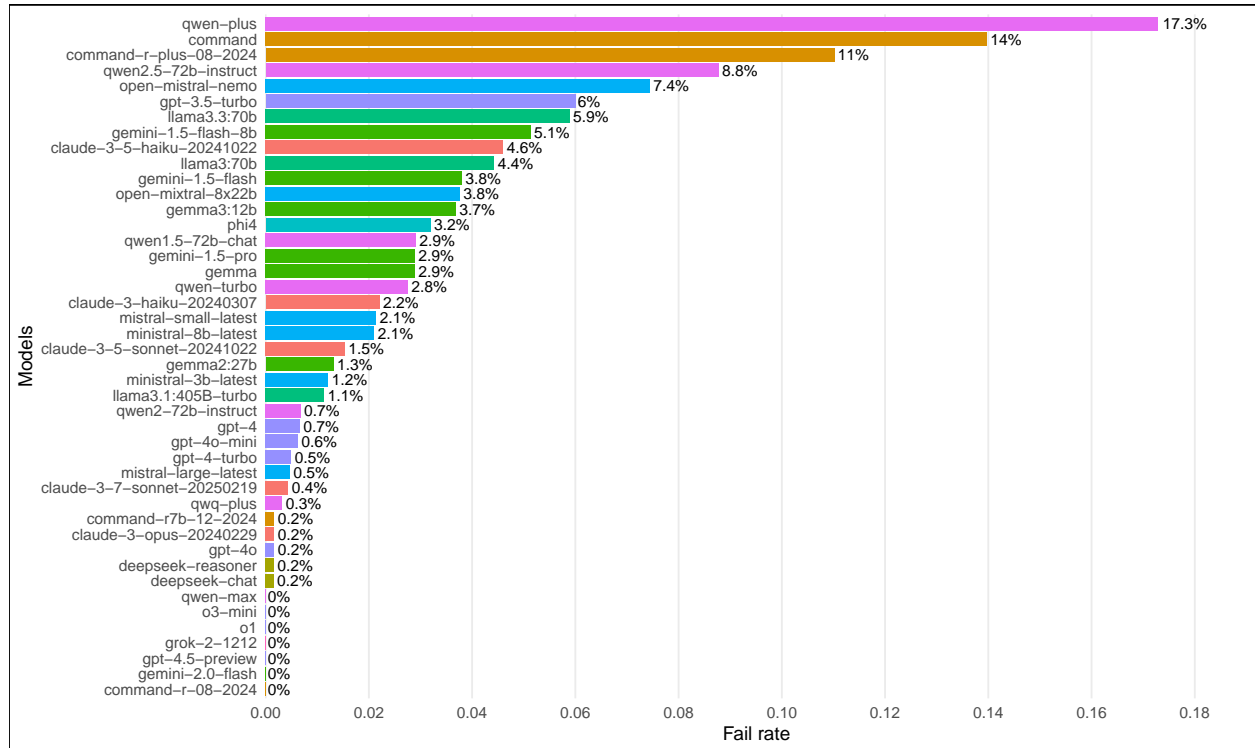
Table 4: Excluded models and reasons

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

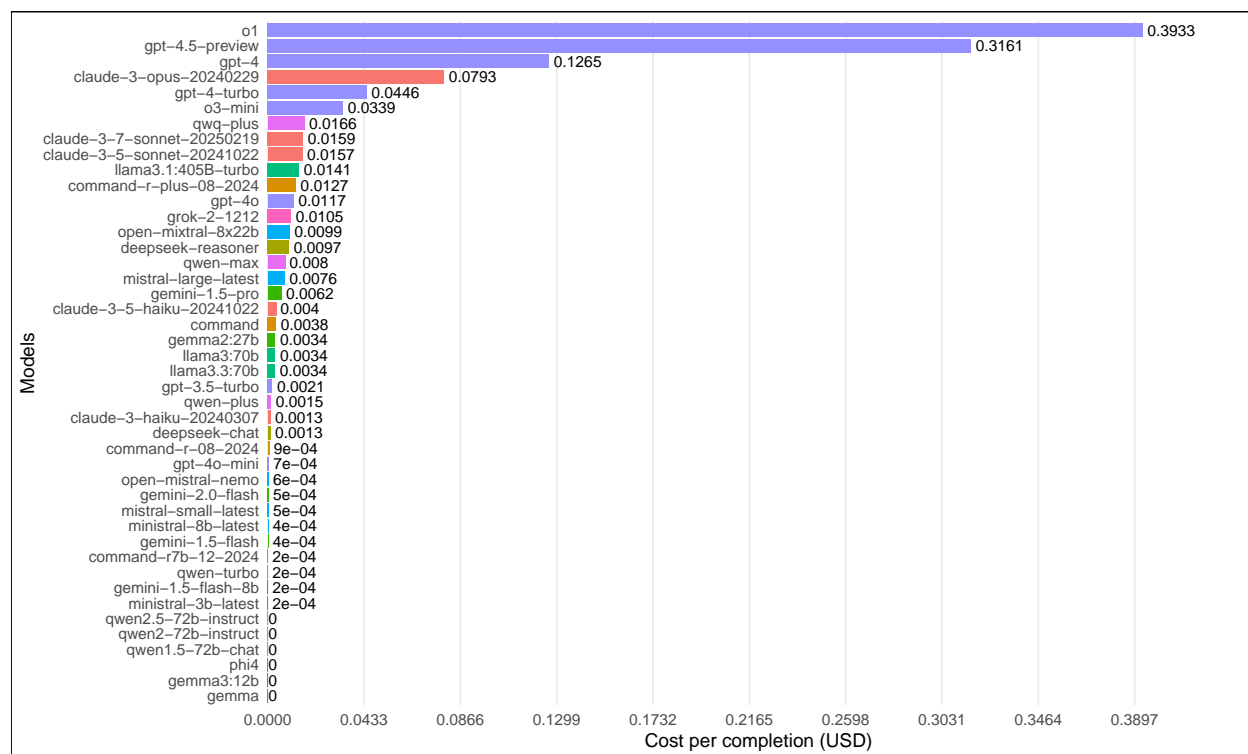
provider	model	reason
mistralai	open-mixtral-8x7b	6% success rate on fremantle only
openai	o1-mini	0% success rate on uppsala_speaks only; responds with “I’m sorry, but I can’t help with that.”
qwen	qwen1.5-110b-chat	has API limit of 10 RPM; too slow

## Execution Summary Plots

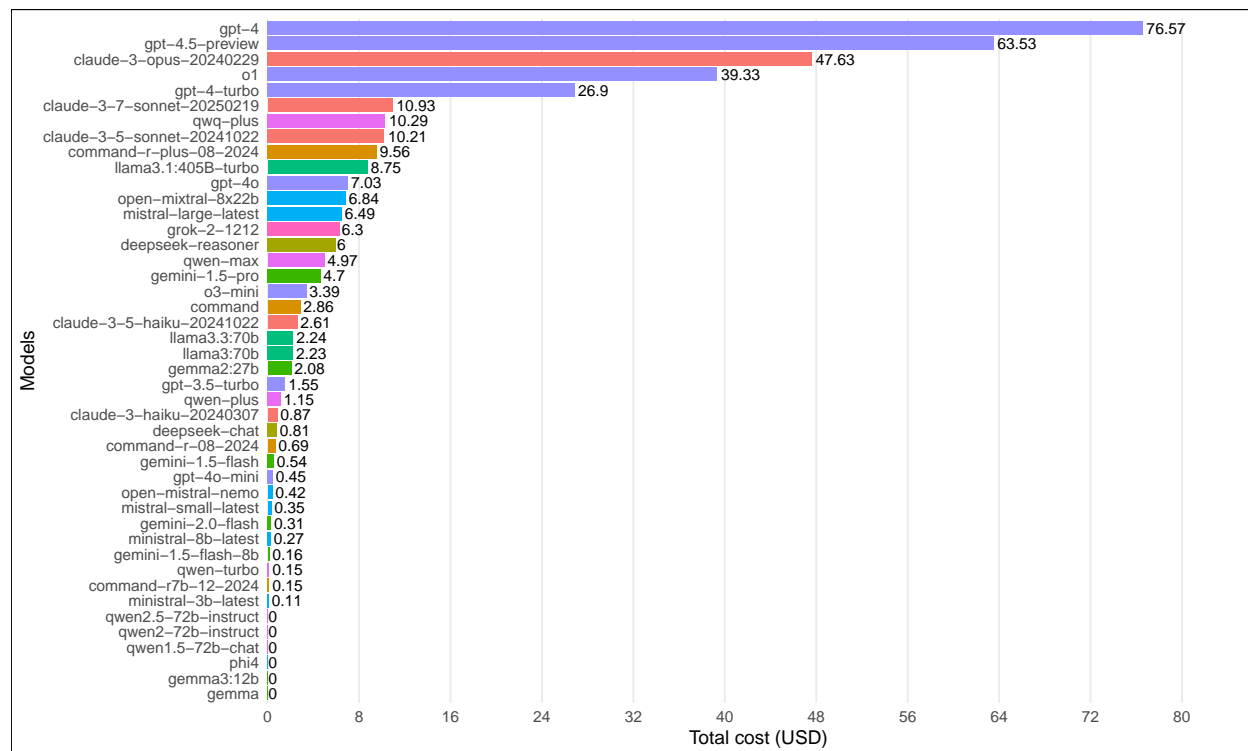
### Fail rate



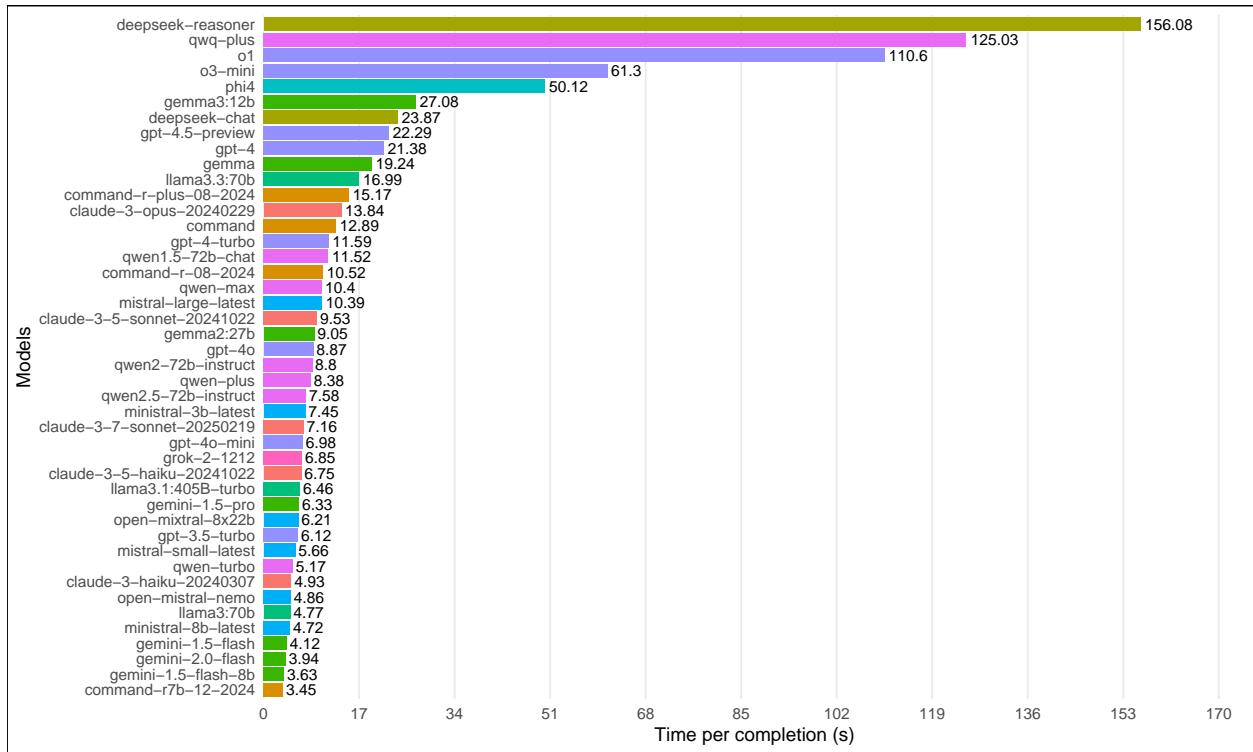
## Cost per completion



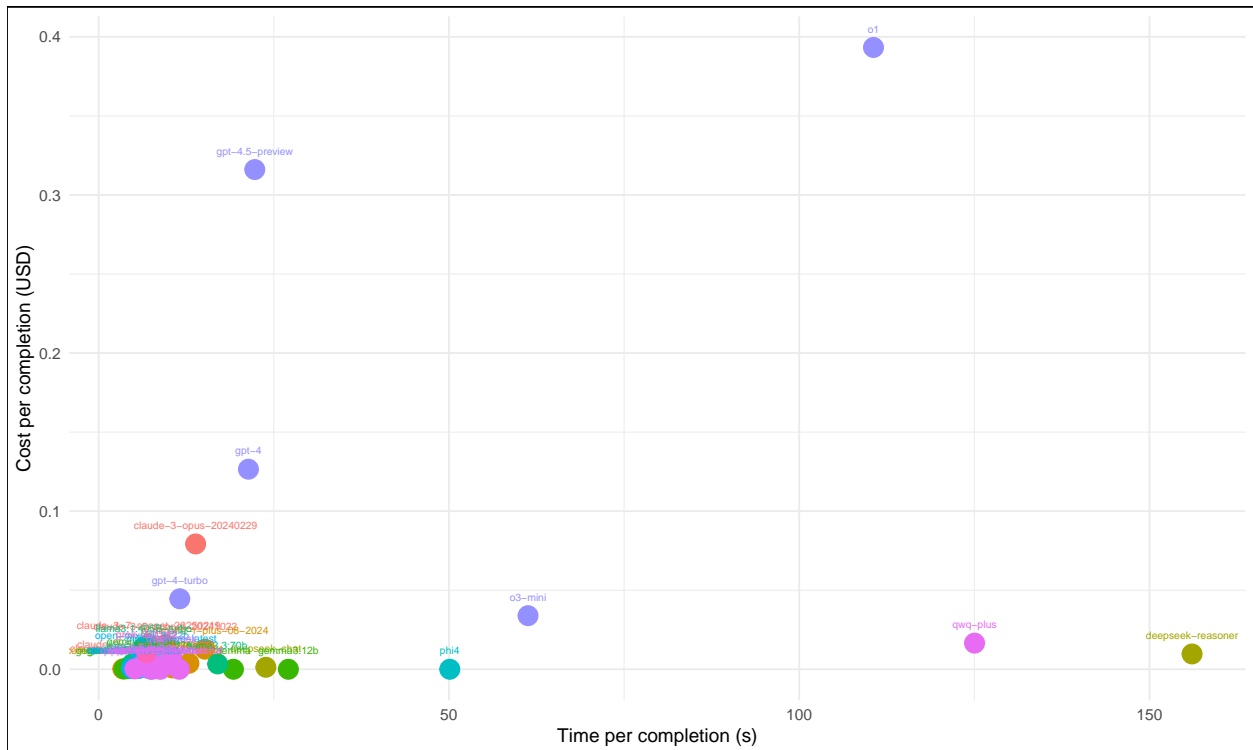
## Total cost



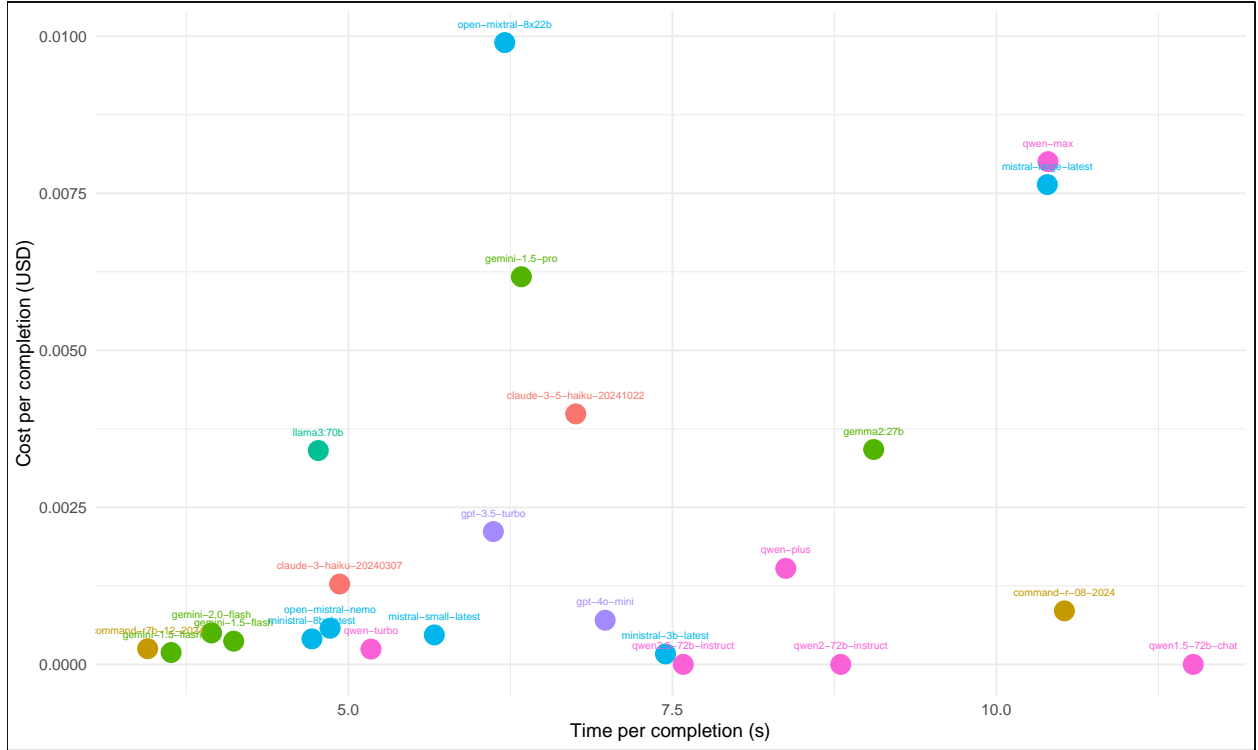
## Time per completion



## 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

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

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



	provider	model	N	all	considerations	policies
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	openai	gpt-4.5-preview	201	0.84	0.87	0.70
38	meta	llama3.1:405B-turbo	600	0.85	0.88	0.49
39	mistralai	ministral-3b-latest	600	0.85	0.86	0.53
40	cohere	command-r7b-12-2024	600	0.86	0.87	0.46
41	cohere	command-r-plus-08-2024	600	0.87	0.89	0.49
42	mistralai	open-mixtral-8x22b	600	0.87	0.90	0.52
43	openai	o1	100	0.92	0.92	0.77
44	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 30703 LLM responses into 988 responses: 1 response per model per survey.

WARNING! All considerations of cohere/command-r-plus-08-2024/fnqcj were aggregated as 1 WARNING!  
All considerations of google/gemma3:12b/valsamoggia were aggregated as 1

## Human Data

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
```

## See help("Deprecated")

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	Thalwil	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.

## 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

```

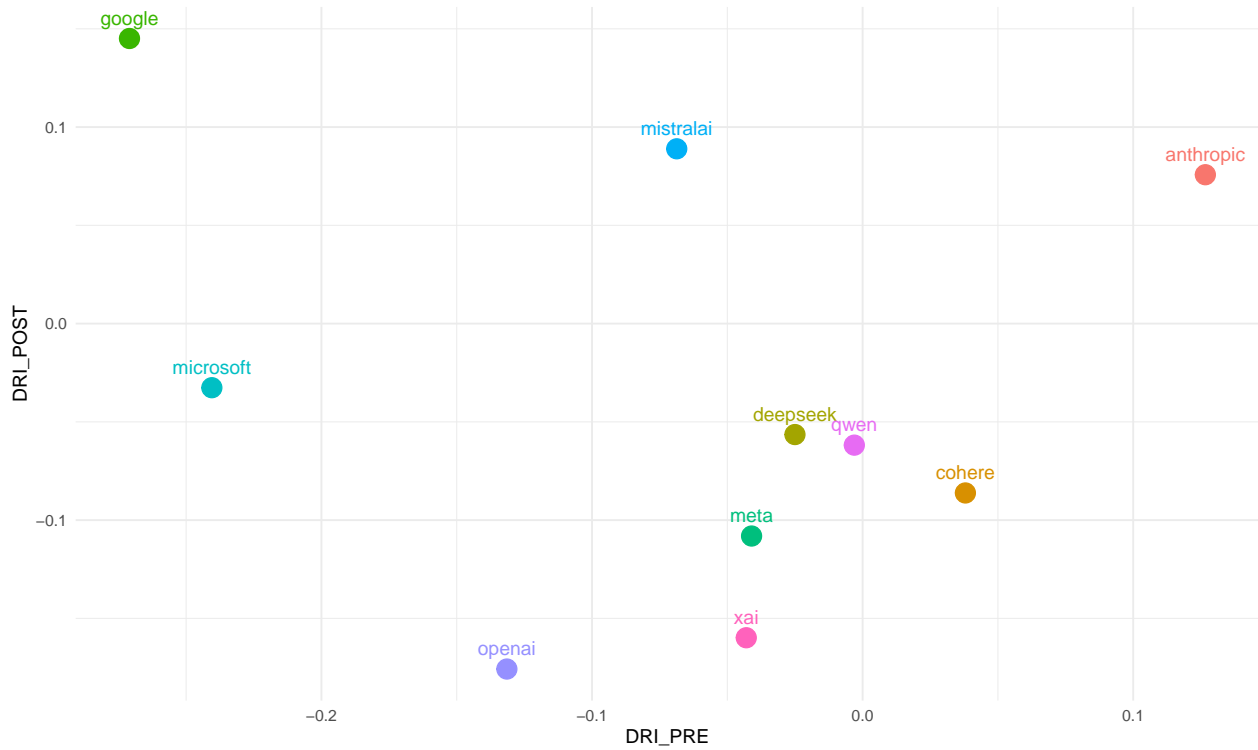
## Warning: Removed 19 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

```

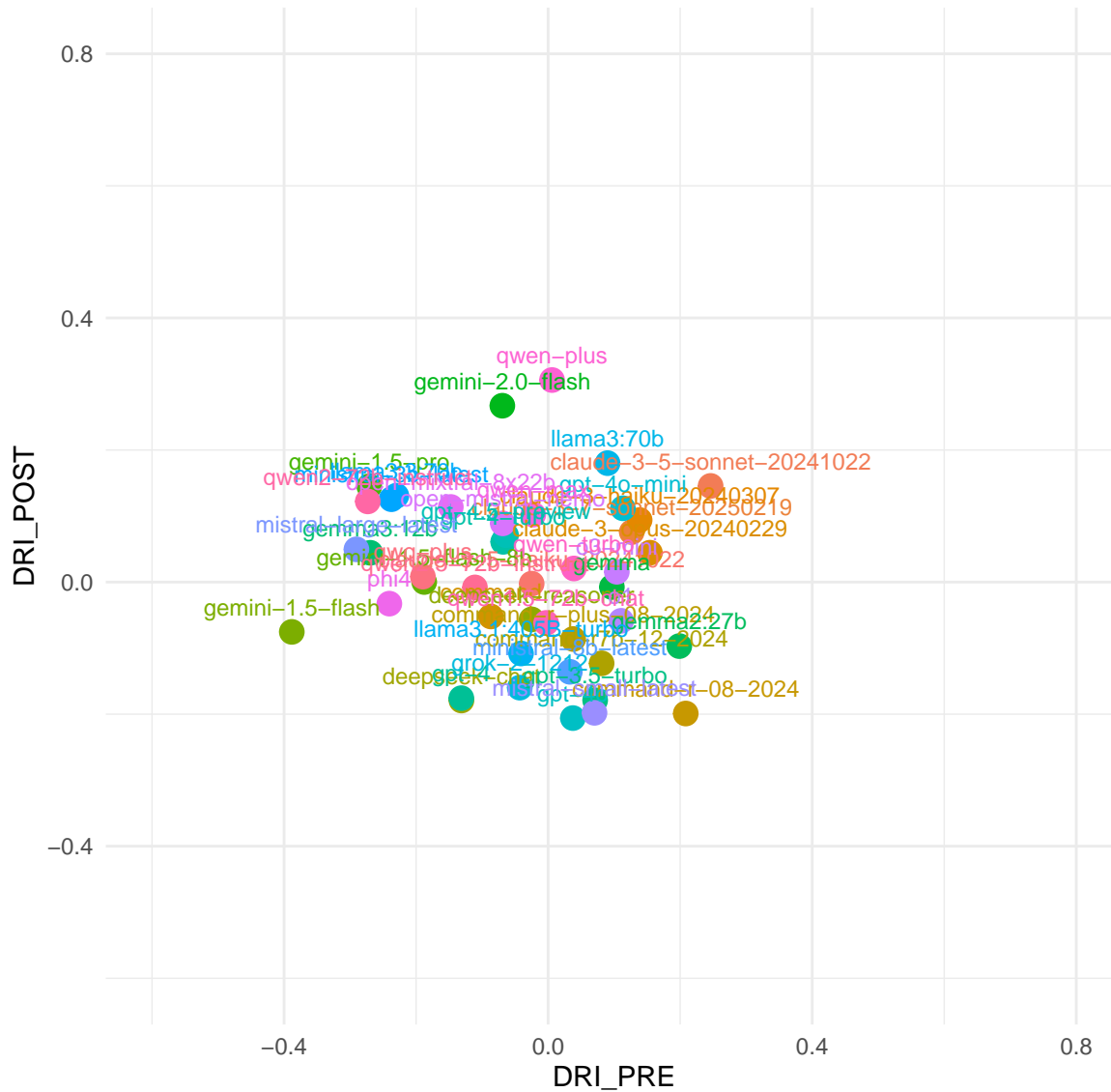
```
## `summarise()` has grouped output by 'provider', 'model'. You can override using
## the `.groups` argument.

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

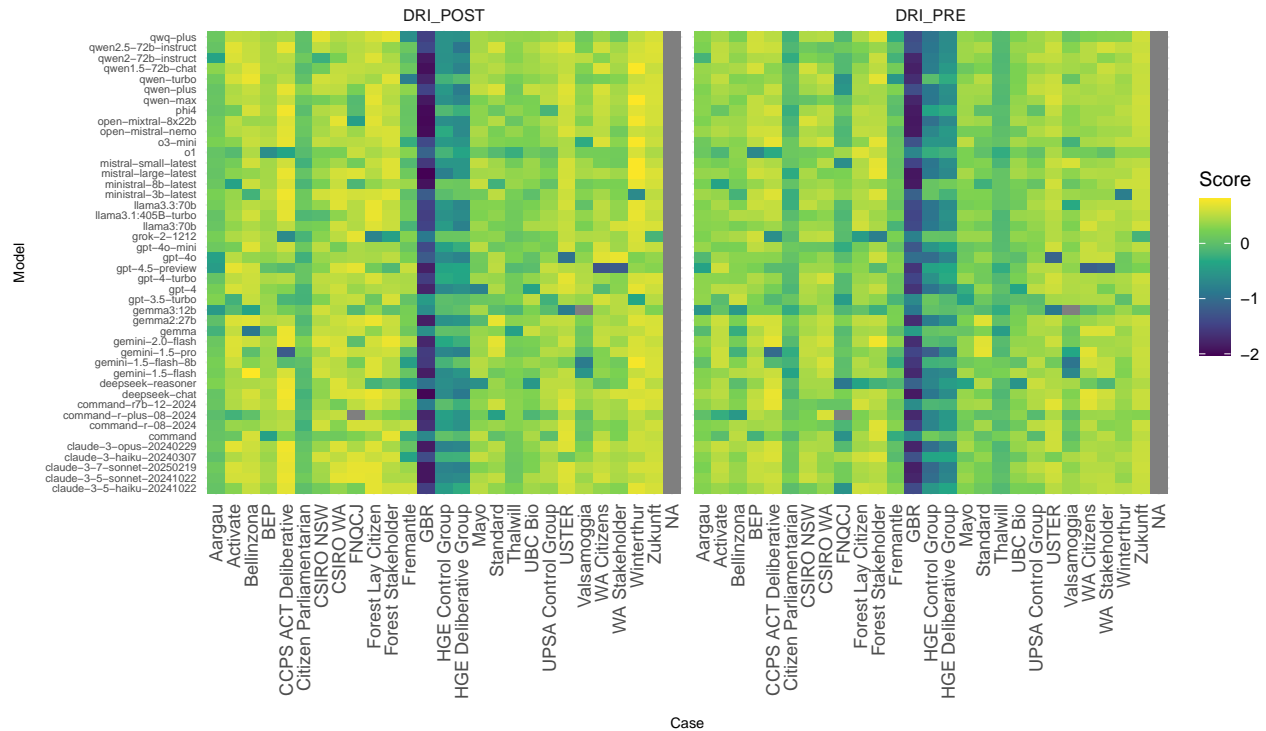
### Comparison PRE and POST DRI by Provider



Comparison PRE and POST DRI by Model



## Heatmap of DRI Scores by Case and Model

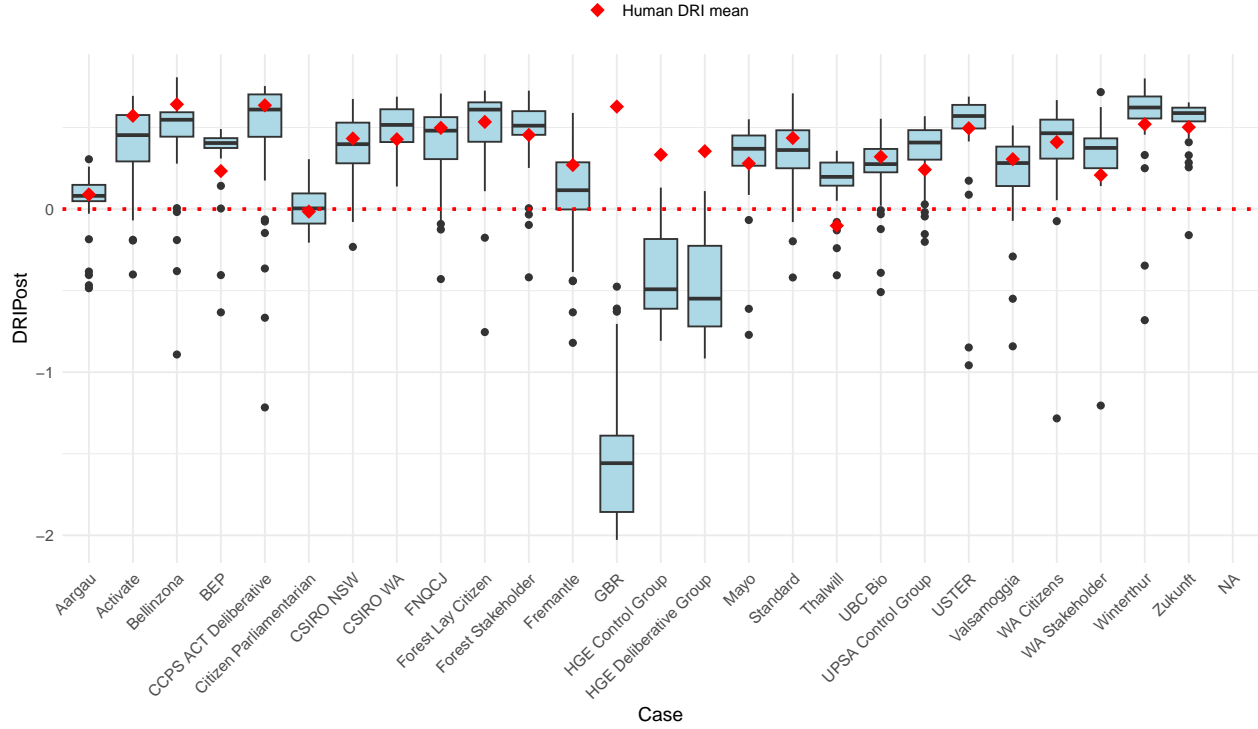


## Boxplot of LLM DRI Post by Case

```
## 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 46 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 44 rows containing non-finite outside the scale range
## (`stat_summary()`).
```



## LLM Performance Metrics Against Human DRI Post-Scores

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 7: LLM Performance Metrics Against Human DRI Post-Scores

Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
command-r7b-12-2024	0.197	0.344	85.810	0.744	0.265	0.463	0.538	-0.041
command	0.283	0.387	89.798	0.744	0.381	0.521	0.406	-0.187
gpt-3.5-turbo	0.310	0.414	128.487	0.744	0.417	0.557	-0.010	-0.185
gemma	0.245	0.424	76.739	0.744	0.330	0.570	0.339	-0.129
claude-3-haiku-20240307	0.254	0.462	98.213	0.744	0.341	0.622	0.475	-0.102
gpt-4o-mini	0.255	0.469	100.318	0.744	0.342	0.631	0.398	-0.137
gpt-4-turbo	0.227	0.478	80.697	0.744	0.306	0.643	0.547	-0.080
ministral-3b-latest	0.289	0.491	111.081	0.744	0.388	0.660	0.220	-0.131
claude-3-5-haiku-20241022	0.268	0.495	76.615	0.744	0.360	0.666	0.371	-0.108
o1	0.318	0.505	92.257	0.744	0.427	0.679	0.309	-0.301
o3-mini	0.292	0.510	95.798	0.744	0.393	0.686	0.454	-0.139
llama3.3:70b	0.275	0.514	111.403	0.744	0.369	0.691	0.521	-0.124
gemma3:12b	0.374	0.520	118.761	0.744	0.502	0.699	-0.137	-0.306
llama3.1:405B-turbo	0.260	0.521	92.533	0.744	0.349	0.701	0.537	-0.155
llama3:70b	0.298	0.526	129.718	0.744	0.400	0.707	0.380	-0.135

Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
qwen2.5-72b-instruct	0.277	0.527	84.711	0.744	0.373	0.709	0.525	-0.092
mistral-small-latest	0.284	0.527	119.671	0.744	0.382	0.709	0.483	-0.172
grok-2-1212	0.317	0.528	109.056	0.744	0.426	0.710	0.063	-0.221
qwen-plus	0.293	0.529	157.093	0.744	0.395	0.711	0.474	-0.067
gemini-2.0-flash	0.283	0.530	142.756	0.744	0.381	0.713	0.469	-0.060
command-r-08-2024	0.279	0.534	122.313	0.744	0.375	0.718	0.394	-0.143
qwq-plus	0.282	0.541	90.107	0.744	0.379	0.728	0.543	-0.153
qwen-turbo	0.267	0.548	85.491	0.744	0.360	0.737	0.562	-0.131
deepseek-reasoner	0.375	0.549	123.108	0.744	0.504	0.739	0.282	-0.258
gemini-1.5-flash-8b	0.328	0.561	97.684	0.744	0.442	0.755	0.227	-0.198
gemma2:27b	0.285	0.567	103.724	0.744	0.383	0.762	0.570	-0.101
phi4	0.287	0.571	83.983	0.744	0.385	0.767	0.426	-0.151
open-mistral-nemo	0.276	0.580	104.933	0.744	0.371	0.780	0.516	-0.120
gpt-4o	0.357	0.586	158.169	0.744	0.481	0.788	0.258	-0.252
ministral-8b-latest	0.309	0.587	109.421	0.744	0.415	0.789	0.208	-0.186
claude-3-opus-20240229	0.284	0.588	92.192	0.744	0.382	0.790	0.548	-0.114
claude-3-5-sonnet-20241022	0.289	0.589	115.990	0.744	0.388	0.791	0.573	-0.072
gemini-1.5-flash	0.307	0.592	102.964	0.744	0.413	0.797	0.521	-0.176
qwen-max	0.313	0.596	111.424	0.744	0.420	0.801	0.390	-0.162
qwen1.5-72b-chat	0.298	0.600	103.533	0.744	0.400	0.807	0.480	-0.117
claude-3-7-sonnet-20250219	0.291	0.601	99.713	0.744	0.391	0.808	0.551	-0.097
qwen2-72b-instruct	0.331	0.602	142.072	0.744	0.445	0.809	0.443	-0.166
mistral-large-latest	0.305	0.616	99.385	0.744	0.410	0.828	0.420	-0.124
command-r-plus-08-2024	0.369	0.617	119.389	0.744	0.497	0.830	0.111	-0.294
open-mixtral-8x22b	0.308	0.623	108.671	0.744	0.415	0.838	0.436	-0.165
gpt-4	0.360	0.624	141.193	0.744	0.484	0.839	0.388	-0.213
deepseek-chat	0.315	0.625	129.052	0.744	0.423	0.840	0.471	-0.106
gemini-1.5-pro	0.389	0.672	138.578	0.744	0.524	0.904	0.179	-0.221
gpt-4.5-preview	0.459	0.727	160.975	0.744	0.617	0.977	0.286	-0.348

## PRE vs. POST Aggregate Scores Correlation Across LLMs

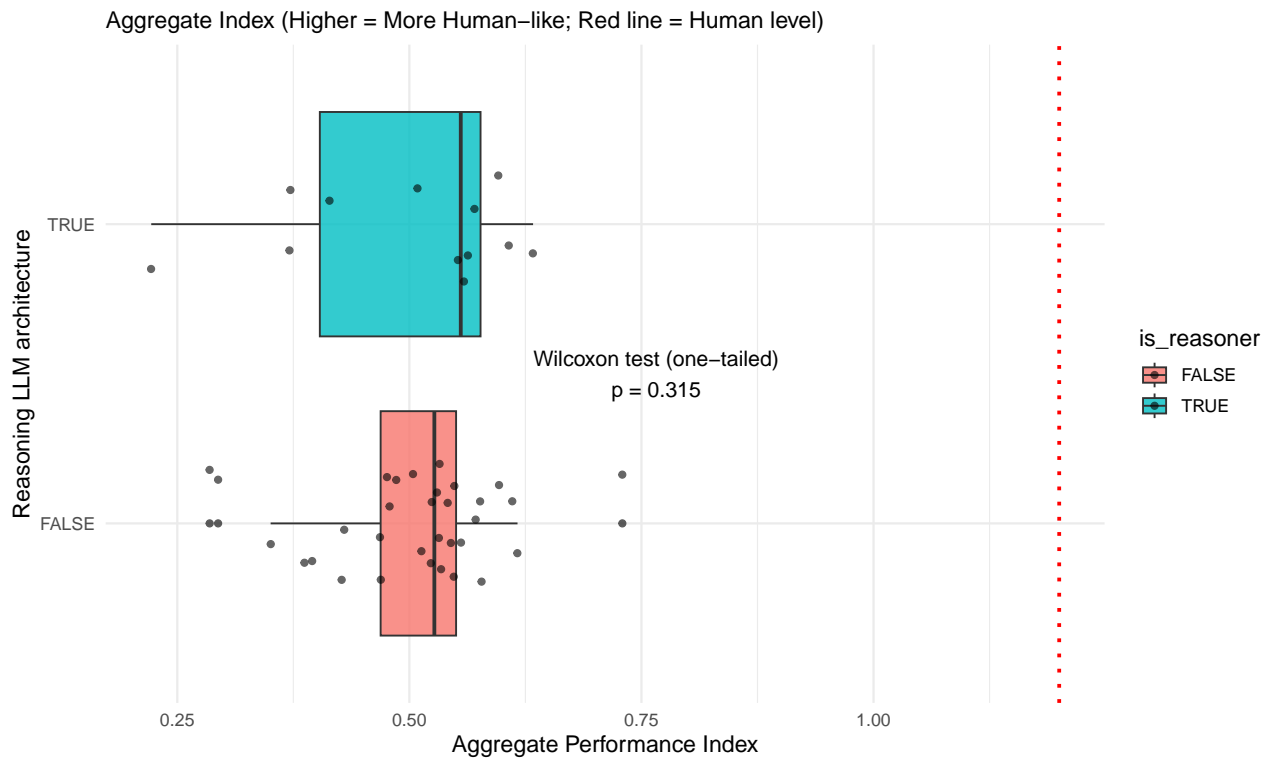
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_vline()`).

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_hline()`).
```





## LLM Performance by Reasoner Classification



## 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.