Triage Against the Machine: Can AI Reason Deliberatively?

Francesco Veri, Gustavo Umbelino

2025-05-22

Large-Language Models (LLMs) Preview

Table 1: LLMs

	Provider	Model	Parameters (B)	Context Length	Architecture	Version
1	anthropic	claude-3-5-haiku-20241022	-	200000	-	2
2	anthropic	claude-3-5-sonnet-20241022	-	200000	-	2
3	anthropic	claude-3-7-sonnet-20250219	-	200000	-	3
4	anthropic	claude-3-7-sonnet-20250219-think=high	-	200000	-	3
5	anthropic	claude-3-7-sonnet-20250219-think=low	-	200000	-	3
6	anthropic	claude-3-haiku-20240307	-	200000	-	1
7	anthropic	claude-3-opus-20240229	-	200000	-	1
8	anthropic	claude-3-sonnet-20240229	-	200000	-	1
9	cohere	command	-	4096	-	1
10	cohere	command-a-03-2025	111	288000	dense, decoder-only	3
11	cohere	command-r-08-2024	32	128000	-	2
12	cohere	command-r-plus-08-2024	104	128000	dense, decoder-only	2
13	cohere	command-r7b-12-2024	7	128000	-	2
14	deepseek	deepseek-chat	671	128000	MoE	3
15	deepseek	deepseek-reasoner	671	128000	MoE	1
16	deepseek	deepseek-v2	NA	128000	-	1
17	deepseek	deepseek-v2.5	NA	128000	-	2
18	google	gemini-1.5-flash	-	1000000	MoE	1
19	google	gemini-1.5-flash-8b	8	1048576	MoE	1
20	google	gemini-1.5-pro	-	2000000	MoE	1
21	google	gemini-2.0-flash	-	1000000	-	2
22	google	gemini-2.0-flash-thinking-exp	NA	NA	NA	2
23	google	gemini-2.5-pro-preview-03-25	-	1048576	-	2
24	google	gemma	-	-	dense, decoder-only	1
25	google	gemma-3-27b-it	27	NA	NA	3
26	google	gemma2:27b	27	8190	dense, decoder-only	2
27	google	gemma3:12b	12	128000	-	3
28	ibm	granite3.3	8	131072	dense	3
29	meta	llama2:13b	13	4100	-	1

			Parameters	Context		
	Provider	Model	(B)	Length	Architecture	Version
30	meta	llama2:70b	70	4100	-	1
31	meta	llama3.1:405B-turbo	405	128000	-	2
32	meta	llama3.2	3	131072	-	4
33	meta	llama3.3:70b	70	128000	-	3
34	meta	llama3:70b	70	8190	-	1
35	meta	llama4-maverick	17	1000000	MoE	4
36	meta	llama4-scout	17	1000000000	MoE	4
37	microsoft	phi	NA	NA	-	1
38	microsoft	phi2	NA	NA	-	2
39	microsoft	phi3	NA	NA	-	3
40	microsoft	phi3.5	NA	NA	-	4
41	microsoft	phi4	14	16000	dense, decoder-only	5
42	mistralai	ministral-3b-latest	3	128000	-	1
43	mistralai	ministral-8b-latest	8	128000	-	1
44	mistralai	mistral-large-latest	123	128000	-	1
45	mistralai	mistral-small-latest	22	32800	-	1
46	mistralai	open-mistral-7b	7	NA	-	NA
47	mistralai	open-mistral-nemo	12	128000	-	1
48	mistralai	open-mixtral-8x22b	39	65400	SMoE	1
49	mistralai	open-mixtral-8x7b	7	NA	SMoE	NA
50	openai	gpt-3.5-turbo	-	16385	-	1
51	openai	gpt-4	-	8192	-	3
52	openai	gpt-4-turbo	-	128000	-	2
53	openai	gpt-4.5-preview	-	128000	-	4
54	openai	gpt-4o	-	128000	-	2
55	openai	gpt-4o-mini	-	128000	-	2
56	openai	o1	-	200000	-	1
57	openai	o1-mini	NA	NA	-	1
58	openai	o3-mini	-	200000	-	2
59	qwen	qwen-max	-	32768	-	1
60	qwen	qwen-plus	-	131072	-	1
61	qwen	qwen-turbo	-	1000000	-	1
62	qwen	qwen1.5-110b-chat	110	NA	-	1
63	qwen	qwen1.5-72b-chat	72	8000	-	1
64	qwen	qwen2-72b-instruct	72	131072	-	2
65	qwen	qwen2.5-72b-instruct	72	131072	-	3
66	qwen	qwq-plus	-	131072	-	1
67	xai	grok-2-1212	-	131072	-	2
68	xai	grok-3-beta	-	131072	-	3
69	xai	grok-3-mini-beta	-	131072	-	3
70	xai	grok-3-mini-beta-r=high	-	131072	-	3
71	xai	grok-3-mini-beta-r=low	-	131072	-	3
72	xai	grok-3-mini-fast-beta	-	131072	-	3
73	xai	grok-3-mini-fast-beta-r=high	-	131072	-	3
74	xai	grok-3-mini-fast-beta-r=low	-	131072	-	3
75	xai	grok-beta	314	131072	MoE	1

We started the analysis with 75 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
20	zukunft	20	7	7	FALSE

LLM Data Collection

Handle special models

command-r7b-12-2024-t=1 grok-3-beta-r=TRUE

We collected a total of 37896 valid LLM responses across 20 surveys.

Cost

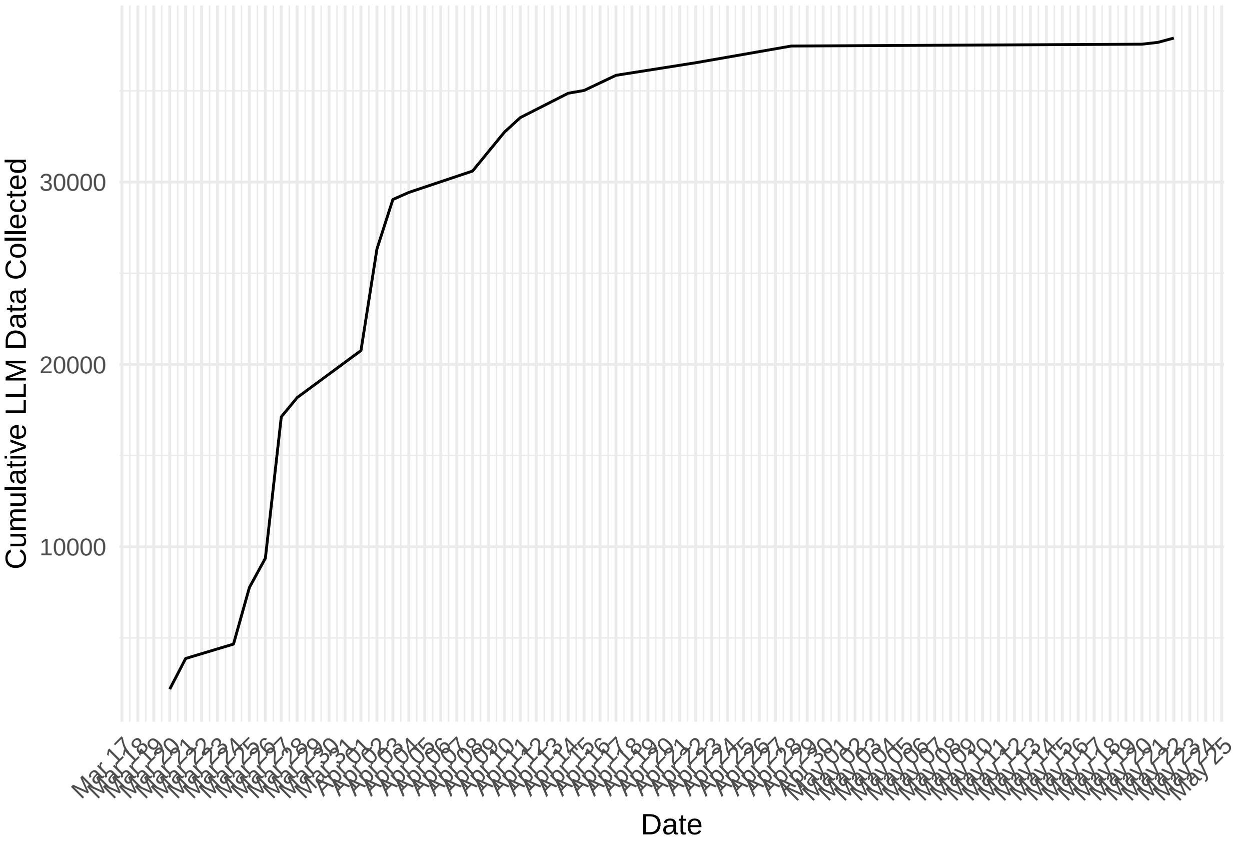
We spent a total of 429.04 USD. The cost breakdown per API is below.

Table 3: Costs by API

api	num_models	credits_paid
OpenAI API	9	225.52
Anthropic API	8	90.00
xAI API	9	30.98
Cohere API	6	20.47
Mistral AI API	8	20.00
Alibaba Cloud	8	17.49
Together AI	8	14.58
DeepSeek API	2	10.00
Google Could	8	NA
ollama	10	NA

Time

It took a total of 186 hours[1] across 63 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, May 22, 2025.



Excluded Models

18 out of 79 were excluded from the analysis for the following reasons.

Table 4: Excluded models and reasons

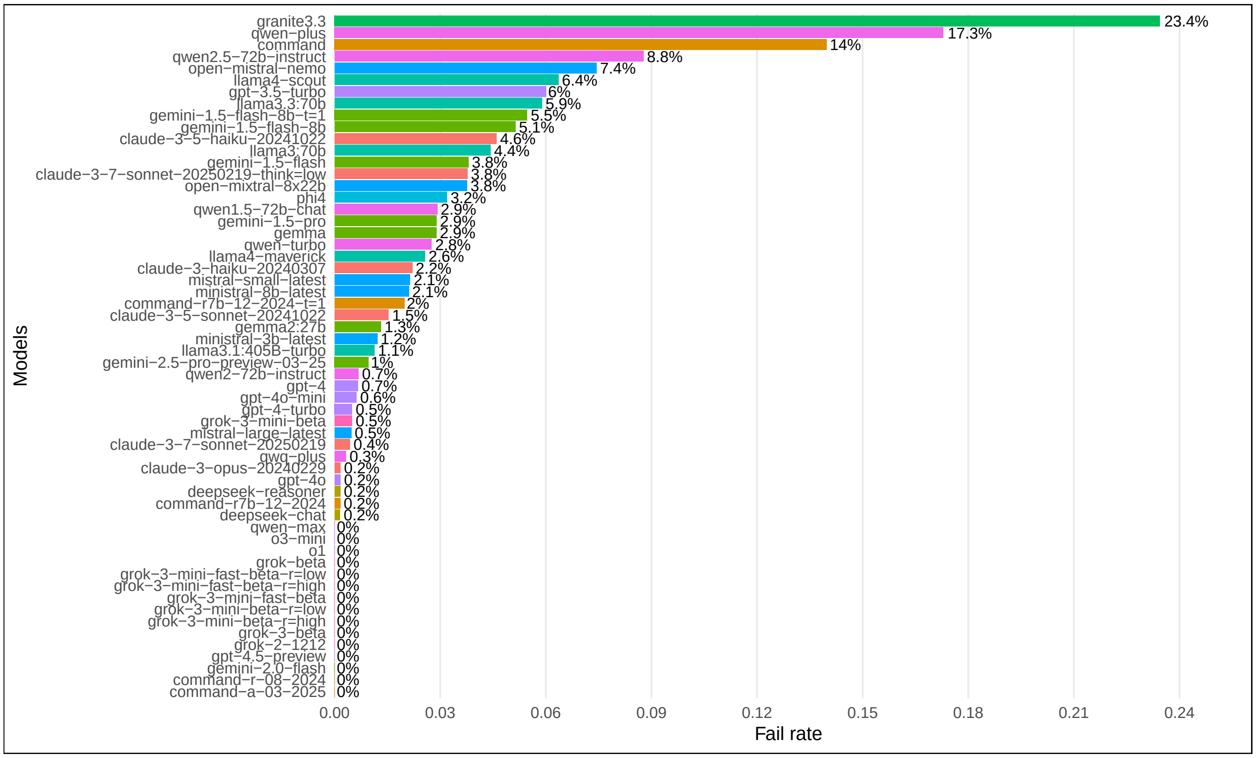
Provider	Model	Reason for exclusion
anthropic	claude-3-sonnet-20240229	not available in Anthropic API anymore
cohere	command-r-plus-08-2024	uniform aggregated considerations (1s)
deepseek	deepseek-v2	high fail rate (85%)
deepseek	deepseek-v2.5	too big to run locally; not available through APIs
google	gemini-2.0-flash-thinking-exp	NA
google	gemma-3-27b-it	low rate limit (15K tokens/min)
google	gemma3:12b	uniform aggregated considerations (1s)
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

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

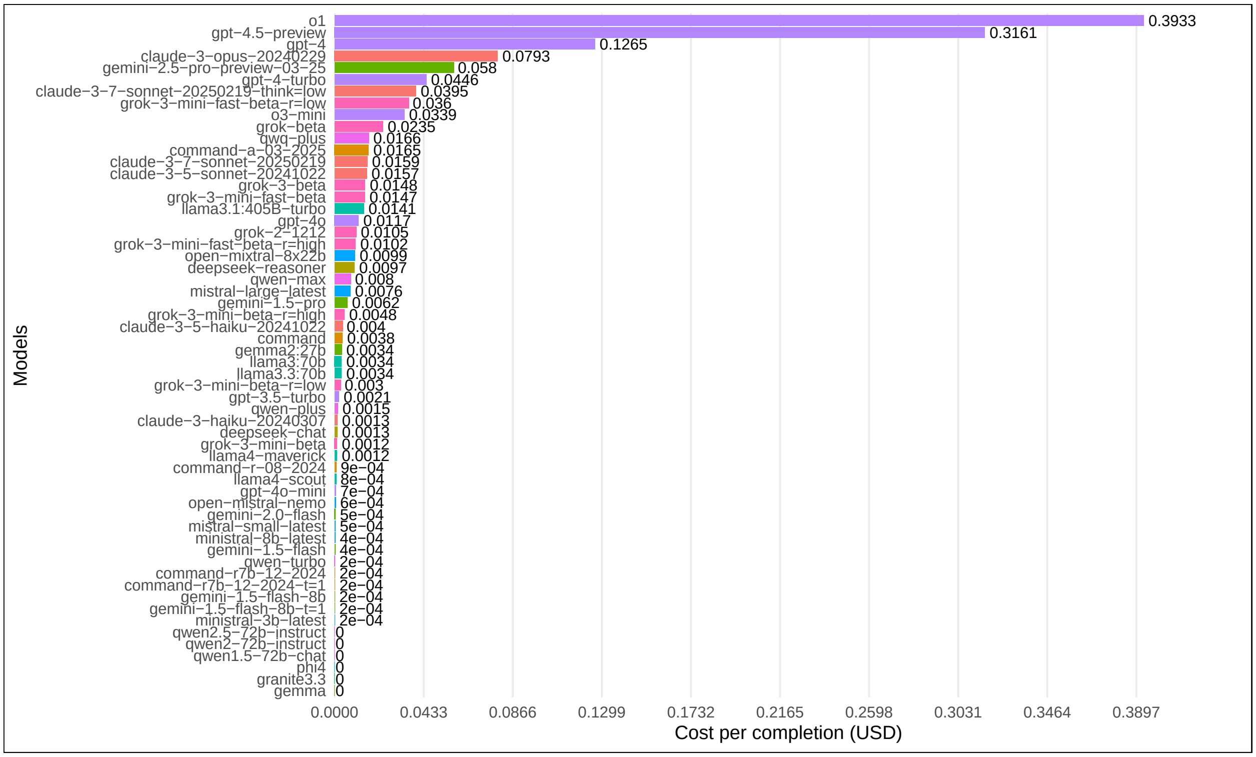
Provider	Model	Reason for exclusion
microsoft	phi2	same model as phi
microsoft	phi3	does not respond to prompts correctly
microsoft	phi3.5	10% success rate for biobanking_wa
mistralai	open-mistral-7b	11% success rate for auscj, uppsala_speaks, and biobanking_wa
mistralai	open-mixtral-8x7b	6% success 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

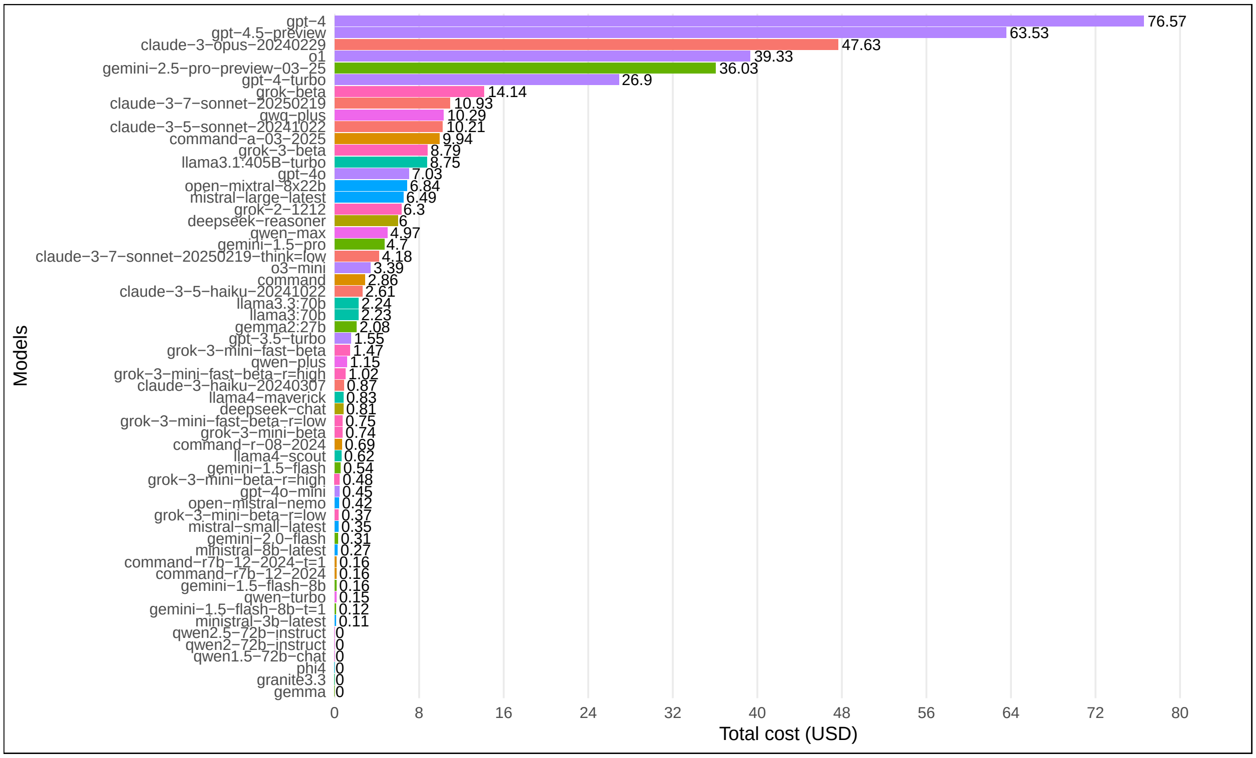
Fail rate



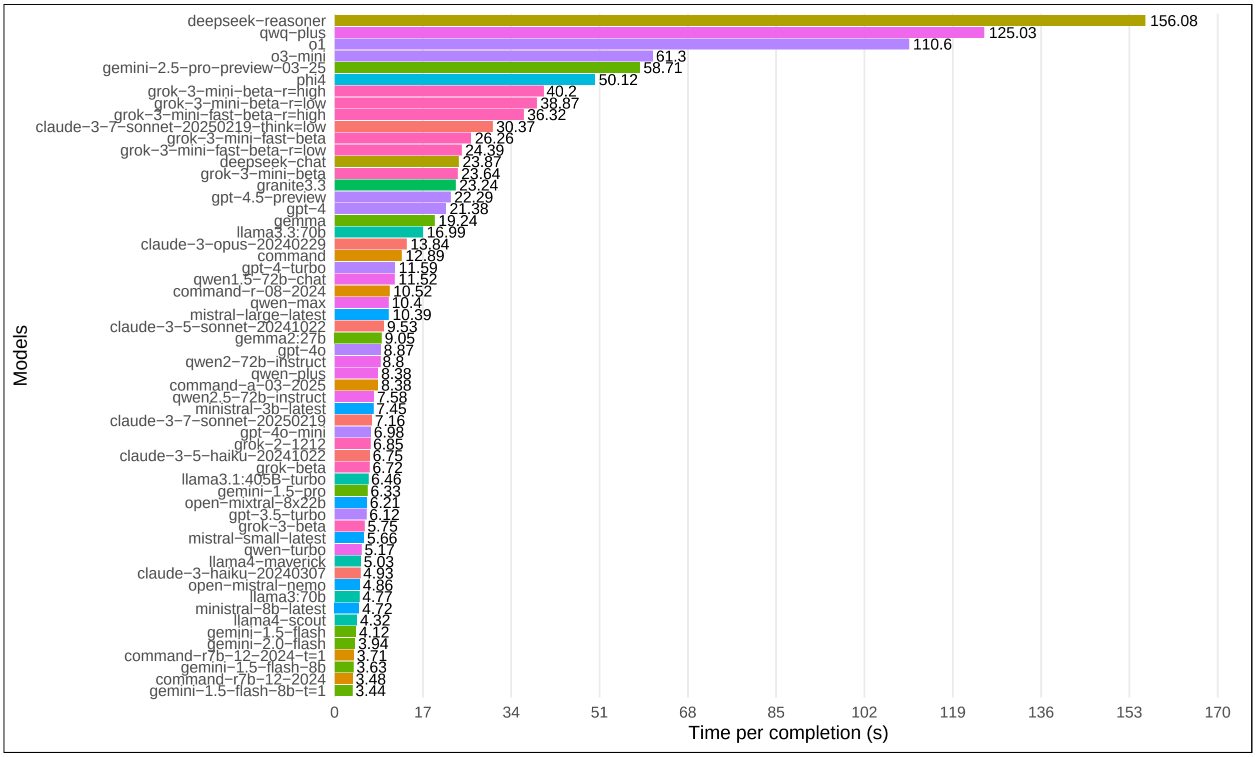
Cost per completion



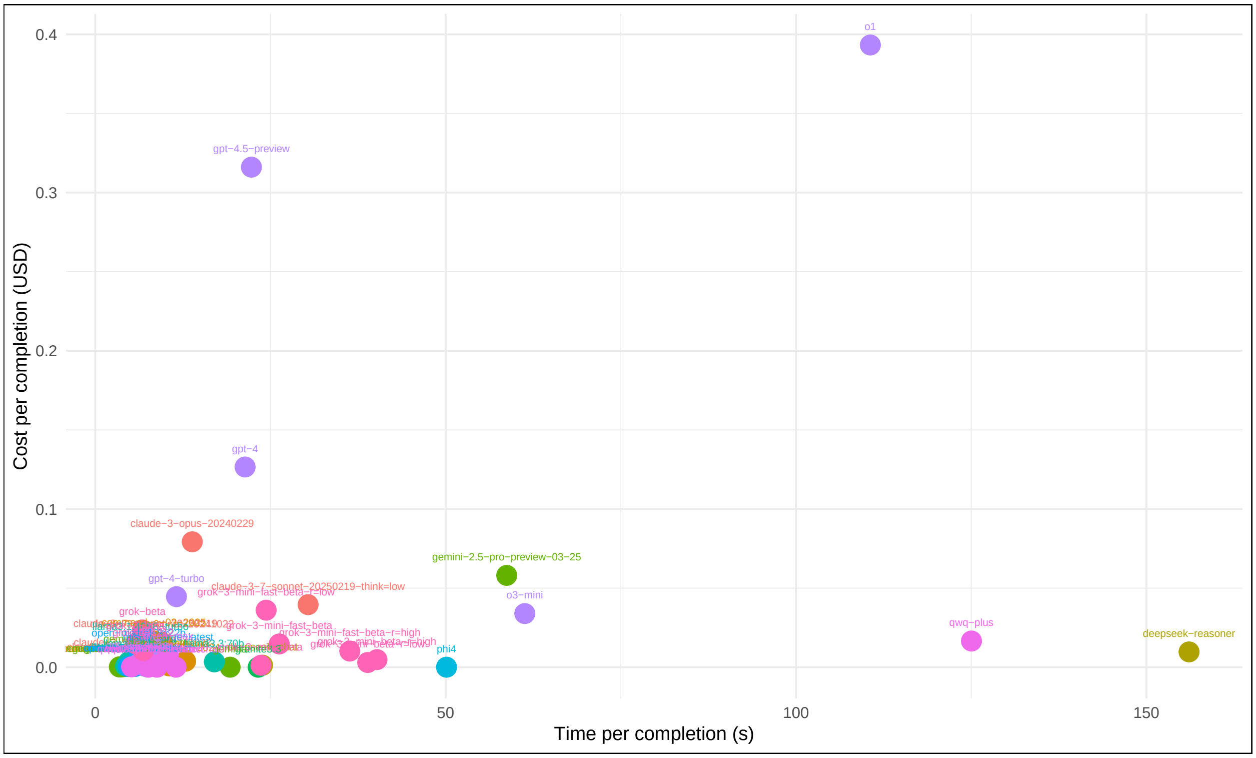
Total cost



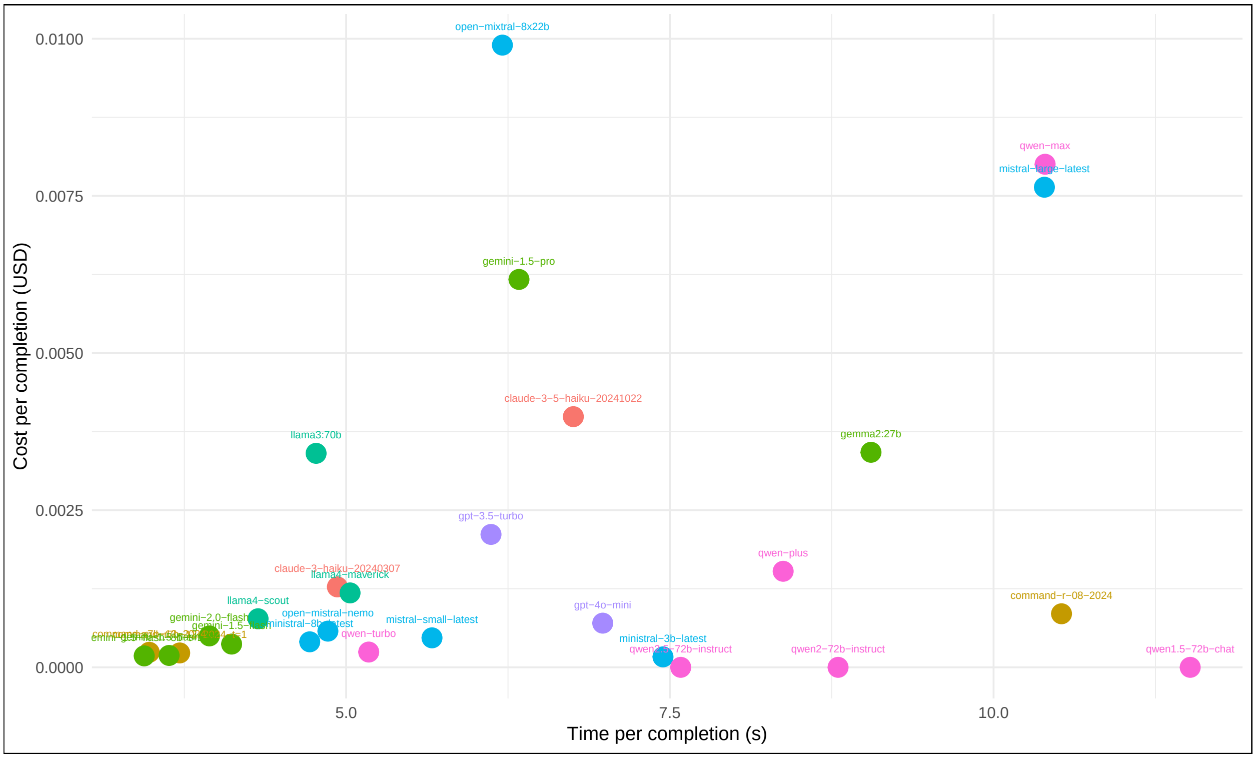
Time per completion



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	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	google	gemini-1.5-flash-8b-t=1	600	0.75	0.81	0.49
10	ibm	granite3.3	600	0.75	0.75	0.47
11	openai	gpt-4	600	0.75	0.82	0.52
12	openai	gpt-4-turbo	600	0.75	0.82	0.53
13	xai	grok-beta	600	0.75	0.85	0.49
14	google	gemini-1.5-pro	600	0.76	0.78	0.57
15	google	gemini-2.5-pro-preview-03-25	600	0.76	0.83	0.67
16	openai	gpt-4o	600	0.76	0.86	0.50
17	cohere	command	600	0.78	0.78	0.44
18	google	gemma	600	0.78	0.80	0.45
19	meta	llama3.3:70b	600	0.78	0.82	0.52
20	mistralai	mistral-small-latest	600	0.78	0.84	0.52

	provider	model	N	all	considerations	policies
21	mistralai	open-mistral-nemo	600	0.78	0.80	0.49
22	qwen	qwq-plus	600	0.78	0.79	0.58
23	xai	grok-2-1212	600	0.78	0.89	0.47
24	cohere	command-a-03-2025	600	0.79	0.86	0.51
25	cohere	command-r-08-2024	600	0.79	0.81	0.50
26	deepseek	deepseek-chat	600	0.79	0.86	0.52
27	google	gemini-1.5-flash-8b	600	0.79	0.84	0.50
28	meta	llama3:70b	600	0.79	0.79	0.52
29	qwen	qwen-turbo	600	0.79	0.83	0.48
30	anthropic	claude-3-7-sonnet-20250219	600	0.80	0.84	0.53
31	meta	llama4-scout	600	0.80	0.85	0.51
32	qwen	qwen-plus	600	0.80	0.82	0.49
33	qwen	qwen2-72b-instruct	600	0.80	0.86	0.48
34	qwen	qwen2.5-72b-instruct	600	0.80	0.84	0.51
35	xai	grok-3-mini-beta	600	0.80	0.78	0.67
36	anthropic	claude-3-5-haiku-20241022	600	0.81	0.86	0.47
37	microsoft	phi4	600	0.81	0.82	0.55
38	xai	grok-3-beta	600	0.81	0.84	0.53
39	mistralai	ministral-8b-latest	600	0.82	0.83	0.51
40	qwen	qwen-max	600	0.82	0.84	0.51
41	anthropic	claude-3-opus-20240229	600	0.83	0.87	0.50
42	mistralai	mistral-large-latest	600	0.83	0.86	0.54
43	google	gemini-2.0-flash	600	0.84	0.84	0.62
44	openai	gpt-3.5-turbo	600	0.84	0.87	0.48
45	openai	gpt-4.5-preview	201	0.84	0.87	0.70
46	cohere	command-r7b-12-2024-t=1	600	0.85	0.86	0.47
47	meta	llama3.1:405B-turbo	600	0.85	0.88	0.49
48	mistralai	ministral-3b-latest	600	0.85	0.86	0.53
49	cohere	command-r7b-12-2024	600	0.86	0.87	0.46
50	mistralai	open-mixtral-8x22b	600	0.87	0.90	0.52
51	anthropic	claude-3-7-sonnet-20250219-think=low	102	0.89	0.90	0.76
52	xai	grok-3-mini-beta-r=high	100	0.91	0.90	0.81
53	xai	grok-3-mini-beta-r=low	124	0.91	0.89	0.80
54	xai	grok-3-mini-fast-beta	100	0.91	0.89	0.86
55	xai	grok-3-mini-fast-beta-r=high	100	0.91	0.90	0.84
56	xai	grok-3-mini-fast-beta-r=low	202	0.91	0.89	0.81
57	openai	o1	100	0.92	0.92	0.77
58	openai	o3-mini	100	0.92	0.91	0.80

Human Data

Handle Swiss cases

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(swiss_C_cols)
##
##   # Now:
##   data %>% select(all_of(swiss_C_cols))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(col)
##
##   # Now:
##   data %>% select(all_of(col))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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	CA	swiss_health	56
20	Activate	uppsala_speaks	26
21	Standard	uppsala_speaks	22
22	UPSA Control Group	uppsala_speaks	20
23	Valsamoggia	valsamoggia	16
24	Thalwill	zh_thalwil	14
25	USTER	zh_uster	15
26	Winterthur	zh_winterthur	16
27	Zukunft	zukunft	63

We collected 1144 human responses across 27 case studies, including pre-post deliberation responses.

Excluded cases

Table 7: Excluded cases

	Case	Survey	Participants	Excluded Reason
1	HGE Control Group	auscj	19	control group, no deliberation
2	GBR	gbr	7	unclear if human survey data is accurate
3	UPSA Control Group	uppsala_speaks	20	control group, no deliberation

We excluded 3 cases due to the reasons listed above.

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-mini-fast-beta / acp
## updating: grok-3-mini-fast-beta / auscj
## updating: grok-3-mini-fast-beta / bep
## updating: grok-3-mini-fast-beta / biobanking_mayo_ubc
## updating: grok-3-mini-fast-beta / biobanking_wa
## updating: grok-3-mini-fast-beta / ccps
## updating: grok-3-mini-fast-beta / ds_aargau
## updating: grok-3-mini-fast-beta / ds_bellinzona
## updating: grok-3-mini-fast-beta / energy_futures
## updating: grok-3-mini-fast-beta / fnqcj
## updating: grok-3-mini-fast-beta / forestera
## updating: grok-3-mini-fast-beta / fremantle
## updating: grok-3-mini-fast-beta / gbr
## updating: grok-3-mini-fast-beta / swiss_health
## updating: grok-3-mini-fast-beta / uppsala_speaks
## updating: grok-3-mini-fast-beta / valsamoggia
## updating: grok-3-mini-fast-beta / zh_thalwil
## updating: grok-3-mini-fast-beta / zh_uster
## updating: grok-3-mini-fast-beta / zh_winterthur
## updating: grok-3-mini-fast-beta / zukunft
```

We aggregated 33572 LLM responses into 1160 responses: 1 response per model per survey.

Randomly Generated Data

Then, we generated 20 random reseponses, one 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) {
  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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 in cor(Q, method = "spearman"): the standard deviation is zero
## Warning in cor(Q, method = "spearman"): the standard deviation is zero

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

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

Select Dependent Variable for Analysis

We are using the average DRI calculated across the iterations of LLM (DRIIndV3_mean).

Hypotheses Testing

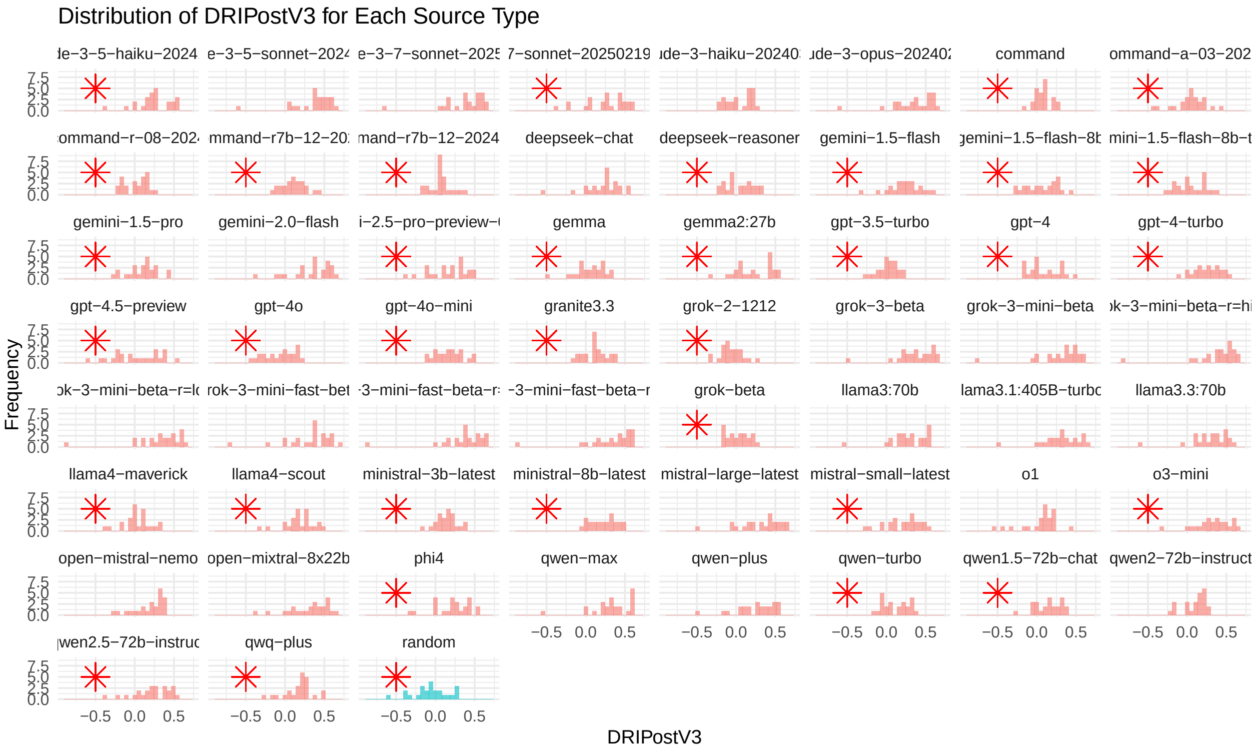
H1. DRI scores of LLMs do not significantly differ from those produced by a random generation process.

Testing assumptions

We employed a one-way ANOVA (or a Kruskal-Wallis test, depending on the results of the exploratory analysis) between subjects to analyze our results. If normality and homogeneity of variance assumptions are met, we will use ANOVA followed by Tukey's HSD post-hoc test for pairwise comparisons between LLM/version DRI and random DRI. If assumptions are violated, we will use the non-parametric Kruskal-Wallis test, followed by Dunn's post-hoc test with Bonferroni correction.

The independent variable is be the type of participant (e.g., random, model). The dependent variable is the individual-level DRI score.

```
## Adding missing grouping variables: `provider`, `model`
```



Testing hypothesis

```
##
```

```
##  Kruskal-Wallis rank sum test
##
## data:  DRIPostV3 by source
## Kruskal-Wallis chi-squared = 442.14, df = 58, p-value < 2.2e-16
```

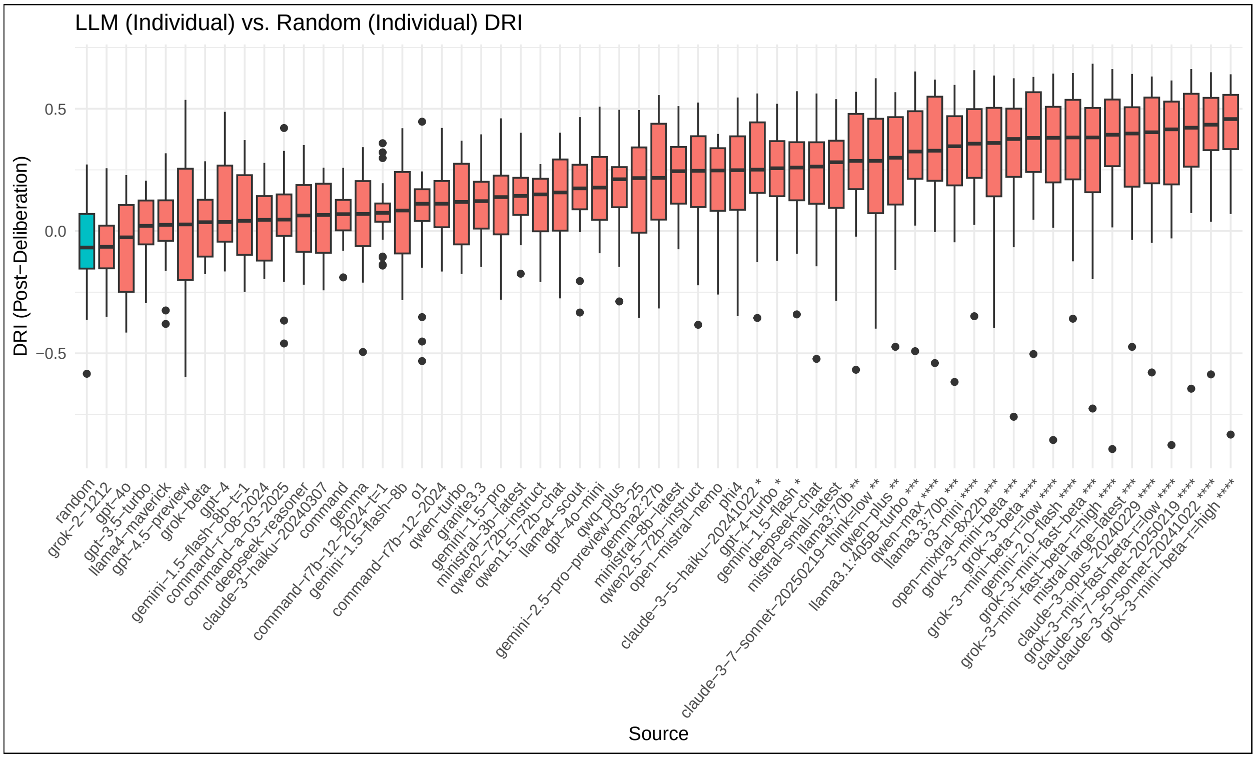
Post-hoc tests

```
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Table 8: Models compared to random

Model	P-adjusted
grok-3-mini-beta-r=high	0*
claude-3-7-sonnet-20250219	0*
claude-3-5-sonnet-20241022	0*
grok-3-beta	0*
grok-3-mini-fast-beta-r=high	0*
claude-3-opus-20240229	1e-05*
grok-3-mini-fast-beta-r=low	1e-05*
grok-3-mini-beta-r=low	2e-05*
gemini-2.0-flash	4e-05*
o3-mini	4e-05*
qwen-max	7e-05*
grok-3-mini-fast-beta	1e-04*
grok-3-mini-beta	1e-04*
llama3.1:405B-turbo	1e-04*
mistral-large-latest	0.00014*
open-mixtral-8x22b	4e-04*
llama3.3:70b	0.00049*
llama3:70b	0.00148*
qwen-plus	0.00543*
claude-3-7-sonnet-20250219-think=low	0.00718*
claude-3-5-haiku-20241022	0.02478*
gpt-4-turbo	0.02505*
gemini-1.5-flash	0.04275*
deepseek-chat	0.05812
ministral-8b-latest	0.07759
qwen2.5-72b-instruct	0.08506
phi4	0.09375
mistral-small-latest	0.10297
gemma2:27b	0.14073
open-mistral-nemo	0.63749
claude-3-haiku-20240307	1
command	1
command-a-03-2025	1
command-r-08-2024	1
command-r7b-12-2024	1
command-r7b-12-2024-t=1	1
deepseek-reasoner	1
gemini-1.5-flash-8b	1
gemini-1.5-flash-8b-t=1	1
gemini-1.5-pro	1

Model	P-adjusted
gemini-2.5-pro-preview-03-25	1
gemma	1
gpt-3.5-turbo	1
gpt-4	1
gpt-4.5-preview	1
gpt-4o	1
gpt-4o-mini	1
granite3.3	1
grok-2-1212	1
grok-beta	1
llama4-maverick	1
llama4-scout	1
ministral-3b-latest	1
o1	1
qwen-turbo	1
qwen1.5-72b-chat	1
qwen2-72b-instruct	1
qwq-plus	1

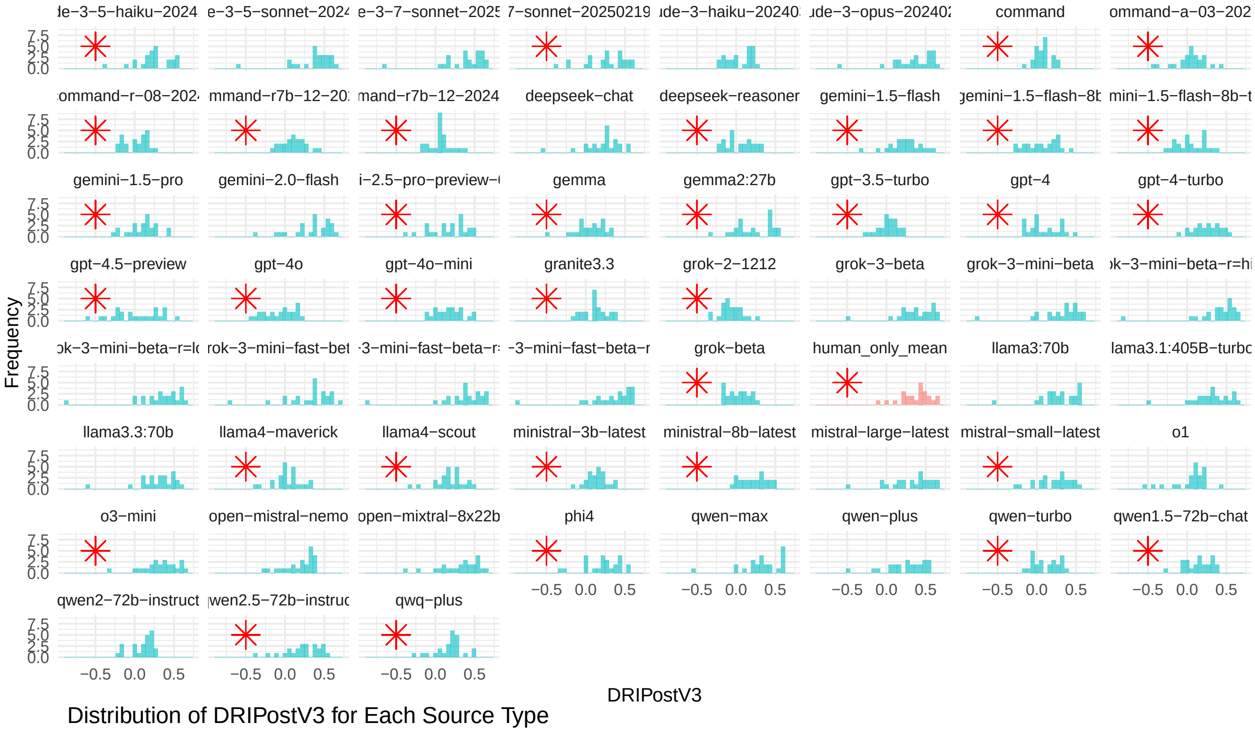


Some models, 23 out of 58, are significantly different than random.

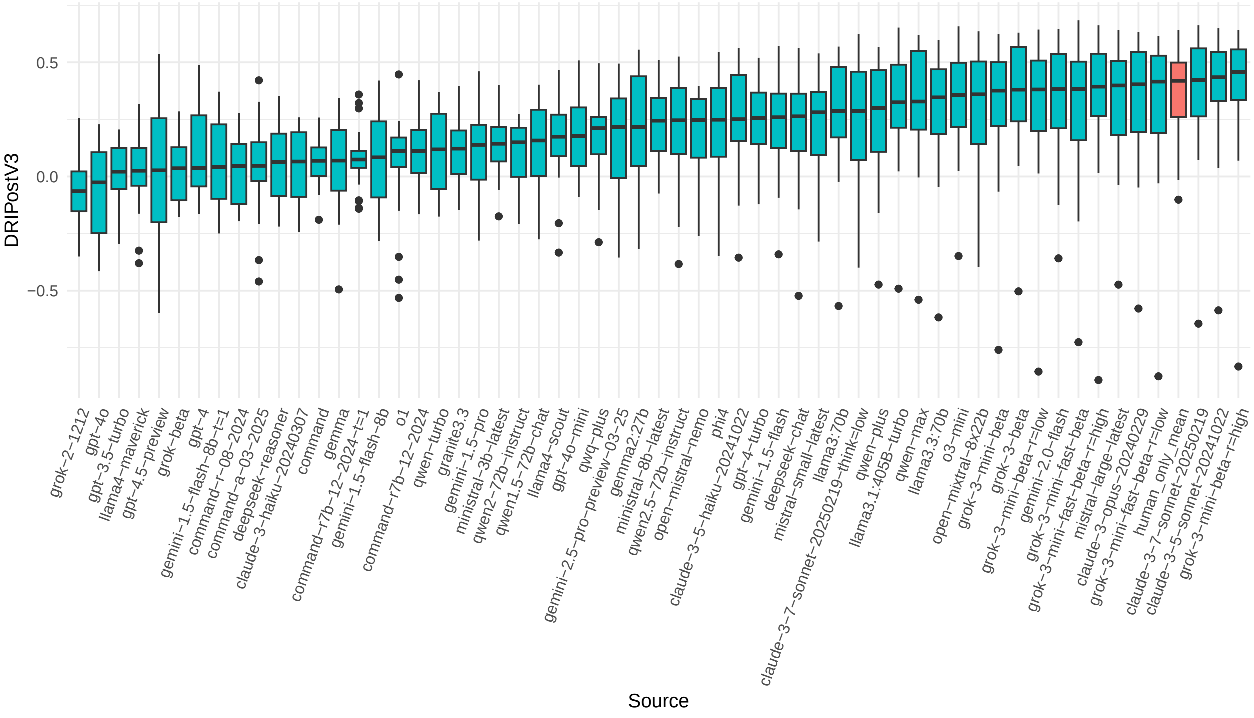
H2. LLMs' DRI scores will be significantly lower than those obtained from human participants after deliberation.

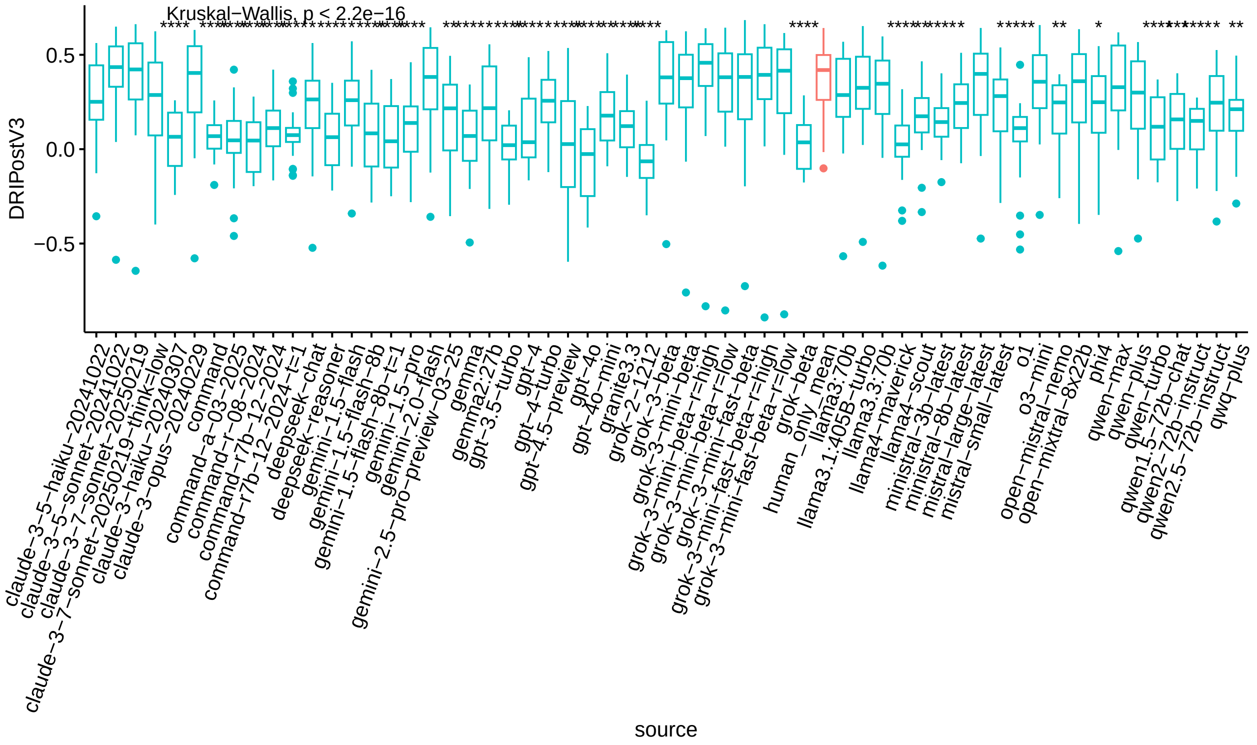
Testing assumptions

Distribution of DRIPostV3 for Each Source Type



Distribution of DRIPostV3 for Each Source Type





Testing hypothesis

To test H2, we will compare the average individual-level, post-deliberation DRI scores obtained by human participants with the individual-level DRI scores obtained by LLMs both across case studies and across LLM/version.

First, for each case study, we will employ a t-test (or non-parametric equivalent, depending on the results of the exploratory analysis) to analyze our results across case studies. The independent variable is participant type (human-only vs. LLM) and the dependent variable is the individual-level DRI scores.

For each case study…

human average

Second, for each LLM/version, we will employ a t-test (or non-parametric equivalent, depending on the results of the exploratory analysis) to analyze our results across LLM/version. The independent variable is participant type (human-only vs. LLM/version) and the dependent variable is the individual-level DRI scores.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  DRIPostV3 by source
## Kruskal-Wallis chi-squared = 437.87, df = 58, p-value < 2.2e-16
```

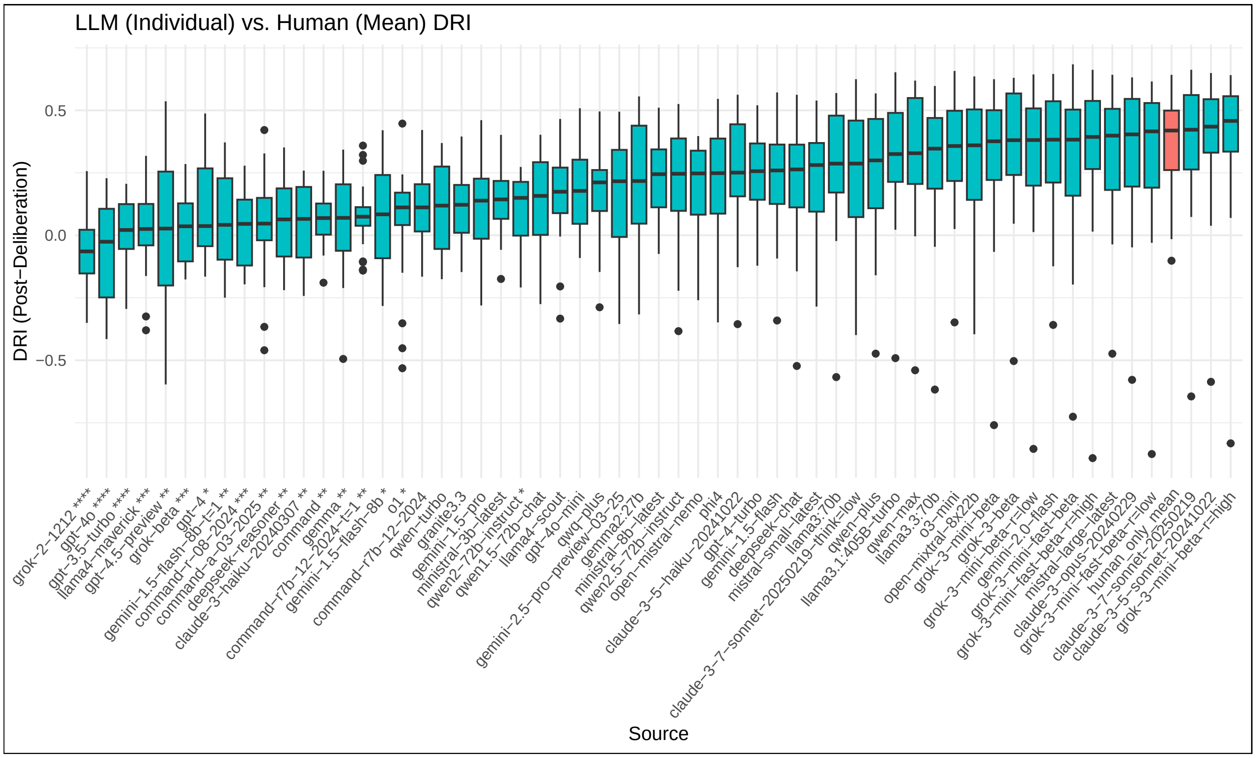
Post-hoc tests

```
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Table 9: Models compared to human

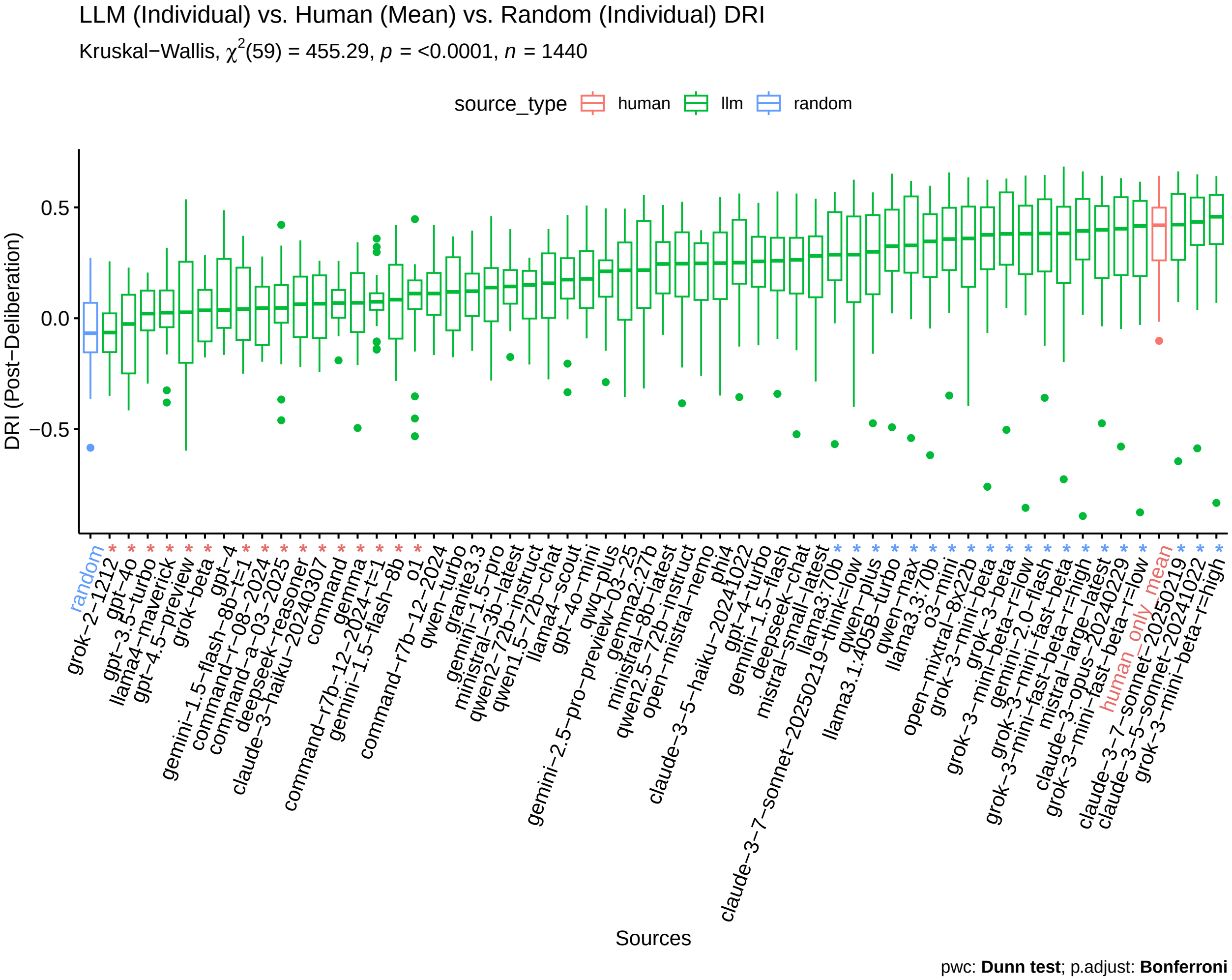
Model	P-adjusted
grok-2-1212	1.80249306139143e-07*
gpt-4o	3.31357802524231e-06*
gpt-3.5-turbo	6.34158704912946e-05*
grok-beta	0.000135524776249451*
command-r-08-2024	0.000174618206032383*
llama4-maverick	0.000460829800658591*
command-a-03-2025	0.00132288150241591*
claude-3-haiku-20240307	0.00137762257056516*
deepseek-reasoner	0.00149370898799689*
command	0.00235657462344917*
gemini-1.5-flash-8b-t=1	0.00320263891162993*
command-r7b-12-2024-t=1	0.00373764138824981*
gemma	0.00430669142086903*
gpt-4.5-preview	0.00833229182663337*
o1	0.0109997319195447*
gemini-1.5-flash-8b	0.015893626293051*
gpt-4	0.0326596801055083*
qwen2-72b-instruct	0.0413502737450386*
command-r7b-12-2024	0.0525169999715211
gemini-1.5-pro	0.087569210391808
qwen-turbo	0.129994827876516
granite3.3	0.157743406547145
ministral-3b-latest	0.453222848457356
claude-3-5-haiku-20241022	1
claude-3-5-sonnet-20241022	1
claude-3-7-sonnet-20250219	1
claude-3-7-sonnet-20250219-think=low	1
claude-3-opus-20240229	1
deepseek-chat	1
gemini-1.5-flash	1
gemini-2.0-flash	1
gemini-2.5-pro-preview-03-25	1
gemma2:27b	1
gpt-4-turbo	1
gpt-4o-mini	1
grok-3-beta	1
grok-3-mini-beta	1
grok-3-mini-beta-r=high	1
grok-3-mini-beta-r=low	1
grok-3-mini-fast-beta	1
grok-3-mini-fast-beta-r=high	1
grok-3-mini-fast-beta-r=low	1
llama3:70b	1
llama3.1:405B-turbo	1
llama3.3:70b	1
llama4-scout	1
ministral-8b-latest	1
mistral-large-latest	1
mistral-small-latest	1
o3-mini	1

Model	P-adjusted
open-mistral-nemo	1
open-mixtral-8x22b	1
phi4	1
qwen-max	1
qwen-plus	1
qwen1.5-72b-chat	1
qwen2.5-72b-instruct	1
qwq-plus	1



H1 and H2

```
## # A tibble: 1 x 6
##   .y.            n statistic    df         p method
## * <chr>      <int>     <dbl> <int>     <dbl> <chr>
## 1 DRIPostV3   1440      455.    59 1.45e-62 Kruskal-Wallis
```



H3. LLMs' DRI scores are improving over time, across each version.

Random slope –

Assume each case Multilevel analysis – each case behave differently

LMER –

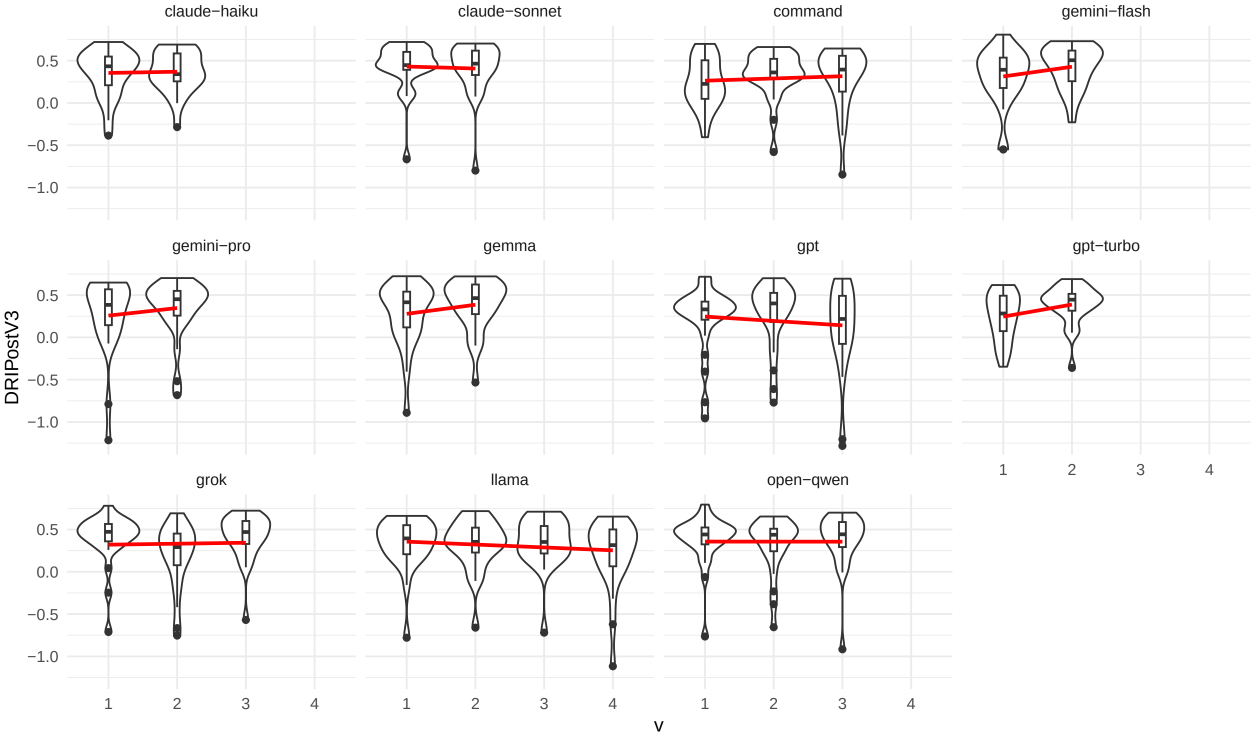
To test H3, we will conduct a repeated measures ANOVA (or Friedman test if the assumptions of normality or sphericity are violated) to test for differences in the mean DRI across all versions (e.g., v1, v2, v3) of an LLM across each case study. We will treat different LLM versions as related groups and the individual-level LLM DRI in each case study as a subject. In this within-subjects design, we can assess whether more recent versions of LLMs have a significant impact on the DRI scores they produce.

We want to assess the effects of Case and Series on weight loss in 10 sedentary individuals.

Dependent variable: - DRIPostV3

Independent variables: - [LLM series (moderator) – which llm?] - [case (moderator) – which case?] – [LATER] - version (focal)

```
## `geom_smooth()` using formula = 'y ~ x'
```

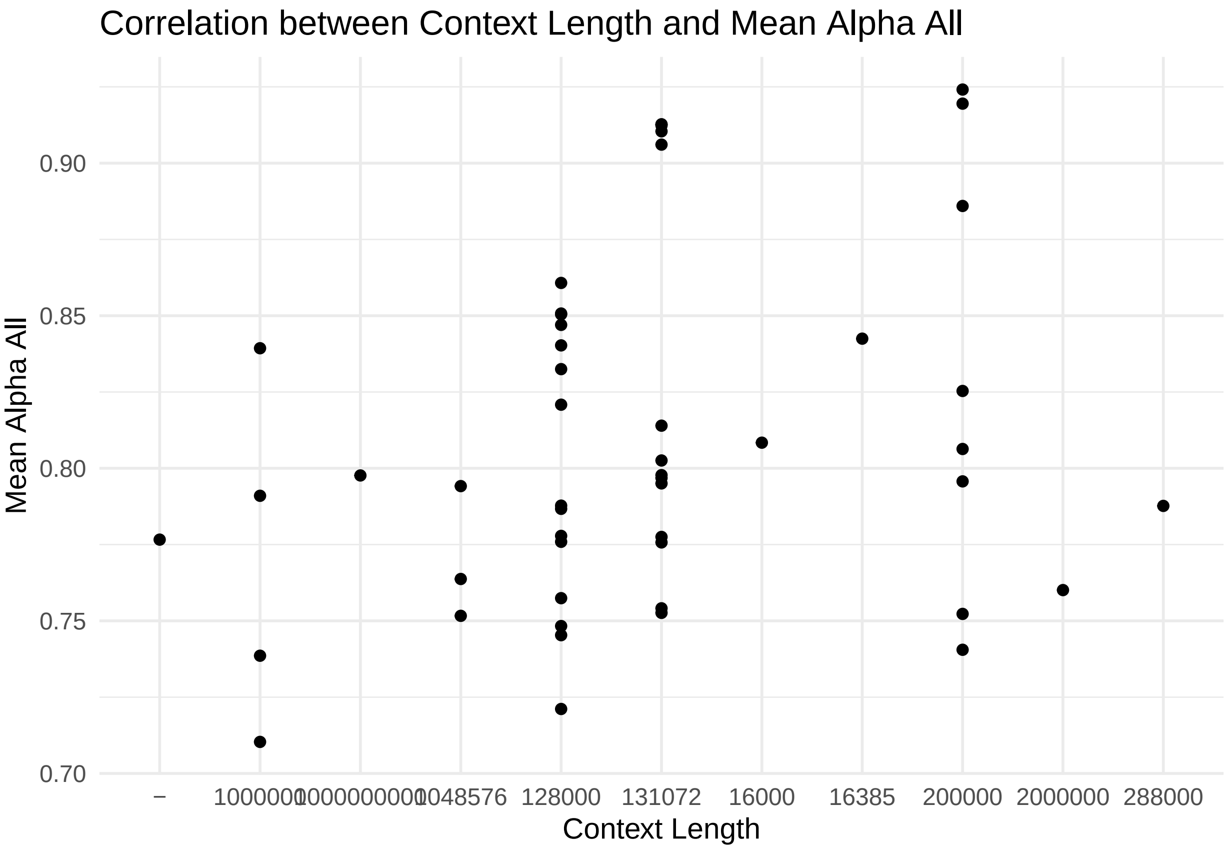


series	p	method
gemini-flash	0.0412268	Friedman test
command	0.1969117	Friedman test
grok	0.1969117	Friedman test
gemma	0.2206714	Friedman test
gpt-turbo	0.2206714	Friedman test
claude-haiku	0.4142162	Friedman test
gemini-pro	0.4142162	Friedman test
open-qwen	0.5134171	Friedman test
llama	0.5418638	Friedman test
claude-sonnet	0.6830914	Friedman test
gpt	0.6872893	Friedman test

If a significant difference is found, we will conduct a post-hoc analysis using paired t-tests (or Wilcoxon signed-rank tests) for pairwise comparisons, with adjustments for multiple comparisons.

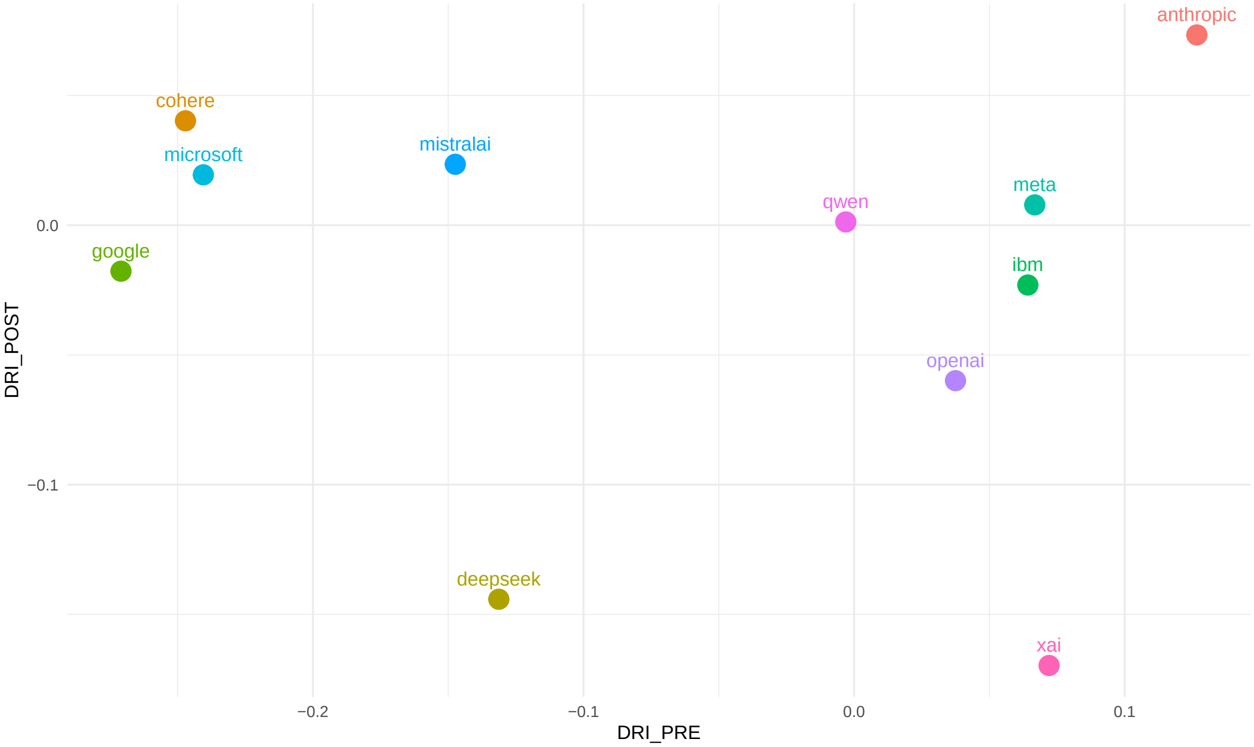
DRI Benchmark

```
## `geom_smooth()` using formula = 'y ~ x'
```

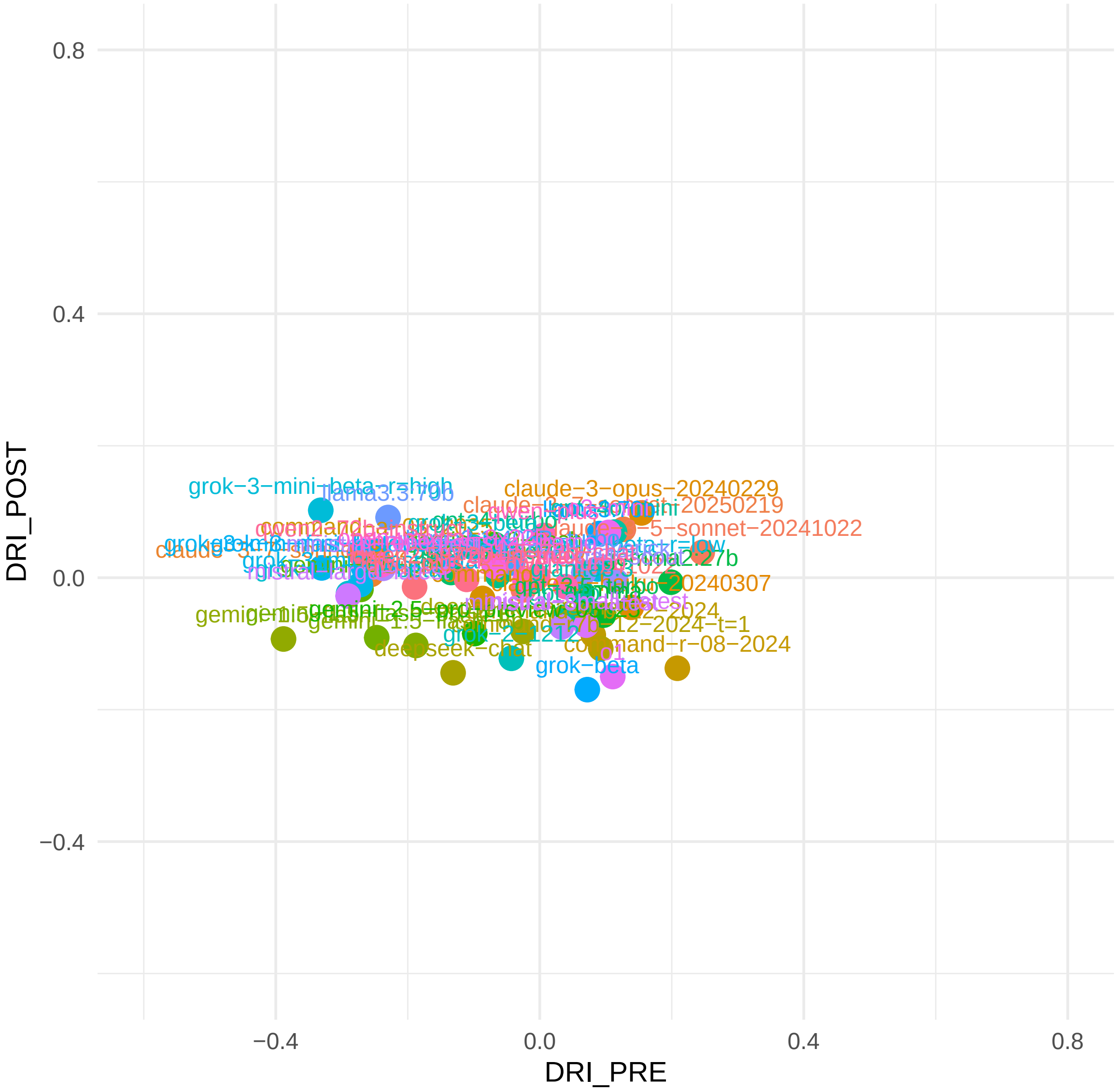


```
## `summarise()` has grouped output by 'provider', 'model'. You can override using
## the `.groups` argument.
```

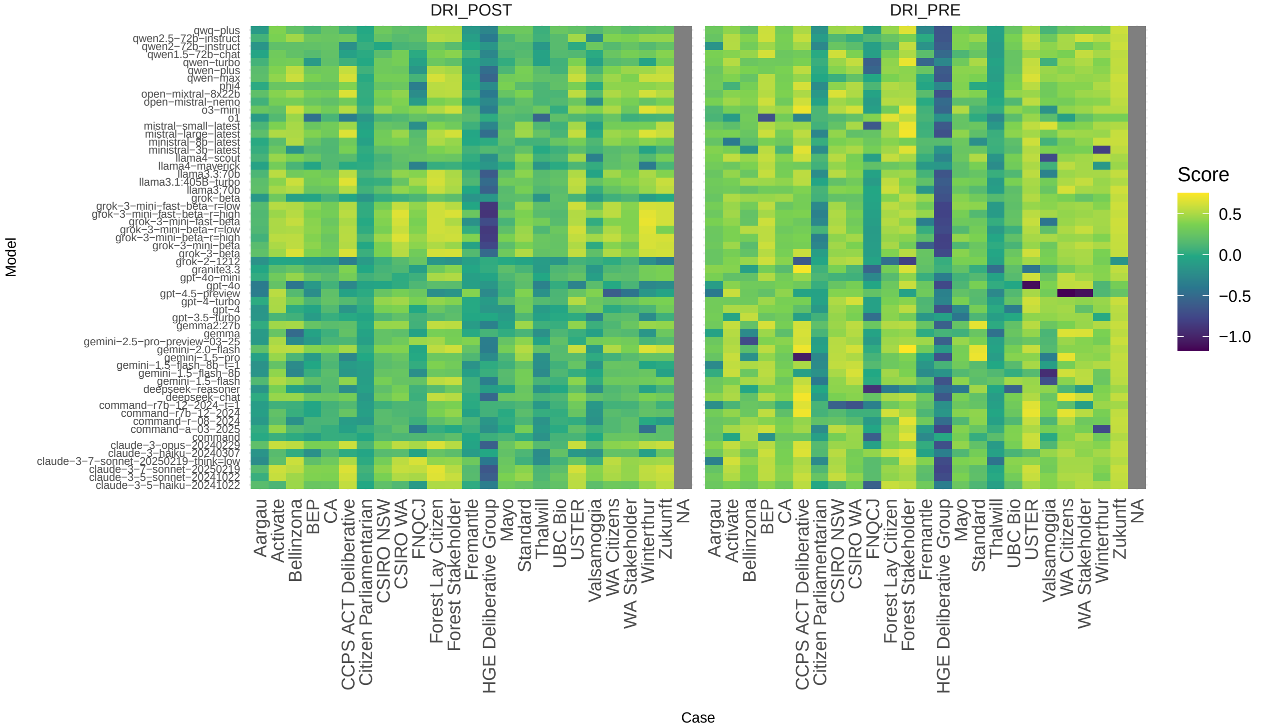
Comparison PRE and POST DRI by Provider



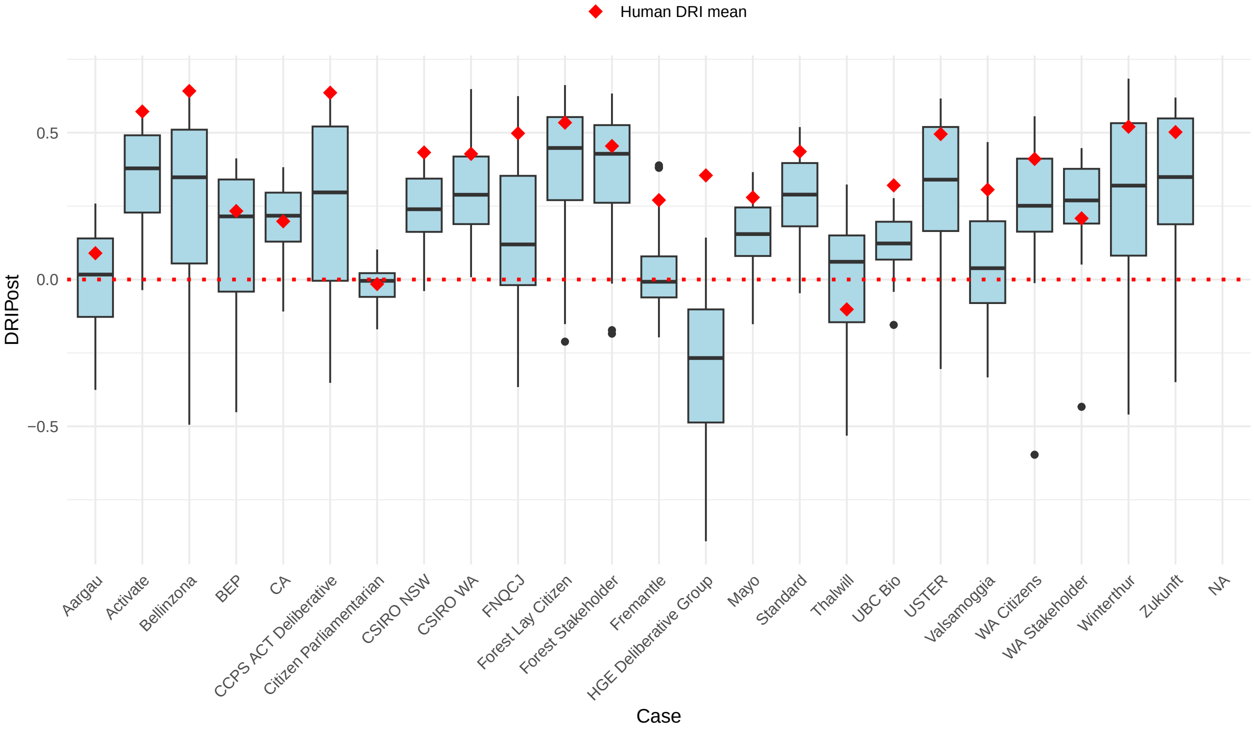
Comparison PRE and POST DRI by Model



Heatmap of DRI Scores by Case and Model



Boxplot of LLM DRI Post by Case



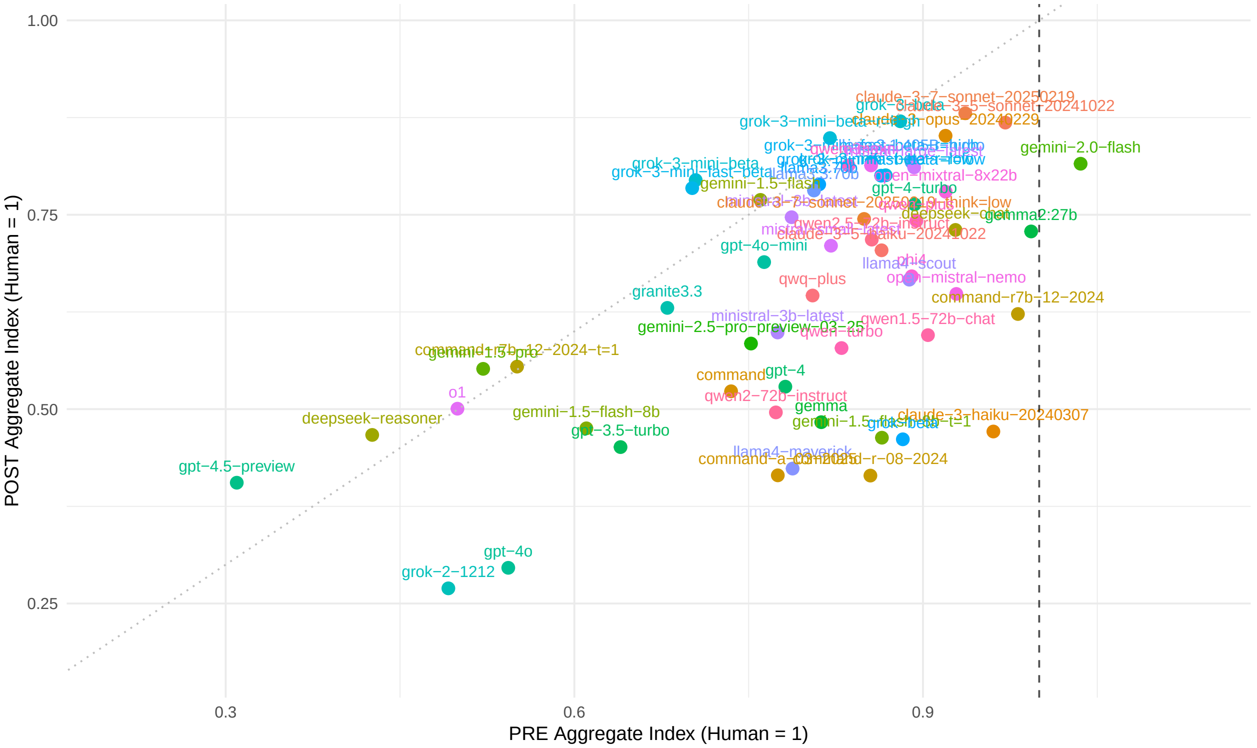
LLM Performance Metrics Against Human DRI Post-Scores

Table 11: LLM Performance Metrics Against Human DRI Post-Scores

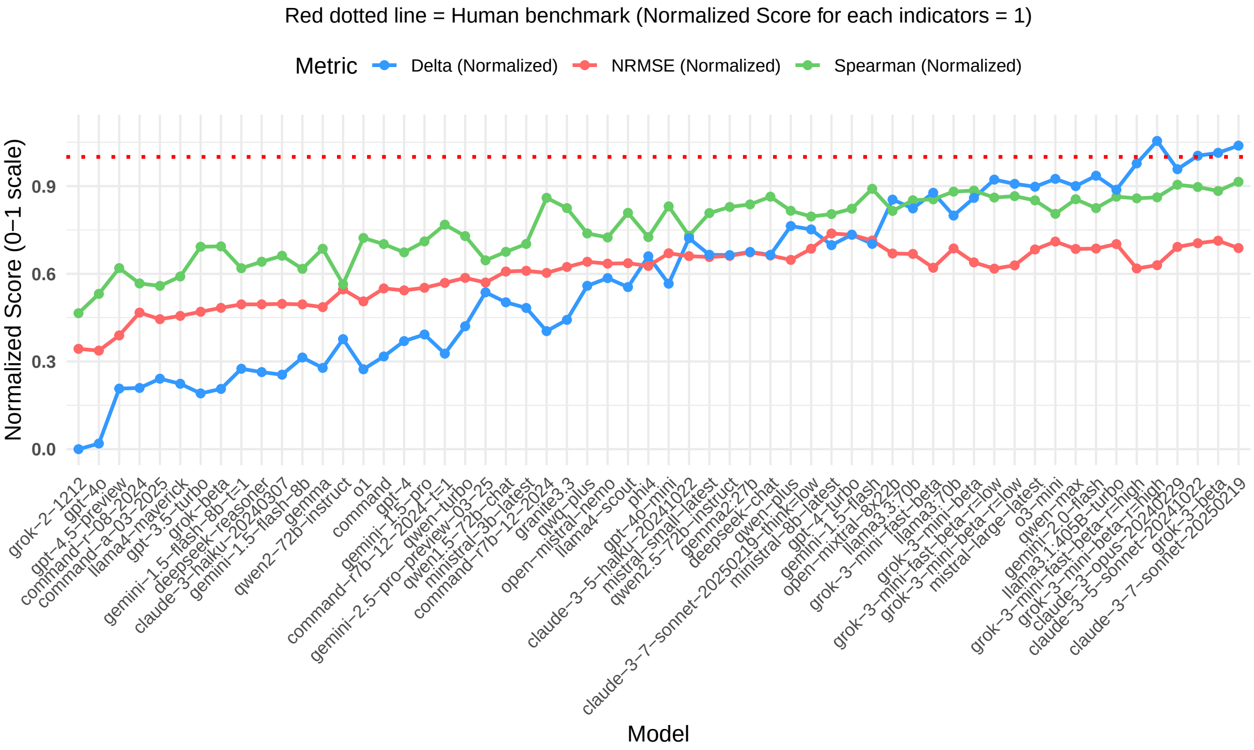
Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
ministral-8b-latest	0.155	0.195	62.849	0.744	0.209	0.262	0.609	-0.129
gpt-4-turbo	0.151	0.198	65.236	0.744	0.203	0.267	0.645	-0.114
gemini-1.5-flash	0.163	0.213	69.953	0.744	0.219	0.286	0.782	-0.127
grok-3-beta	0.136	0.213	68.669	0.744	0.183	0.287	0.767	0.006
o3-mini	0.156	0.215	81.443	0.744	0.210	0.290	0.610	-0.032
claude-3-5-sonnet-20241022	0.125	0.220	59.621	0.744	0.168	0.295	0.794	0.002
llama3.1:405B-turbo	0.142	0.222	61.675	0.744	0.191	0.298	0.728	-0.048
claude-3-opus-20240229	0.137	0.229	82.329	0.744	0.184	0.308	0.810	-0.018
claude-3-7-sonnet-20250219	0.133	0.232	71.006	0.744	0.179	0.312	0.830	0.016
llama3:70b	0.150	0.233	74.105	0.744	0.202	0.313	0.763	-0.086
gemini-2.0-flash	0.162	0.233	67.816	0.744	0.218	0.314	0.650	-0.027
claude-3-7-sonnet-20250219-think=low	0.166	0.234	64.740	0.744	0.224	0.314	0.592	-0.106
qwen-max	0.149	0.234	57.339	0.744	0.201	0.315	0.710	-0.043
mistral-large-latest	0.159	0.235	62.932	0.744	0.214	0.317	0.703	-0.044
gemma2:27b	0.169	0.243	54.209	0.744	0.227	0.327	0.675	-0.139
gpt-4o-mini	0.205	0.245	77.398	0.744	0.276	0.330	0.662	-0.185
open-mixtral-8x22b	0.156	0.246	63.580	0.744	0.210	0.331	0.630	-0.062
llama3.3:70b	0.150	0.247	78.423	0.744	0.202	0.332	0.703	-0.075
deepseek-chat	0.177	0.251	85.957	0.744	0.238	0.337	0.728	-0.143
qwen2.5-72b-instruct	0.183	0.252	58.312	0.744	0.246	0.339	0.657	-0.143
claude-3-5-haiku-20241022	0.172	0.252	54.449	0.744	0.231	0.339	0.461	-0.119
mistral-small-latest	0.188	0.255	71.347	0.744	0.253	0.342	0.616	-0.143
qwen-plus	0.176	0.262	81.573	0.744	0.237	0.352	0.631	-0.101
qwq-plus	0.207	0.267	62.292	0.744	0.279	0.359	0.477	-0.188
grok-3-mini-beta	0.153	0.268	59.971	0.744	0.205	0.361	0.770	-0.060
llama4-scout	0.212	0.271	60.900	0.744	0.285	0.364	0.617	-0.190
open-mistral-nemo	0.211	0.272	68.387	0.744	0.283	0.365	0.449	-0.177
grok-3-mini-beta-r=high	0.155	0.276	85.964	0.744	0.208	0.371	0.723	0.023
grok-3-mini-beta-r=low	0.158	0.276	60.158	0.744	0.212	0.371	0.731	-0.039
phi4	0.199	0.277	68.061	0.744	0.267	0.373	0.451	-0.145
granite3.3	0.238	0.280	65.064	0.744	0.320	0.377	0.650	-0.238
grok-3-mini-fast-beta	0.176	0.282	67.153	0.744	0.236	0.379	0.710	-0.052
grok-3-mini-fast-beta-r=high	0.157	0.284	60.918	0.744	0.211	0.382	0.717	-0.009
grok-3-mini-fast-beta-r=low	0.163	0.284	66.056	0.744	0.219	0.382	0.722	-0.033
ministral-3b-latest	0.231	0.290	65.515	0.744	0.310	0.390	0.404	-0.221
qwen1.5-72b-chat	0.230	0.292	64.693	0.744	0.309	0.392	0.350	-0.212
command-r7b-12-2024	0.265	0.295	95.817	0.744	0.357	0.397	0.720	-0.255
qwen-turbo	0.251	0.308	63.909	0.744	0.337	0.414	0.458	-0.247
gemini-2.5-pro-preview-03-25	0.220	0.320	80.122	0.744	0.296	0.430	0.292	-0.198
command-r7b-12-2024-t=1	0.287	0.321	102.722	0.744	0.387	0.431	0.537	-0.287
gemini-1.5-pro	0.262	0.333	71.434	0.744	0.352	0.448	0.422	-0.259
command	0.296	0.335	81.385	0.744	0.398	0.450	0.403	-0.292
qwen2-72b-instruct	0.276	0.337	86.093	0.744	0.371	0.453	0.129	-0.266
gpt-4	0.271	0.340	82.377	0.744	0.364	0.457	0.347	-0.269
o1	0.310	0.368	132.473	0.744	0.417	0.494	0.445	-0.310
claude-3-haiku-20240307	0.319	0.374	97.381	0.744	0.429	0.503	0.323	-0.318
deepseek-reasoner	0.314	0.375	103.805	0.744	0.423	0.505	0.283	-0.314
gemini-1.5-flash-8b	0.295	0.375	103.753	0.744	0.396	0.505	0.234	-0.293

Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
gemini-1.5-flash-8b-t=1	0.312	0.375	107.329	0.744	0.419	0.505	0.238	-0.310
gemma	0.308	0.382	95.375	0.744	0.415	0.514	0.371	-0.308
grok-beta	0.339	0.384	131.998	0.744	0.456	0.516	0.388	-0.339
gpt-3.5-turbo	0.348	0.394	103.851	0.744	0.468	0.530	0.385	-0.346
command-r-08-2024	0.338	0.396	120.165	0.744	0.454	0.533	0.134	-0.338
llama4-maverick	0.336	0.405	92.630	0.744	0.452	0.544	0.182	-0.331
command-a-03-2025	0.329	0.413	95.028	0.744	0.442	0.555	0.117	-0.324
gpt-4.5-preview	0.357	0.454	114.327	0.744	0.480	0.611	0.239	-0.339
grok-2-1212	0.427	0.488	140.106	0.744	0.574	0.657	-0.070	-0.427
gpt-4o	0.419	0.493	132.587	0.744	0.563	0.663	0.063	-0.419

PRE vs. POST Aggregate Scores Correlation Across LLMs



Human-Normalized Performance

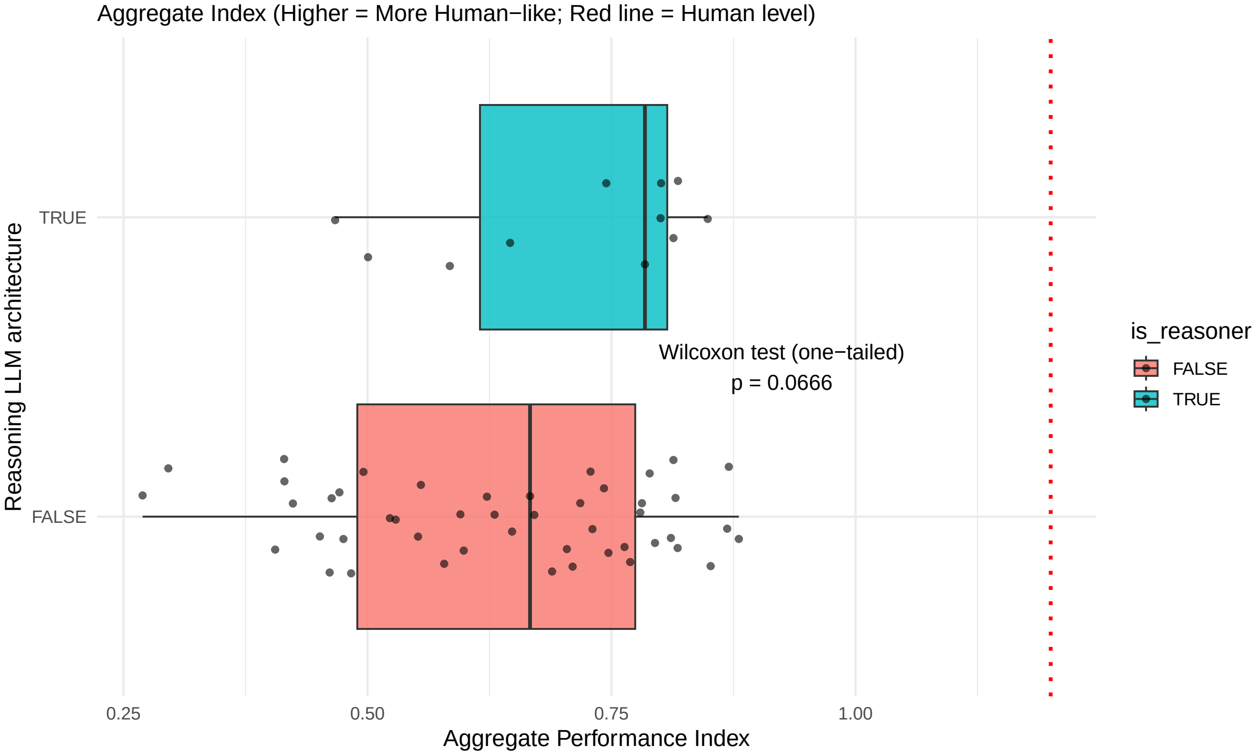


LLM Performance by Reasoner Classification

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



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.