Triage Against the Machine: Can AI Reason Deliberatively?

Francesco Veri, Gustavo Umbelino

2025-04-11

Large-Language Models (LLMs) Preview

Table 1: LLMs

	Provider	Model	Series	Parameters (B)	Context Length	Architecture	Version
1	anthropic	claude-3-5-haiku-20241022	claude	NA	200000	transformer	2.0
2	anthropic	claude-3-5-sonnet-20241022	claude	NA	200000	transformer	2.0
3	anthropic	claude-3-7-sonnet-20250219	claude	NA	200000	transformer	3.0
4	anthropic	claude-3-haiku-20240307	claude	NA	200000	transformer	1.0
5	anthropic	claude-3-opus-20240229	claude	NA	200000	transformer	1.0
6	anthropic	claude-3-sonnet-20240229	claude	NA	200000	transformer	1.0
7	cohere	command	command	NA	4096	transformer	1.0
8	cohere	command-a-03-2025	command	111	288000	transformer	3.0
9	cohere	command-r-08-2024	command	32	128000	transformer	2.0
10	cohere	command-r-plus-08-2024	command	104	128000	transformer	2.0
11	cohere	command-r7b-12-2024	command	7	128000	open-weights	2.0
12	deepseek	deepseek-chat	deepseek-chat	NA	128000	transformer	3.0
13	deepseek	deepseek-reasoner	deepseek-reasoner	NA	128000	MoE	1.0
14	deepseek	deepseek-v2	deepseek-chat	NA	128000	transformer	2.0
15	deepseek	deepseek-v2.5	deepseek-chat	NA	128000	transformer	2.5
16	google	gemini-1.5-flash	gemini	NA	1000000	transformer	1.5
17	google	gemini-1.5-flash-8b	gemini	8	1048576	transformer	1.5
18	google	gemini-1.5-pro	gemini	NA	2000000	transformer	1.5
19	google	gemini-2.0-flash	gemini	NA	1000000	transformer	2.0
20	google	gemma	gemma	NA	NA	transformer	1.0
21	google	gemma2:27b	gemma	27	8190	transformer	2.0
22	google	gemma3:12b	gemma	12	128000	transformer	3.0
23	meta	llama2:13b	llama	13	4100	transformer	2.0
24	meta	llama2:70b	llama	70	4100	transformer	2.0
25	meta	llama3.1:405B-turbo	llama	405	128000	transformer	3.1
26	meta	llama3.2	llama	3	131072	transformer	3.1

	Provider	Model	Series	Parameters (B)	Context Length	Architecture	Version
27	meta	llama3.3:70b	llama	70	128000	transformer	3.3
28	meta	llama3:70b	llama	70	8190	transformer	3.0
29	meta	llama4-maverick	llama	17	1000000	MoE	4.0
30	meta	llama4-scout	llama	17	1000000000	MoE	4.0
31	microsoft	phi	phi	NA	NA	transformer	1.0
32	microsoft	phi2	phi	NA	NA	transformer	2.0
33	microsoft	phi3	phi	NA	NA	transformer	3.0
34	microsoft	phi3.5	phi	NA	NA	transformer	3.5
35	microsoft	phi4	phi	14	16000	decoder-only	4.0
36	mistralai	ministral-3b-latest	ministral	3	128000	transformer	NA
37	mistralai	ministral-8b-latest	ministral	8	128000	transformer	NA
38	mistralai	mistral-large-latest	mistral	123	128000	transformer	NA
39	mistralai	mistral-small-latest	mistral	22	32800	transformer	NA
40	mistralai	open-mistral-7b	mistral	7	NA	transformer	NA
41	mistralai	open-mistral-nemo	mistral	12	128000	transformer	NA
42	mistralai	open-mixtral-8x22b	mixtral	39	65400	SMoE	NA
43	mistralai	open-mixtral-8x7b	mixtral	7	NA	SMoE	NA
44	openai	gpt-3.5-turbo	gpt	NA	16385	transformer	3.5
45	openai	gpt-4	gpt	NA	8192	transformer	4.0
46	openai	gpt-4-turbo	gpt	NA	128000	transformer	4.0
47	openai	gpt-4.5-preview	gpt	NA	128000	transformer	4.5
48	openai	gpt-4o	gpt	NA	128000	transformer	5.0
49	openai	gpt-4o-mini	gpt	NA	128000	transformer	5.0
50	openai	o1	o	NA	200000	transformer	1.0
51	openai	o1-mini	o	NA	NA	transformer	NA
52	openai	o3-mini	o	NA	200000	transformer	3.0
53	qwen	qwen-max	qwen	NA	32768	transformer	NA
54	qwen	qwen-plus	qwen	NA	131072	transformer	NA
55	qwen	qwen-turbo	qwen	NA	1000000	transformer	NA
56	qwen	qwen1.5-110b-chat	qwen	110	NA	transformer	1.5
57	qwen	qwen1.5-72b-chat	qwen	72	8000	transformer	1.5
58	qwen	qwen2-72b-instruct	qwen	72	131072	transformer	2.0
59	qwen	qwen2.5-72b-instruct	qwen	72	131072	transformer	2.5
60	qwen	qwq-plus	qwq	NA	131072	transformer	NA
61	xai	grok-2-1212	grok	NA	131072	transformer	2.0
62	xai	grok-3-beta	grok	NA	131072	transformer	3.0
63	xai	grok-3-mini-beta	grok	NA	131072	transformer	3.0

We started the analysis with 63 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

	survey	considerations	policies	scale_max	q_method
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
20	zukunft	20	7	7	FALSE

LLM Data Collection

We collected a total of 32765 valid LLM responses across 20 surveys.

Cost

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

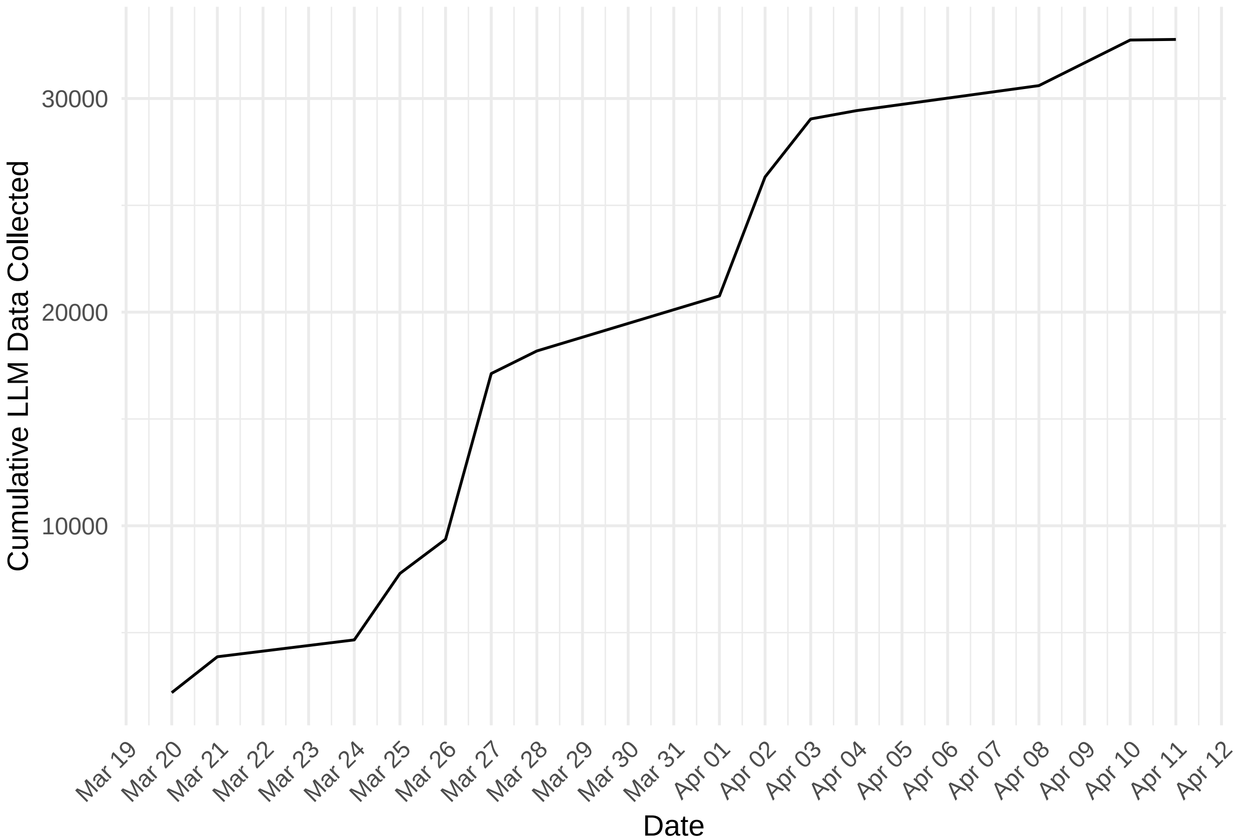
Table 3: Costs by API

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	8	13.00
Cohere API	5	12.70
DeepSeek API	2	10.00
xAI API	3	10.00
Google Could	4	NA
ollama	9	NA

Time

It took a total of 157 hours[1] across 22 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 Friday, Apr 11, 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

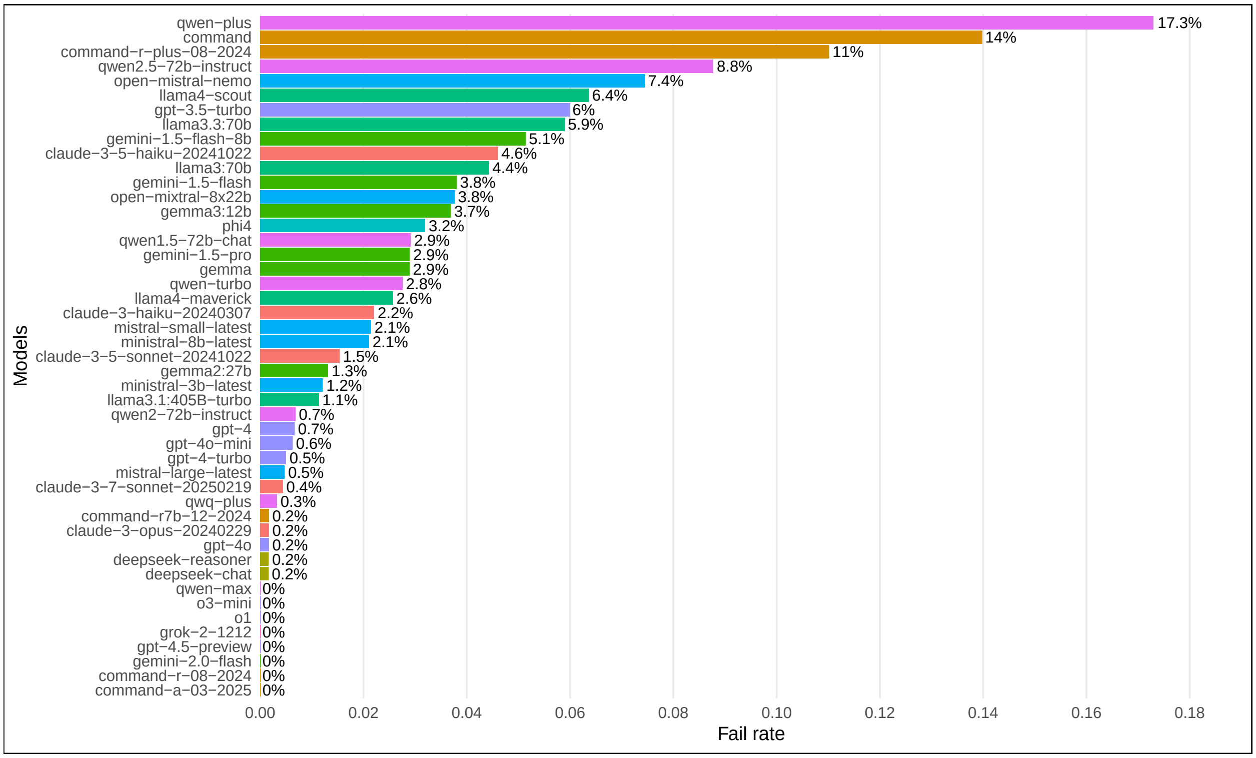
16 out of 65 were excluded from the analysis for the following reasons.

Table 4: Excluded models and reasons

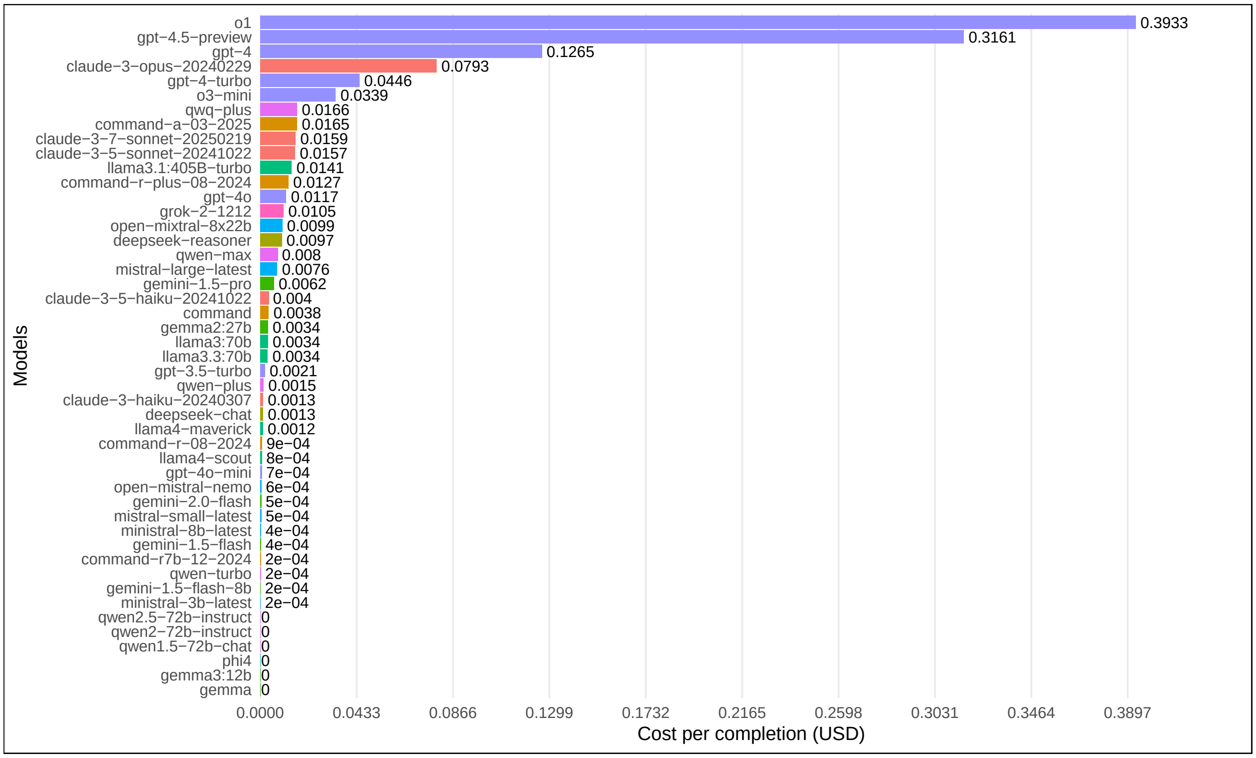
provider	model	reason
anthropic	claude-3-sonnet-20240229	not available in Anthropic API anymore
deepseek	deepseek-v2	high fail rate (85%)
deepseek	deepseek-v2.5	too big to run locally; not available through APIs
meta	llama2:13b	does not respond to prompts correctly
meta	llama2:70b	does not respond to 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
xai	grok-3-beta	NA
xai	grok-3-mini-beta	NA

Execution Summary Plots

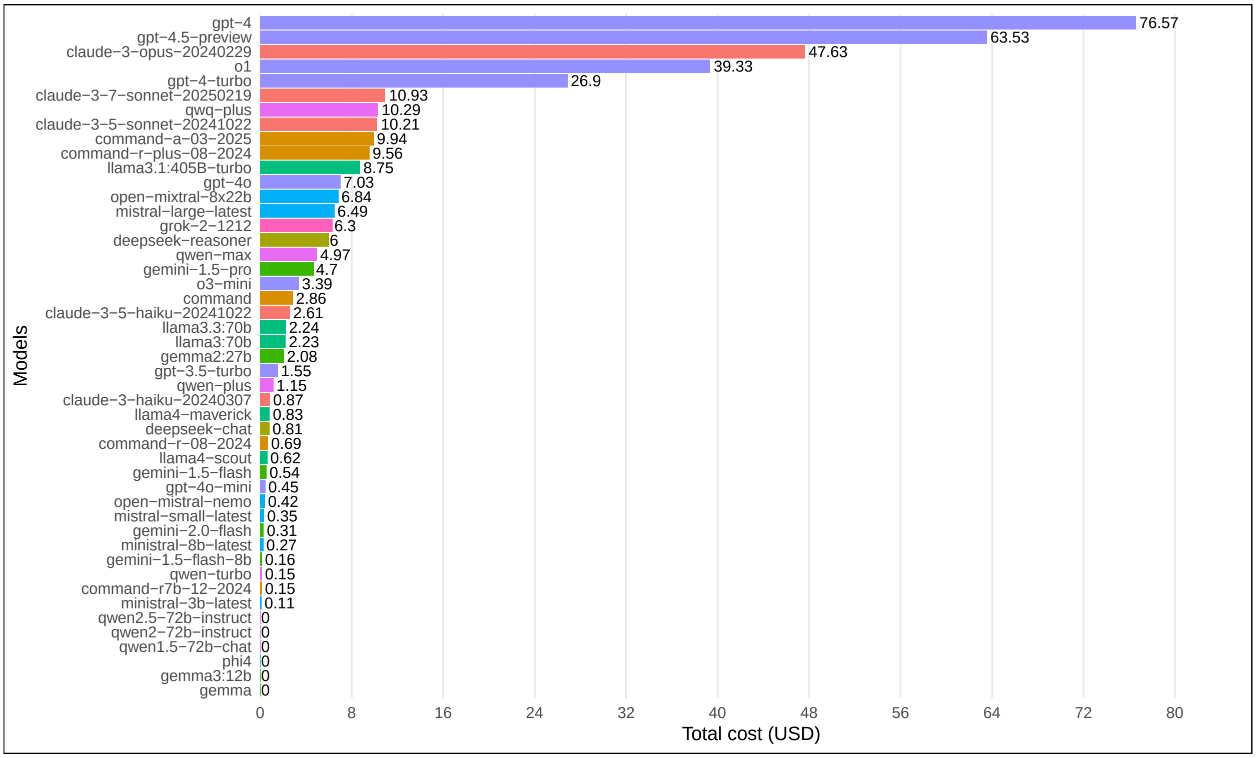
Fail rate



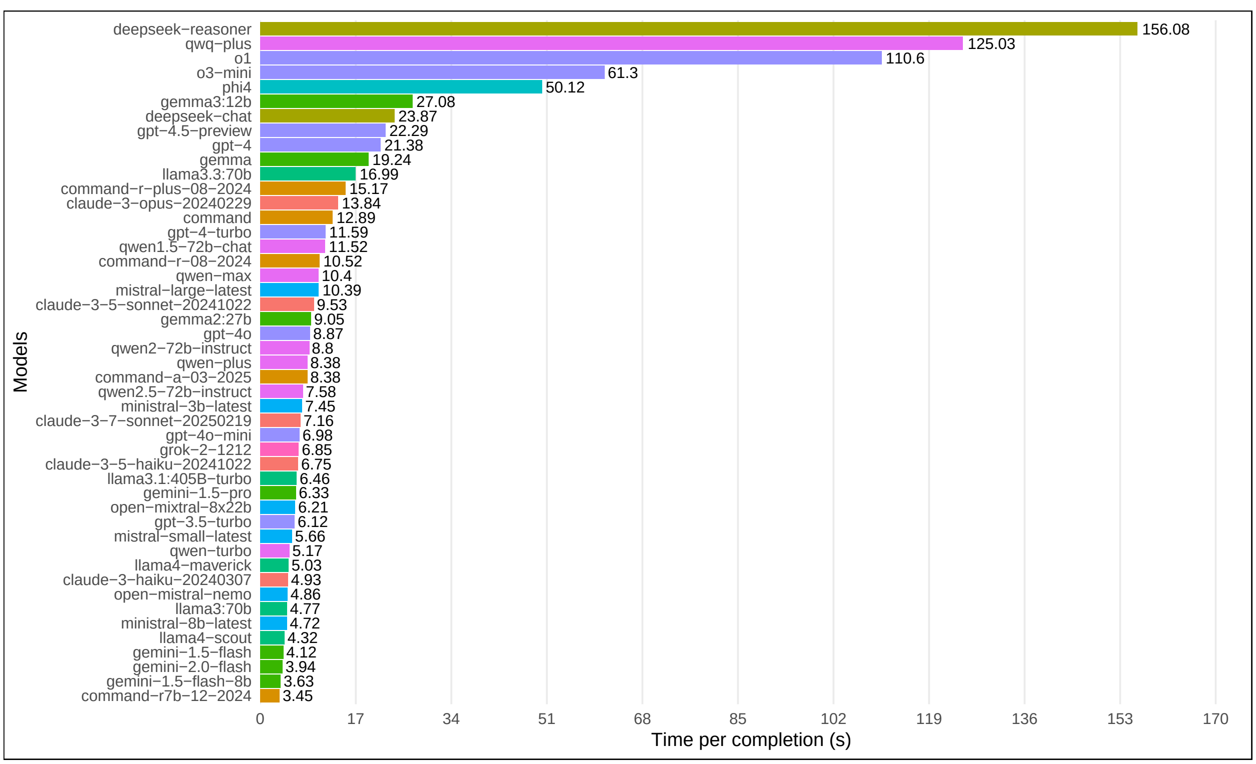
Cost per completion



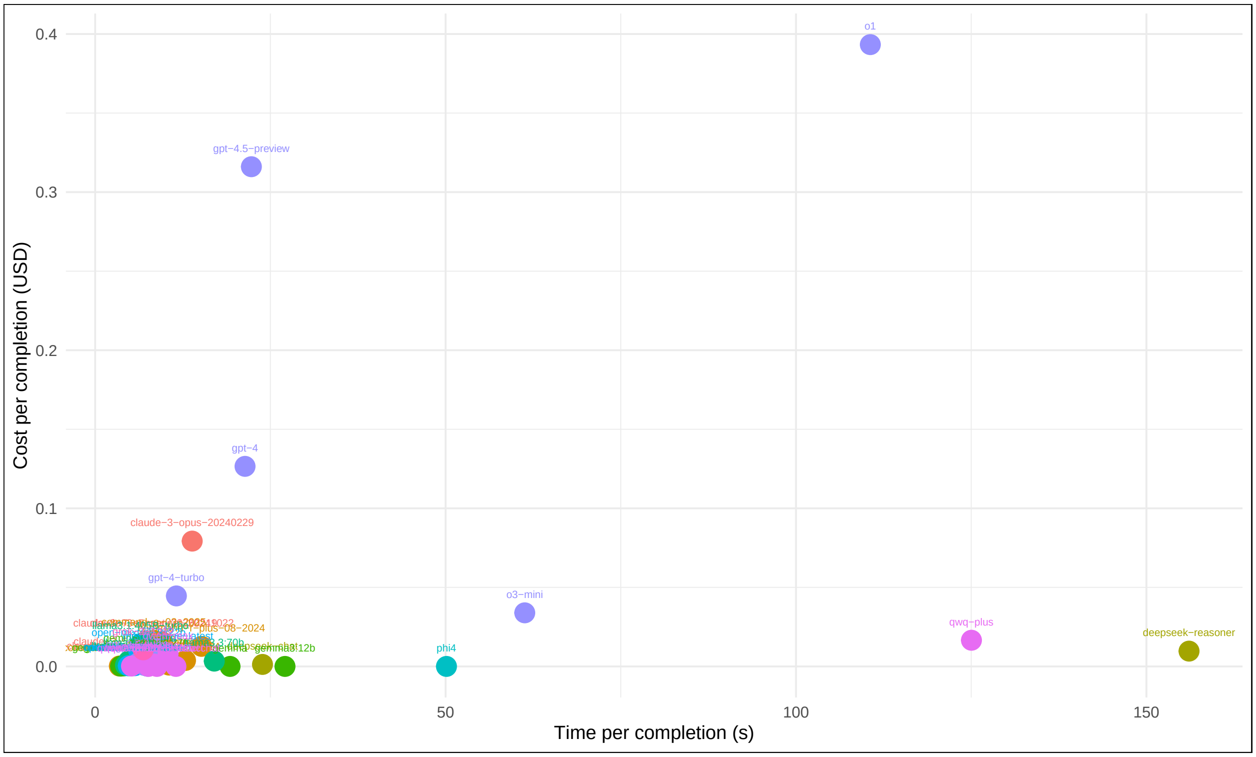
Total cost



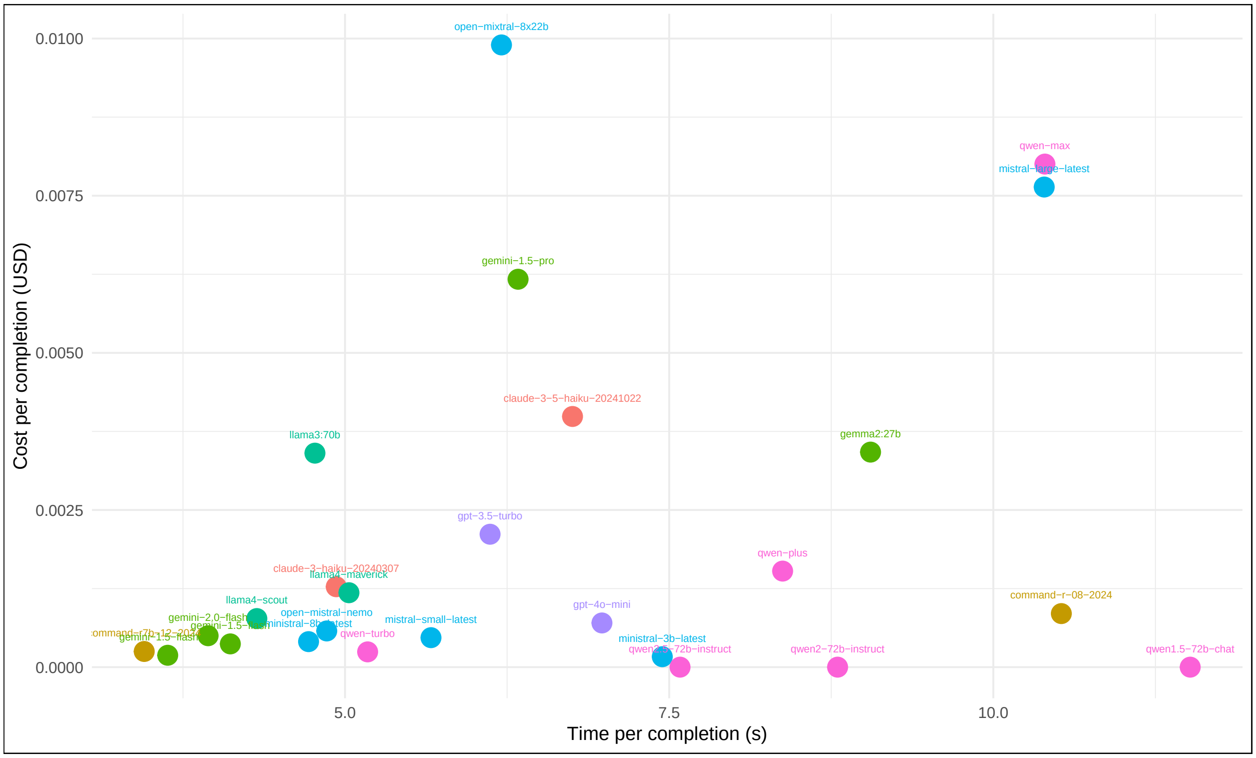
Time per completion



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.

```
## updating: grok-3-beta / acp

## updating: grok-3-beta / fnqcj
```

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	meta	llama4-maverick	600	0.71	0.78	0.44
4	openai	gpt-4o-mini	600	0.72	0.74	0.45
5	anthropic	claude-3-haiku-20240307	600	0.74	0.82	0.44
6	google	gemini-1.5-flash	600	0.74	0.76	0.52
7	anthropic	claude-3-5-sonnet-20241022	600	0.75	0.81	0.58
8	deepseek	deepseek-reasoner	600	0.75	0.79	0.55
9	openai	gpt-4	600	0.75	0.82	0.52
10	openai	gpt-4-turbo	600	0.75	0.82	0.53
11	google	gemini-1.5-pro	600	0.76	0.78	0.57
12	openai	gpt-4o	600	0.76	0.86	0.50
13	cohere	command	600	0.78	0.78	0.44
14	google	gemma	600	0.78	0.80	0.45
15	meta	llama3.3:70b	600	0.78	0.82	0.52
16	mistralai	mistral-small-latest	600	0.78	0.84	0.52
17	mistralai	open-mistral-nemo	600	0.78	0.80	0.49
18	qwen	qwq-plus	600	0.78	0.79	0.58
19	xai	grok-2-1212	600	0.78	0.89	0.47
20	cohere	command-a-03-2025	600	0.79	0.86	0.51
21	cohere	command-r-08-2024	600	0.79	0.81	0.50
22	deepseek	deepseek-chat	600	0.79	0.86	0.52
23	google	gemini-1.5-flash-8b	600	0.79	0.84	0.50
24	meta	llama3:70b	600	0.79	0.79	0.52
25	qwen	qwen-turbo	600	0.79	0.83	0.48
26	anthropic	claude-3-7-sonnet-20250219	600	0.80	0.84	0.53
27	meta	llama4-scout	600	0.80	0.85	0.51
28	qwen	qwen-plus	600	0.80	0.82	0.49
29	qwen	qwen2-72b-instruct	600	0.80	0.86	0.48
30	qwen	qwen2.5-72b-instruct	600	0.80	0.84	0.51
31	anthropic	claude-3-5-haiku-20241022	600	0.81	0.86	0.47
32	google	gemma3:12b	600	0.81	0.81	0.47
33	microsoft	phi4	600	0.81	0.82	0.55
34	mistralai	ministral-8b-latest	600	0.82	0.83	0.51
35	qwen	qwen-max	600	0.82	0.84	0.51
36	anthropic	claude-3-opus-20240229	600	0.83	0.87	0.50
37	mistralai	mistral-large-latest	600	0.83	0.86	0.54
38	google	gemini-2.0-flash	600	0.84	0.84	0.62
39	openai	gpt-3.5-turbo	600	0.84	0.87	0.48
40	openai	gpt-4.5-preview	201	0.84	0.87	0.70
41	meta	llama3.1:405B-turbo	600	0.85	0.88	0.49
42	mistralai	ministral-3b-latest	600	0.85	0.86	0.53

	provider	model	N	all	considerations	policies
43	cohere	command-r7b-12-2024	600	0.86	0.87	0.46
44	cohere	command-r-plus-08-2024	600	0.87	0.89	0.49
45	mistralai	open-mixtral-8x22b	600	0.87	0.90	0.52
46	openai	o1	100	0.92	0.92	0.77
47	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

```
## updating: grok-3-beta / acp
## updating: grok-3-beta / fnqcj
```

We aggregated 32765 LLM responses into 1053 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

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

	Case	survey	participants
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.

```r
# 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) {
```

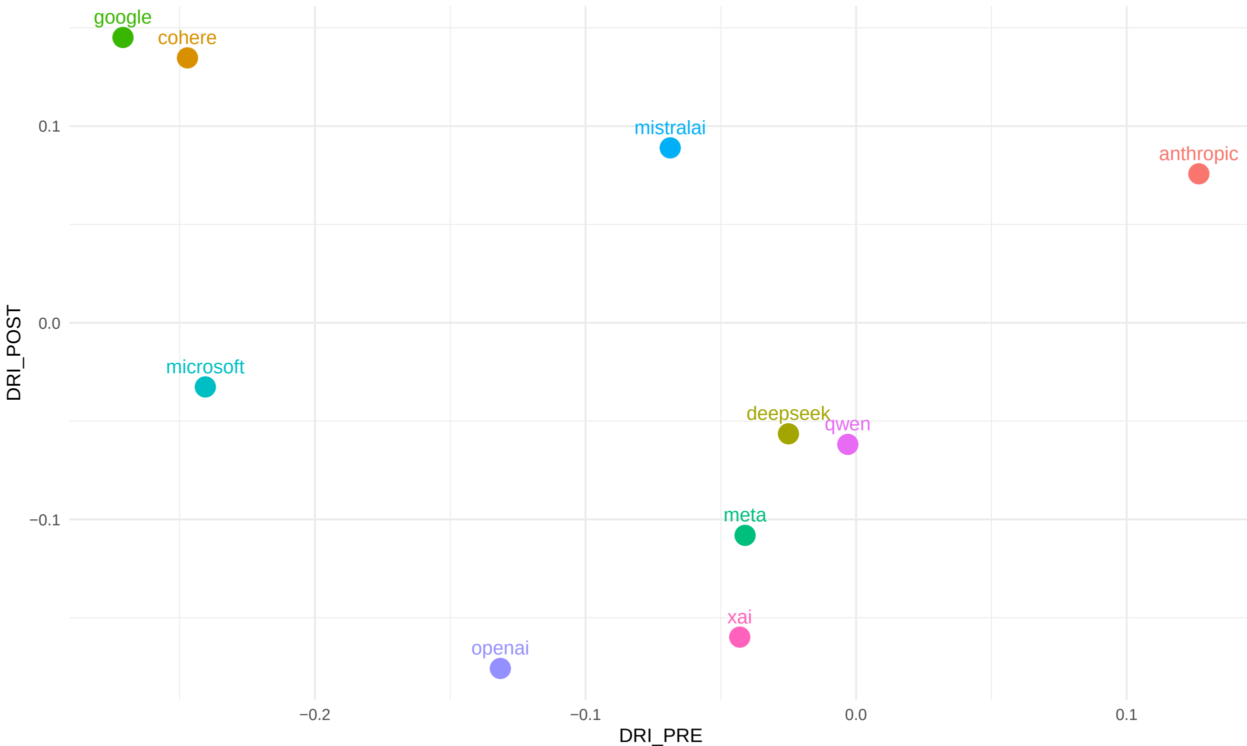
```r
  d <- abs((data[[v1]] - data[[v2]]) / sqrt(2))
  lambda <- 1 - (sqrt(2) / 2)

  # 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]]

  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!
```

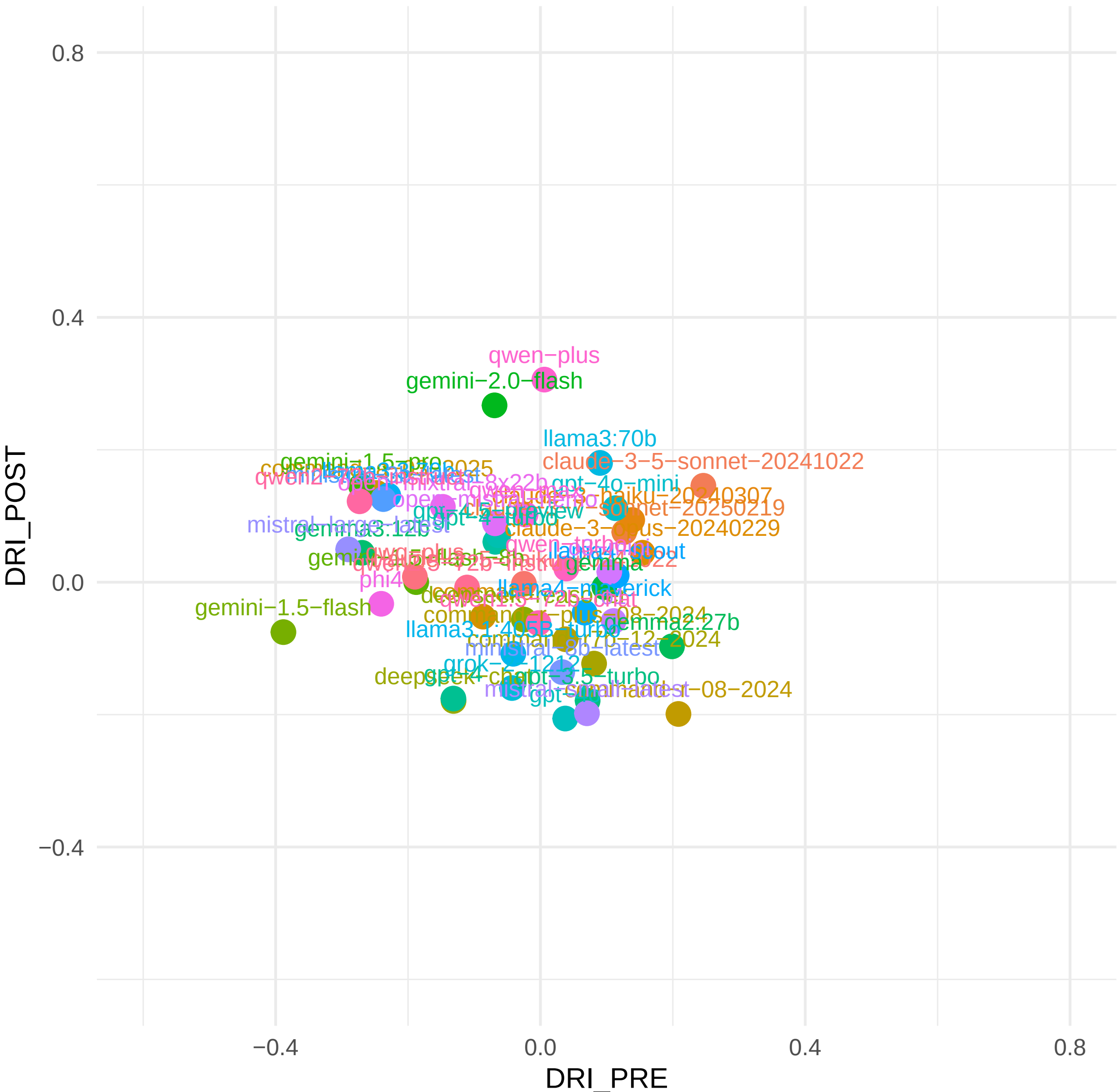
DRI Benchmark

```
## `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
```

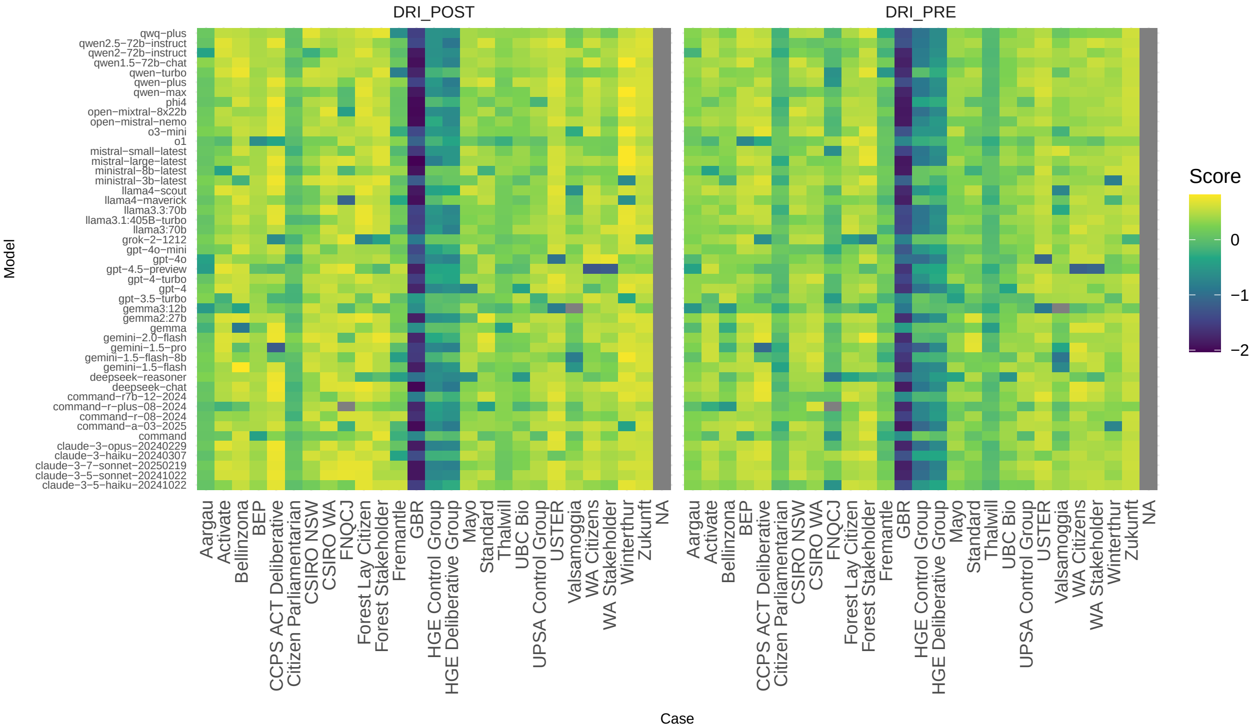
Comparison PRE and POST DRI by Provider



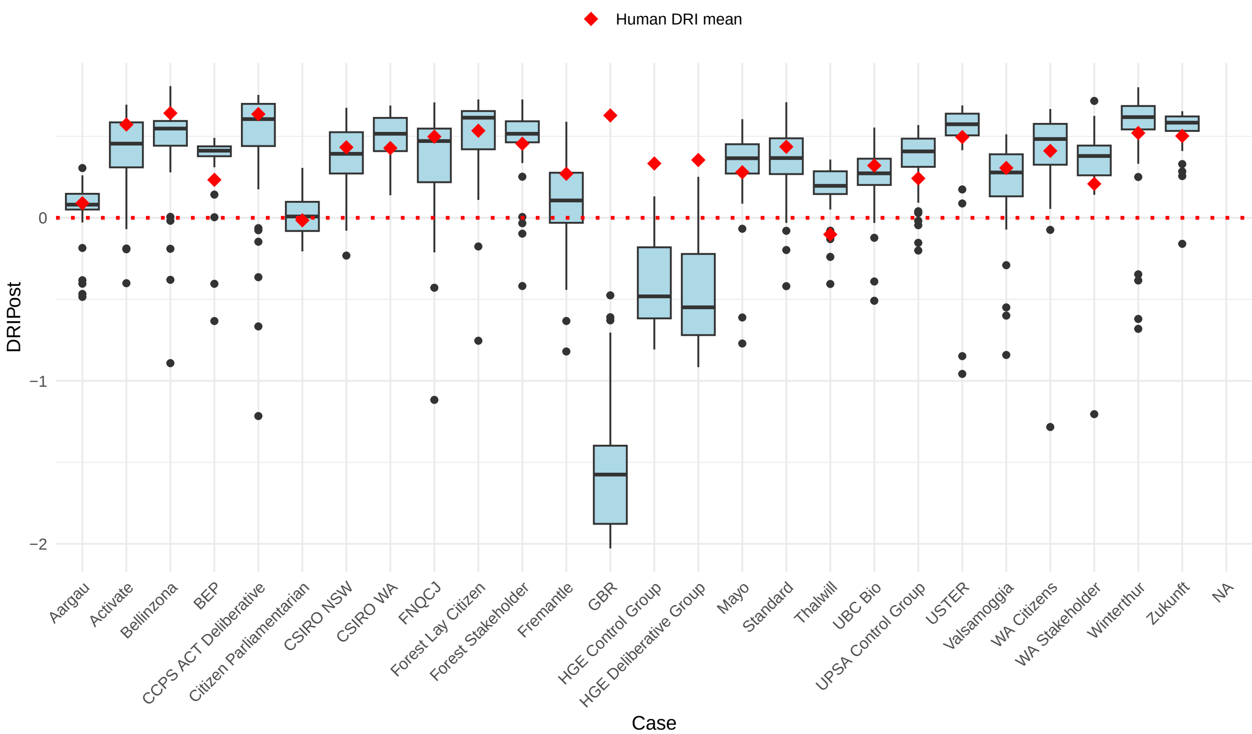
Comparison PRE and POST DRI by Model



Heatmap of DRI Scores by Case and Model



Boxplot of LLM DRI Post by Case



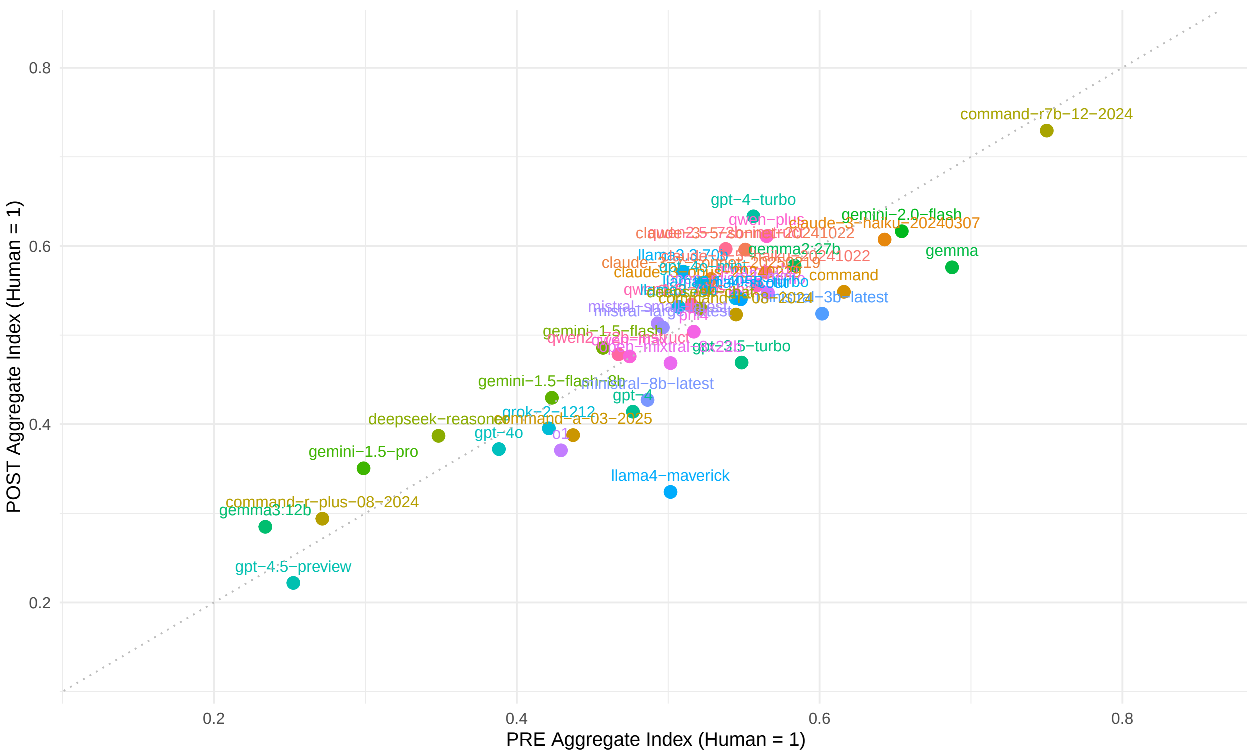
LLM Performance Metrics Against Human DRI Post-Scores

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

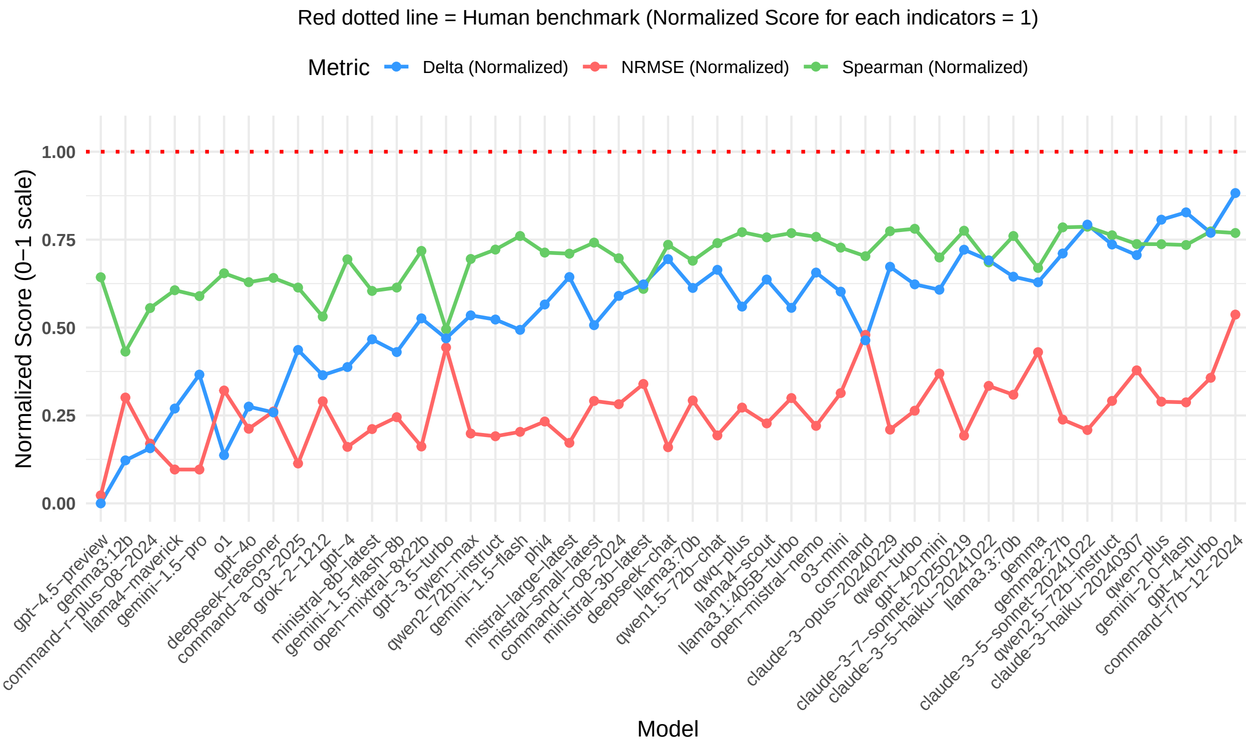
Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
command-r7b-12-2024	0.197	0.344	85.810	0.744	0.265	0.463	0.538	-0.041
command	0.283	0.387	89.798	0.744	0.381	0.521	0.406	-0.187
gpt-3.5-turbo	0.310	0.414	128.487	0.744	0.417	0.557	-0.010	-0.185
gemma	0.245	0.424	76.739	0.744	0.330	0.570	0.339	-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
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
llama4-scout	0.287	0.575	86.507	0.744	0.386	0.773	0.513	-0.127
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
command-a-03-2025	0.375	0.659	140.325	0.744	0.504	0.887	0.227	-0.196
llama4-maverick	0.358	0.672	98.374	0.744	0.482	0.904	0.212	-0.254
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

Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta

PRE vs. POST Aggregate Scores Correlation Across LLMs



Human-Normalized Performance

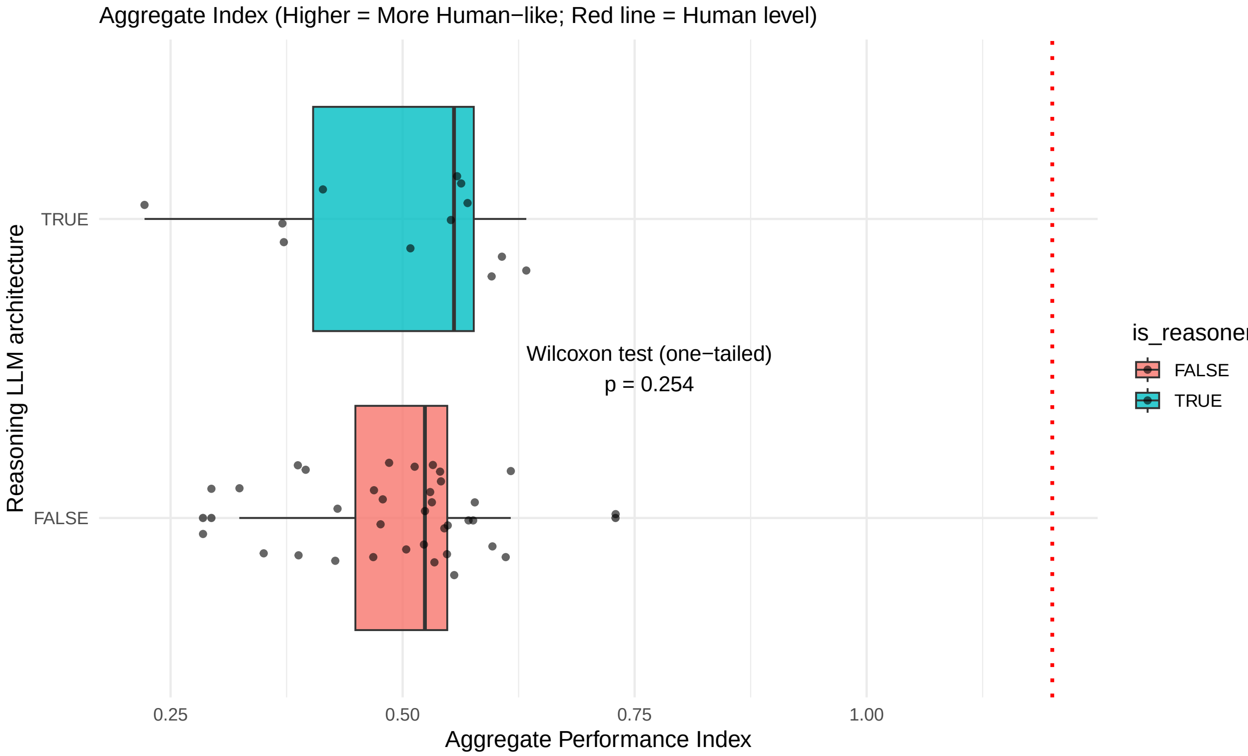


LLM Performance by Reasoner Classification

Architecture types:

- Transformer-based models (Vaswani et al. 2017).

Some models are considered "reasoning" models, like , reason using chain-of-thought (CoT) – this is not a difference in architecture, but in how



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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.