Triage Against the Machine: Can AI Reason Deliberatively?

Francesco Veri, Gustavo Umbelino

2025-04-28

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

	Provider	Model	Parameters (B)	Context Length	Architecture	Version
28	meta	llama2:70b	70	4100	-	1
29	meta	llama3.1:405B-turbo	405	128000	-	3
30	meta	llama3.2	3	131072	-	4
31	meta	llama3.3:70b	70	128000	-	5
32	meta	llama3:70b	70	8190	-	2
33	meta	llama4-maverick	17	1000000	MoE	6
34	meta	llama4-scout	17	1000000000	MoE	6
35	microsoft	phi	NA	NA	-	1
36	microsoft	phi2	NA	NA	-	2
37	microsoft	phi3	NA	NA	-	3
38	microsoft	phi3.5	NA	NA	-	4
39	microsoft	phi4	14	16000	dense, decoder-only	5
40	mistralai	ministral-3b-latest	3	128000	-	1
41	mistralai	ministral-8b-latest	8	128000	-	1
42	mistralai	mistral-large-latest	123	128000	-	1
43	mistralai	mistral-small-latest	22	32800	-	1
44	mistralai	open-mistral-7b	7	NA	-	NA
45	mistralai	open-mistral-nemo	12	128000	-	1
46	mistralai	open-mixtral-8x22b	39	65400	SMoE	1
47	mistralai	open-mixtral-8x7b	7	NA	SMoE	NA
48	openai	gpt-3.5-turbo	-	16385	-	1
49	openai	gpt-4	-	8192	-	3
50	openai	gpt-4-turbo	-	128000	-	3
51	openai	gpt-4.5-preview	-	128000	-	4
52	openai	gpt-4o	-	128000	-	2
53	openai	gpt-4o-mini	-	128000	-	2
54	openai	o1	-	200000	-	1
55	openai	o1-mini	NA	NA	-	1
56	openai	o3-mini	-	200000	-	2
57	qwen	qwen-max	-	32768	-	1
58	qwen	qwen-plus	-	131072	-	1
59	qwen	qwen-turbo	-	1000000	-	1
60	qwen	qwen1.5-110b-chat	110	NA	-	1
61	qwen	qwen1.5-72b-chat	72	8000	-	1
62	qwen	qwen2-72b-instruct	72	131072	-	2
63	qwen	qwen2.5-72b-instruct	72	131072	-	3
64	qwen	qwq-plus	-	131072	-	1
65	xai	grok-2-1212	-	131072	-	2
66	xai	grok-3-beta	-	131072	-	3
67	xai	grok-3-mini-beta	-	131072	-	3
68	xai	grok-3-mini-beta-r=high	-	131072	-	3
69	xai	grok-3-mini-beta-r=low	-	131072	-	3
70	xai	grok-beta	314	131072	MoE	1

We started the analysis with 70 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 37460 valid LLM responses across 20 surveys.

Cost

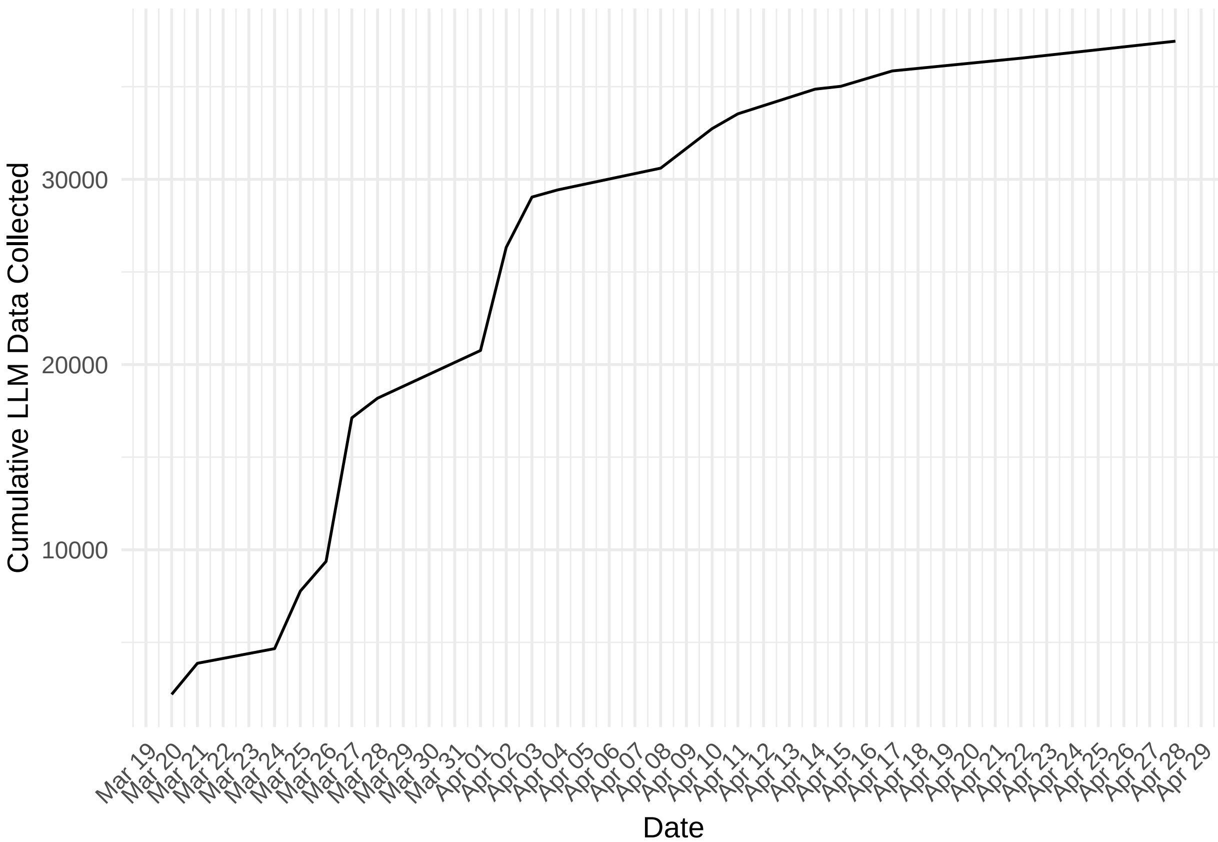
We spent a total of 411.3 USD. The cost breakdown per API is below.

Table 3: Costs by API

api	num_models	credits_paid
OpenAI API	9	225.52
Anthropic API	6	75.00
xAI API	6	29.95
Cohere API	6	20.34
Mistral AI API	8	20.00
Alibaba Cloud	8	17.49
Together AI	8	13.00
DeepSeek API	2	10.00
Google Could	8	NA
ollama	10	NA

Time

It took a total of 183 hours[1] across 39 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 Monday, Apr 28, 2025.



Excluded Models

17 out of 74 were excluded from the analysis for the following reasons.

Table 4: Excluded models and reasons

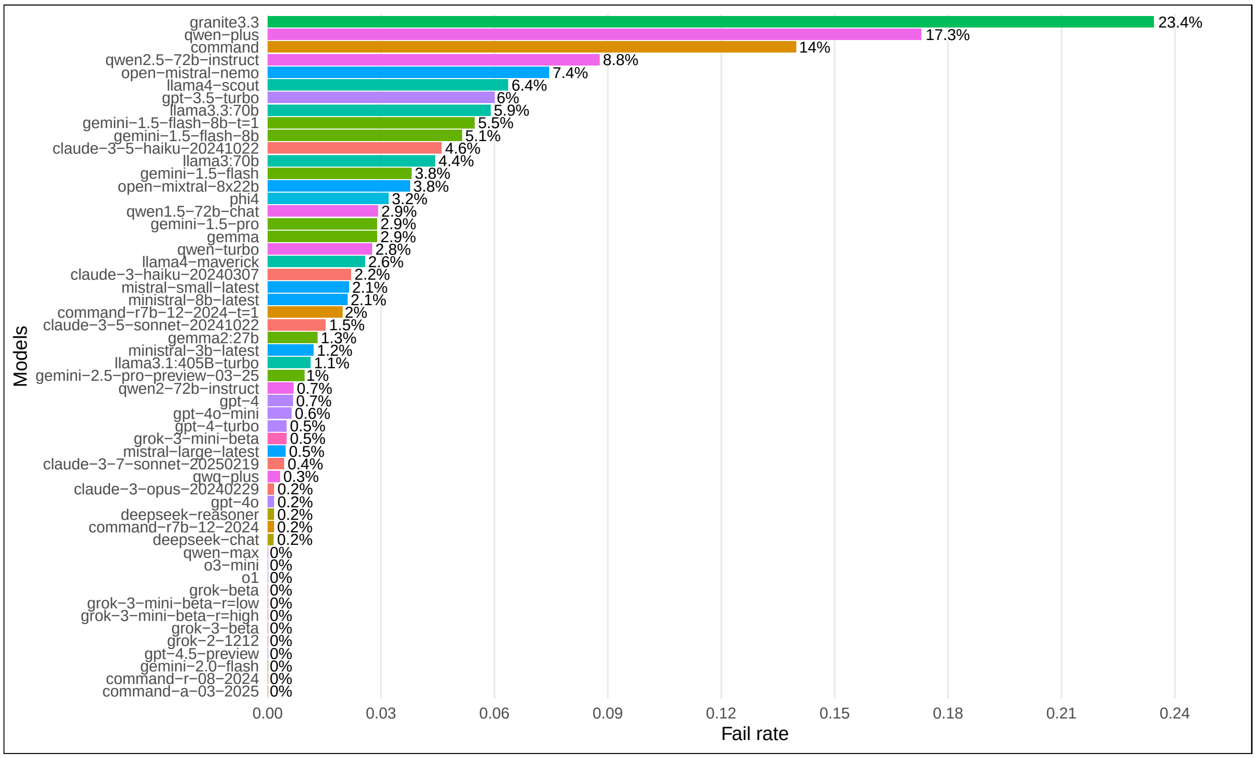
Provider	Model	Reason for exclusion
anthropic	claude-3-sonnet-20240229	not available in Anthropic API anymore
cohere	command-r-plus-08-2024	uniform aggregated considerations (1s)
deepseek	deepseek-v2	high fail rate (85%)
deepseek	deepseek-v2.5	too big to run locally; not available through APIs
google	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
microsoft	phi2	same model as phi
microsoft	phi3	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.

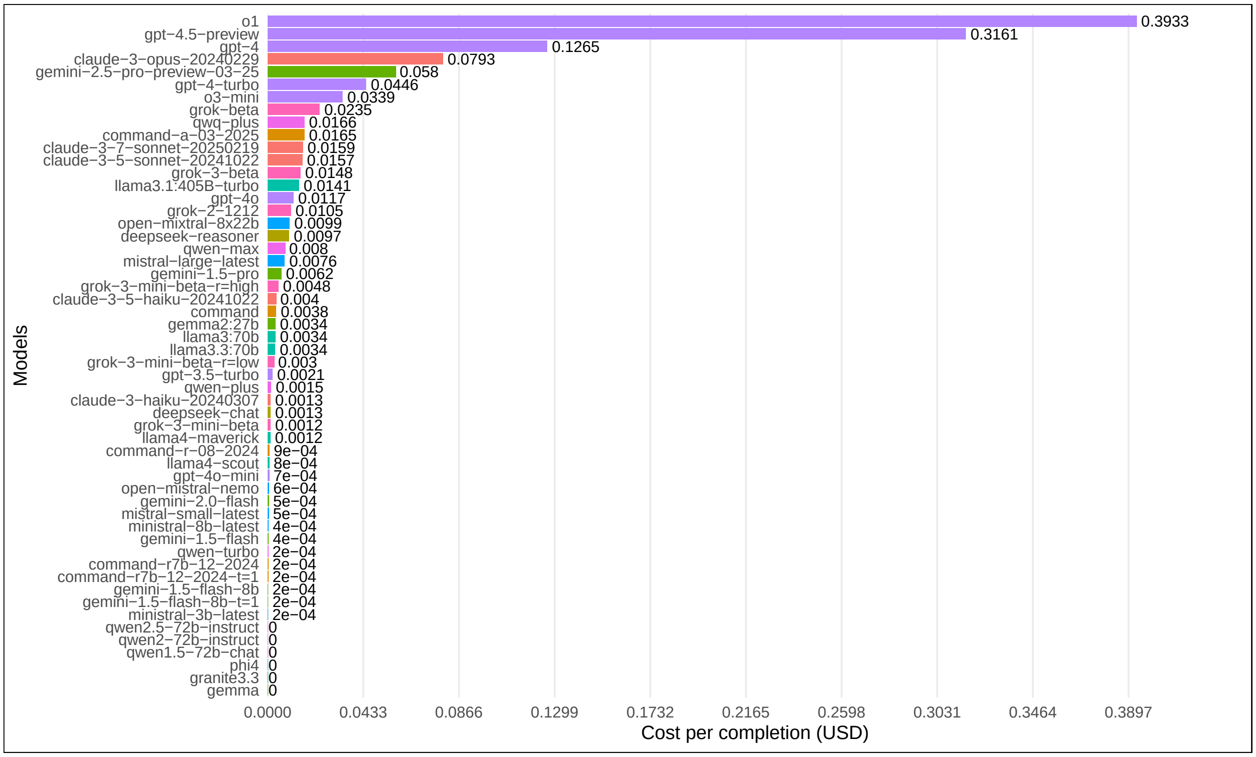
Provider	Model	Reason for exclusion
microsoft	phi3.5	10% success rate for biobanking_wa
mistralai	open-mistral-7b	11% success rate for auscj, uppsala_speaks, and biobanking_wa
mistralai	open-mixtral-8x7b	6% success rate on fremantle only
openai	o1-mini	0% success rate on uppsala_speaks only; responds 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

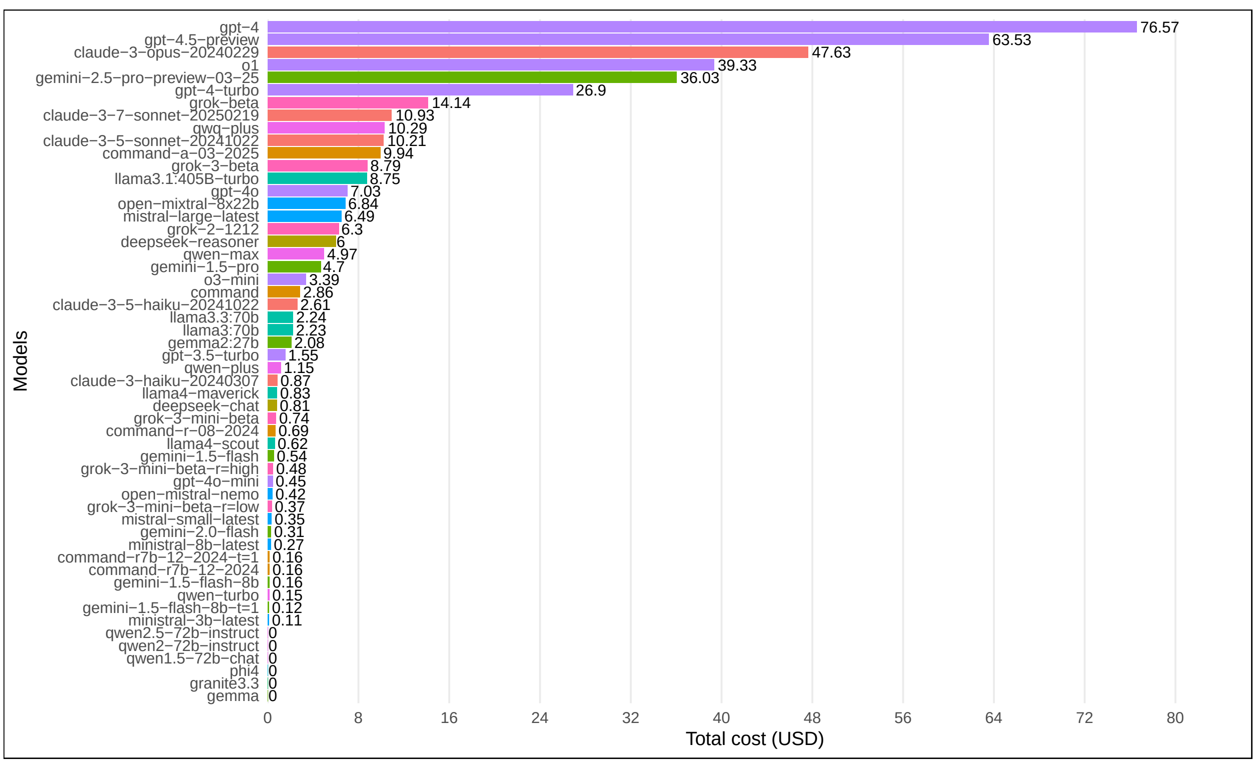
Fail rate



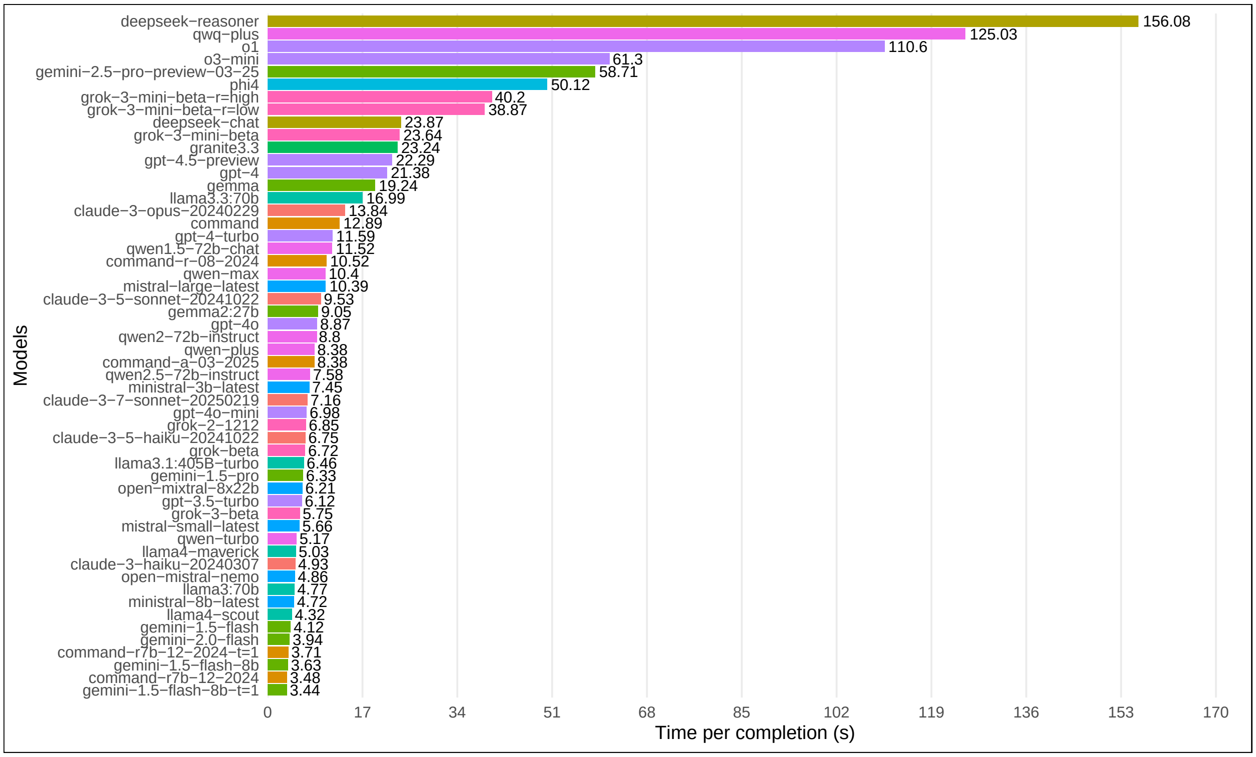
Cost per completion



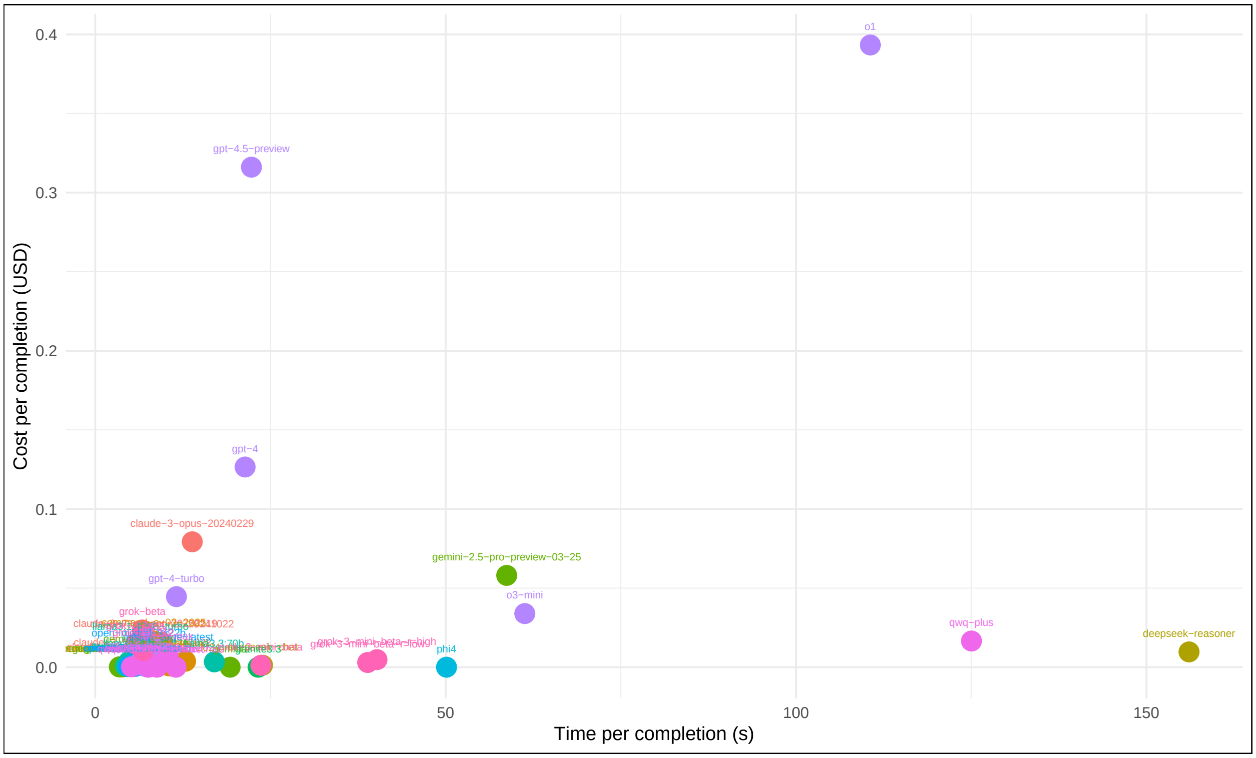
Total cost



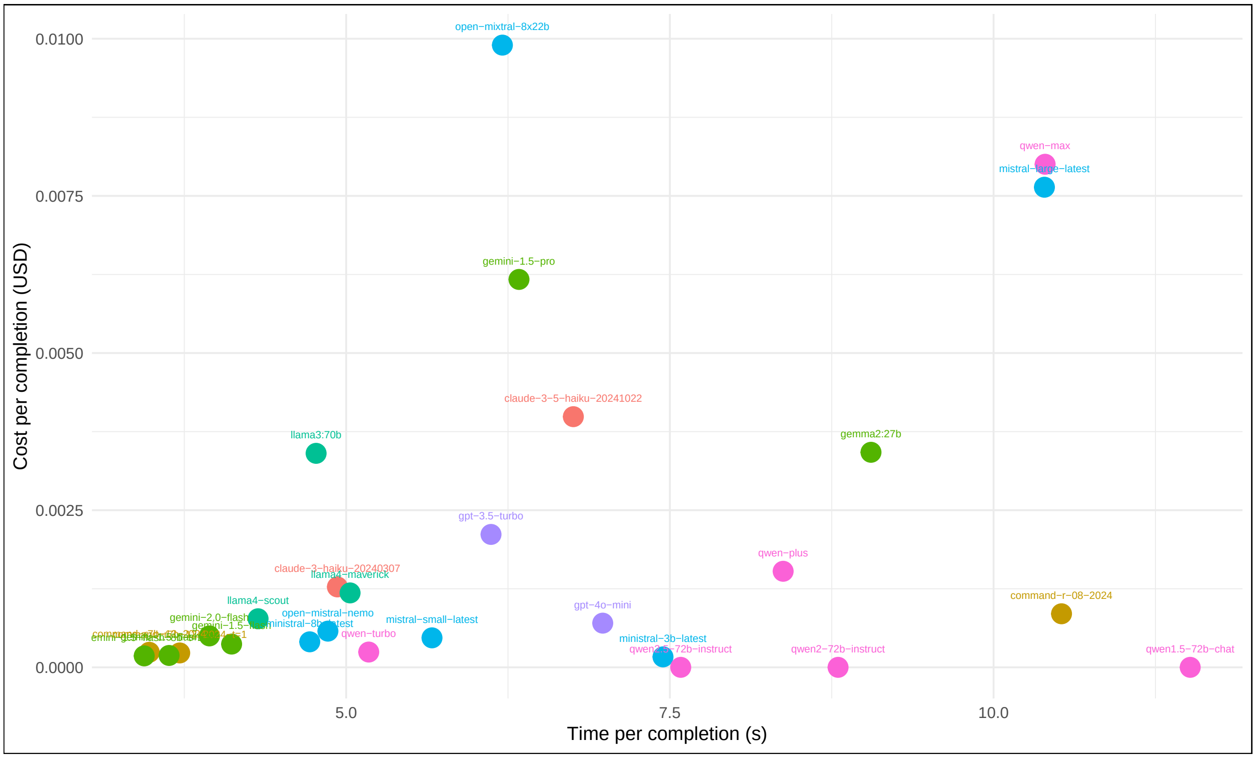
Time per completion



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	xai	grok-3-mini-beta-r=high	100	0.91	0.90	0.81
52	xai	grok-3-mini-beta-r=low	124	0.91	0.89	0.80
53	openai	o1	100	0.92	0.92	0.77
54	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 33169 LLM responses into 1080 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, 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) {
```

```r
  # 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
## 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 swiss_health from DRIInd.LLMs!
```

Hypotheses Testing

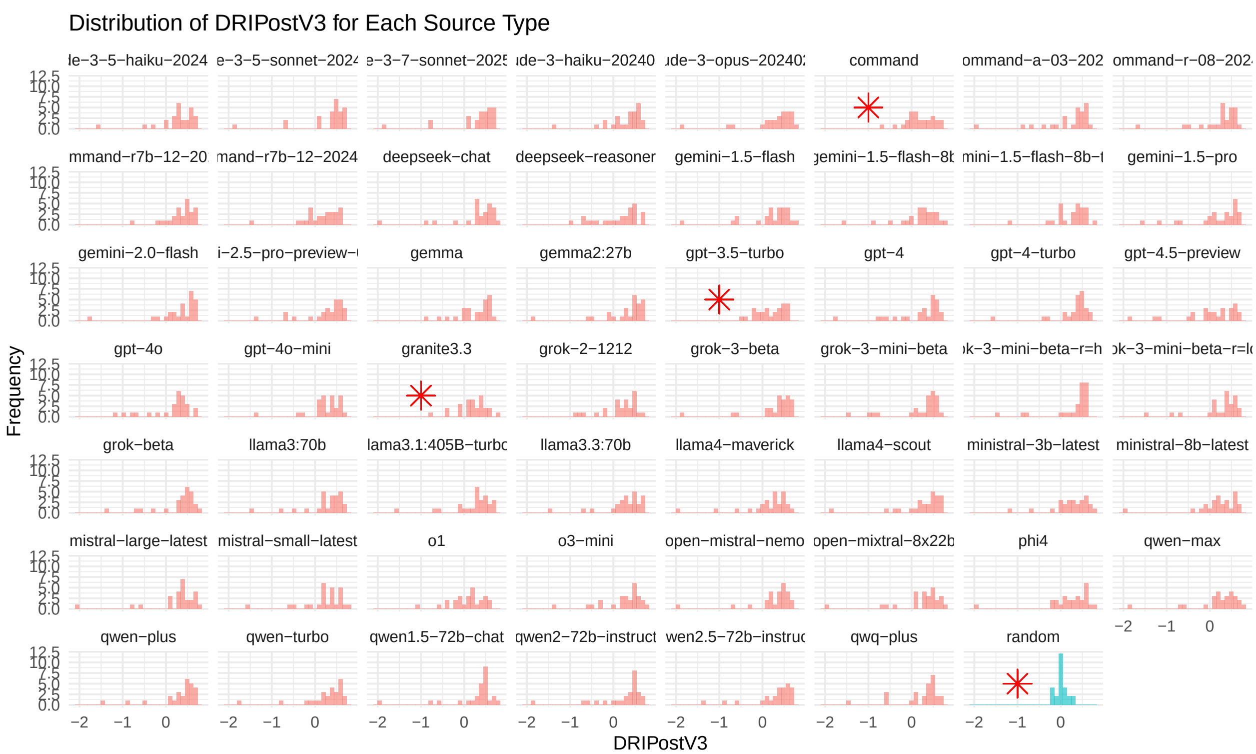
H1. DRI scores of LLMs do not significantly differ from those produced by a random generation process.

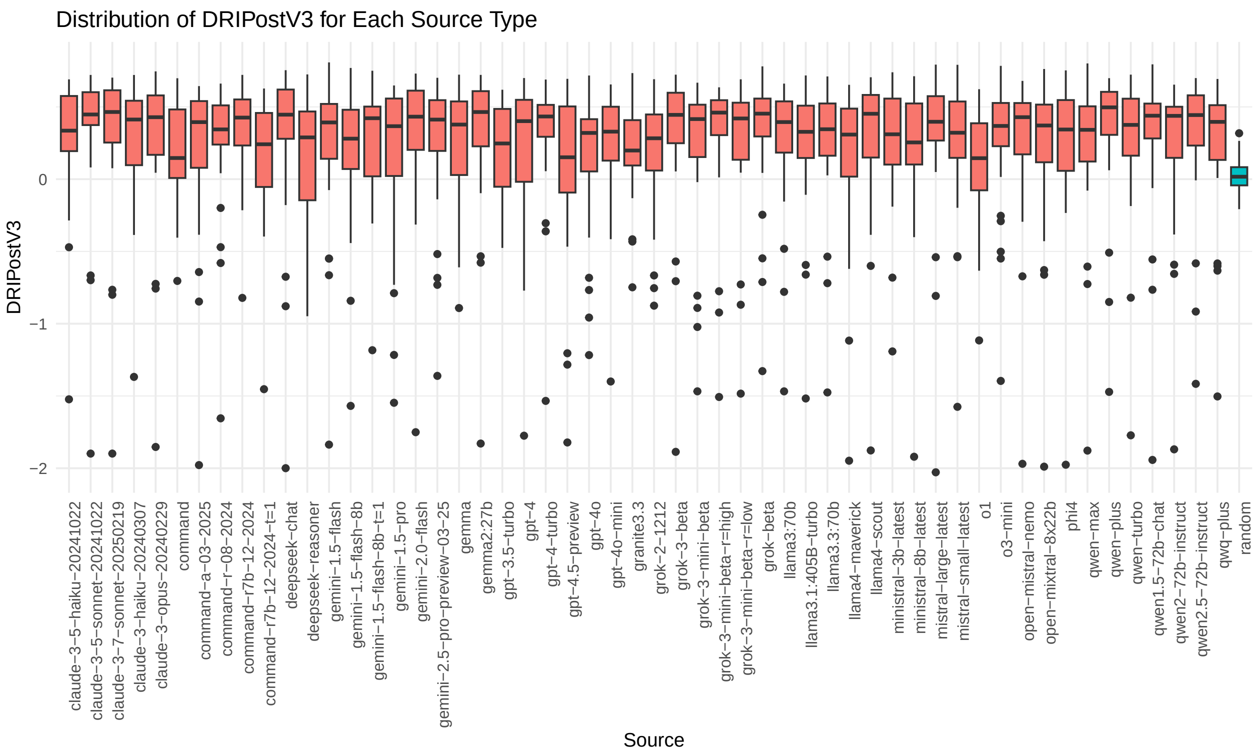
Testing assumptions

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 = 83.061, df = 54, p-value = 0.006719
```

Post-hoc tests

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

Table 7: Models compared to random

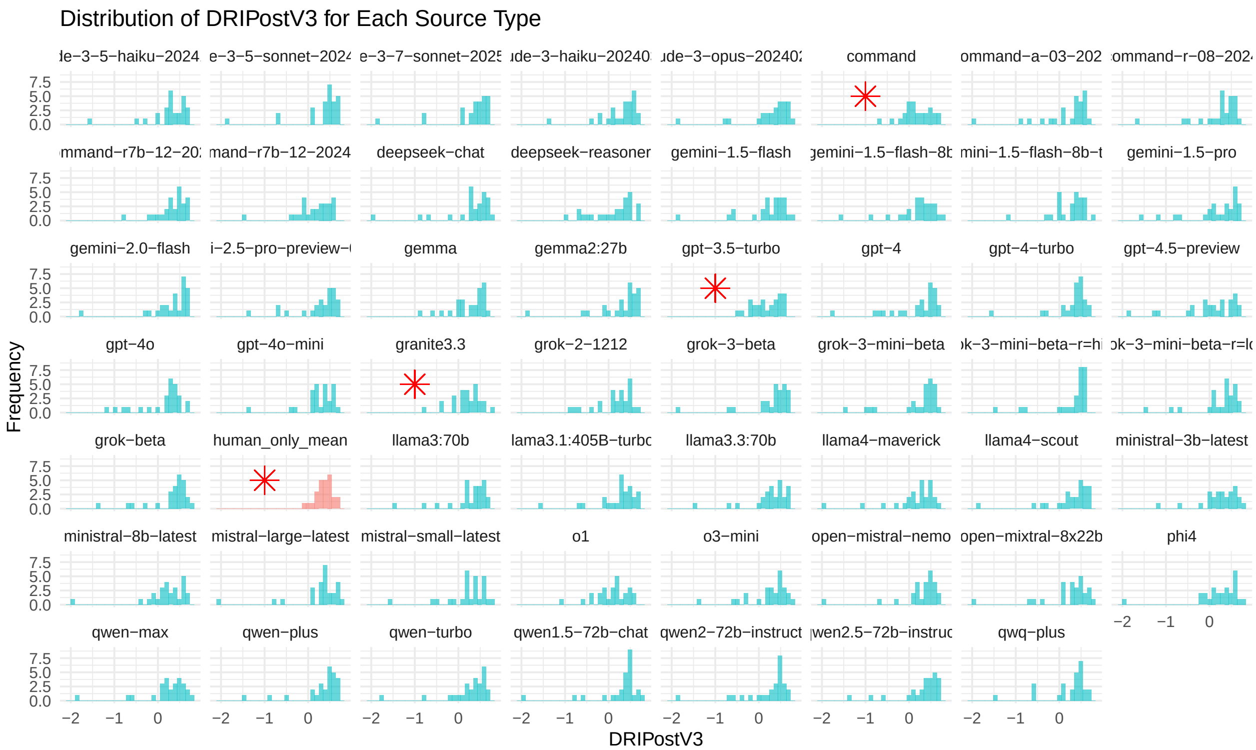
Model	P-adjusted
claude-3-5-sonnet-20241022	0.001*
qwen-plus	0.002*
gemini-2.0-flash	0.003*
claude-3-7-sonnet-20250219	0.003*
deepseek-chat	0.003*
grok-3-beta	0.005*
gemma2:27b	0.006*
qwen2.5-72b-instruct	0.007*
claude-3-opus-20240229	0.012*
grok-beta	0.013*
grok-3-mini-beta-r=high	0.013*
command-r7b-12-2024	0.02*
qwen1.5-72b-chat	0.027*

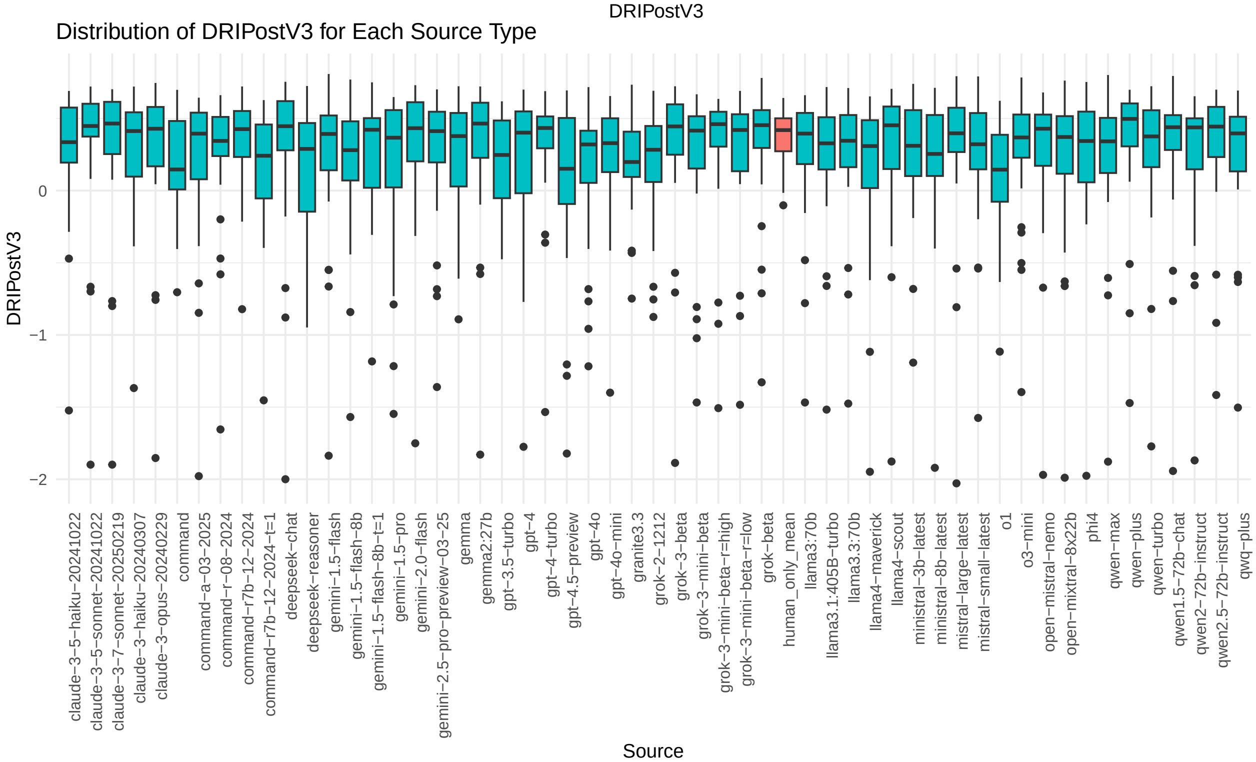
Model	P-adjusted
llama4-scout	0.028*
gpt-4-turbo	0.031*
mistral-large-latest	0.043*
open-mistral-nemo	0.053
gemini-2.5-pro-preview-03-25	0.057
claude-3-haiku-20240307	0.079
claude-3-5-haiku-20241022	0.11
llama3.3:70b	0.111
grok-3-mini-beta-r=low	0.122
qwen-turbo	0.123
qwen2-72b-instruct	0.139
grok-3-mini-beta	0.145
llama3:70b	0.161
o3-mini	0.175
open-mixtral-8x22b	0.178
qwq-plus	0.206
command-a-03-2025	0.302
command-r-08-2024	0.312
gemma	0.317
gemini-1.5-flash	0.338
qwen-max	0.343
ministral-3b-latest	0.39
phi4	0.434
gpt-4	0.449
gemini-1.5-flash-8b-t=1	0.458
llama3.1:405B-turbo	0.462
gemini-1.5-pro	0.501
gpt-4o-mini	0.849
mistral-small-latest	0.992
command	1
command-r7b-12-2024-t=1	1
deepseek-reasoner	1
gemini-1.5-flash-8b	1
gpt-3.5-turbo	1
gpt-4.5-preview	1
gpt-4o	1
granite3.3	1
grok-2-1212	1
llama4-maverick	1
ministral-8b-latest	1
o1	1

Some models, 16 out of 54, are significantly different than random.

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

Testing assumptions





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 = 59.173, df = 54, p-value = 0.2924
```

Post-hoc tests

```
## Kruskal-Wallis test is not significant; no need for post-hoc testing.
```

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.

Dependent variable: - DRIPostV3

Independent variable: - case - series

- Levels
- version

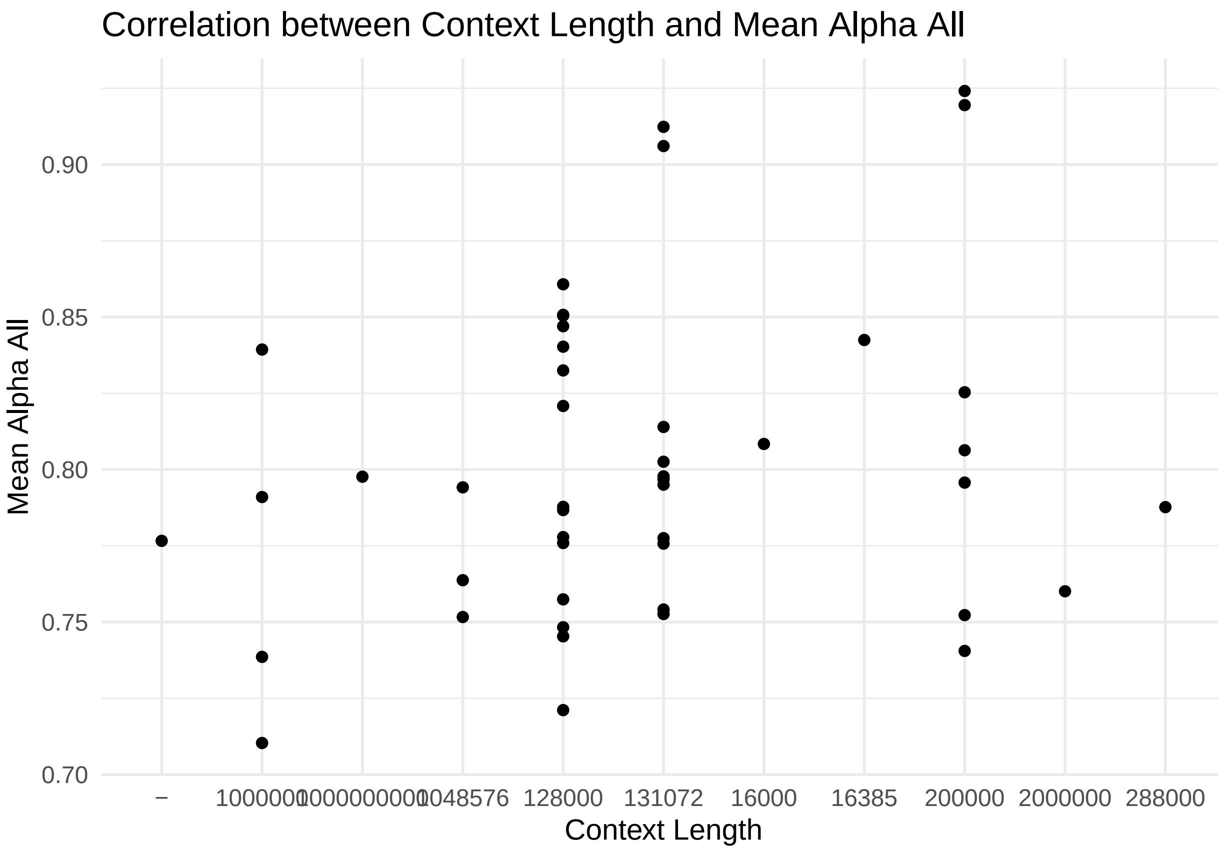
gemini

```
## Joining with `by = join_by(provider, model)`
```

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

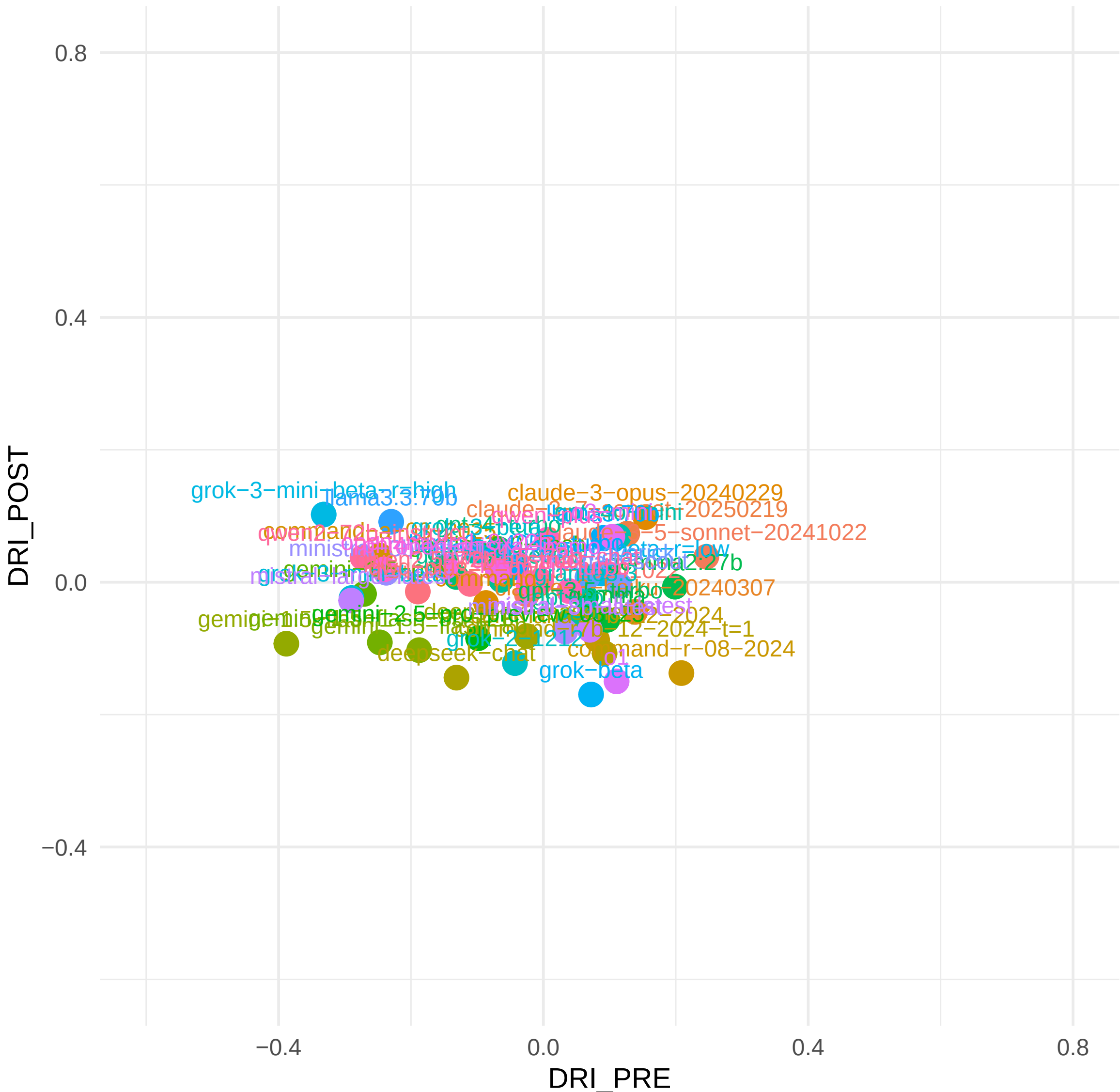
```
## `geom_smooth()` using formula = 'y ~ x'
```



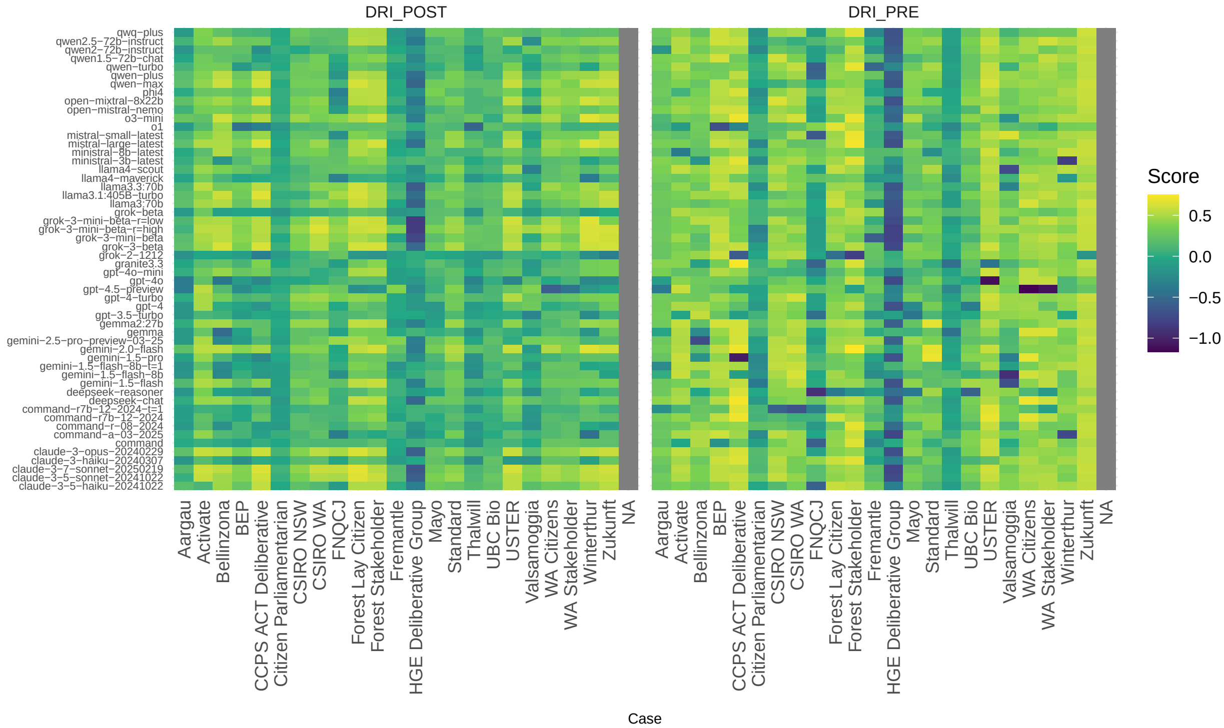

```
## `summarise()` has grouped output by 'provider', 'model'. You can override using
## the `.groups` argument.
```

Comparison PRE and POST DRI by Provider

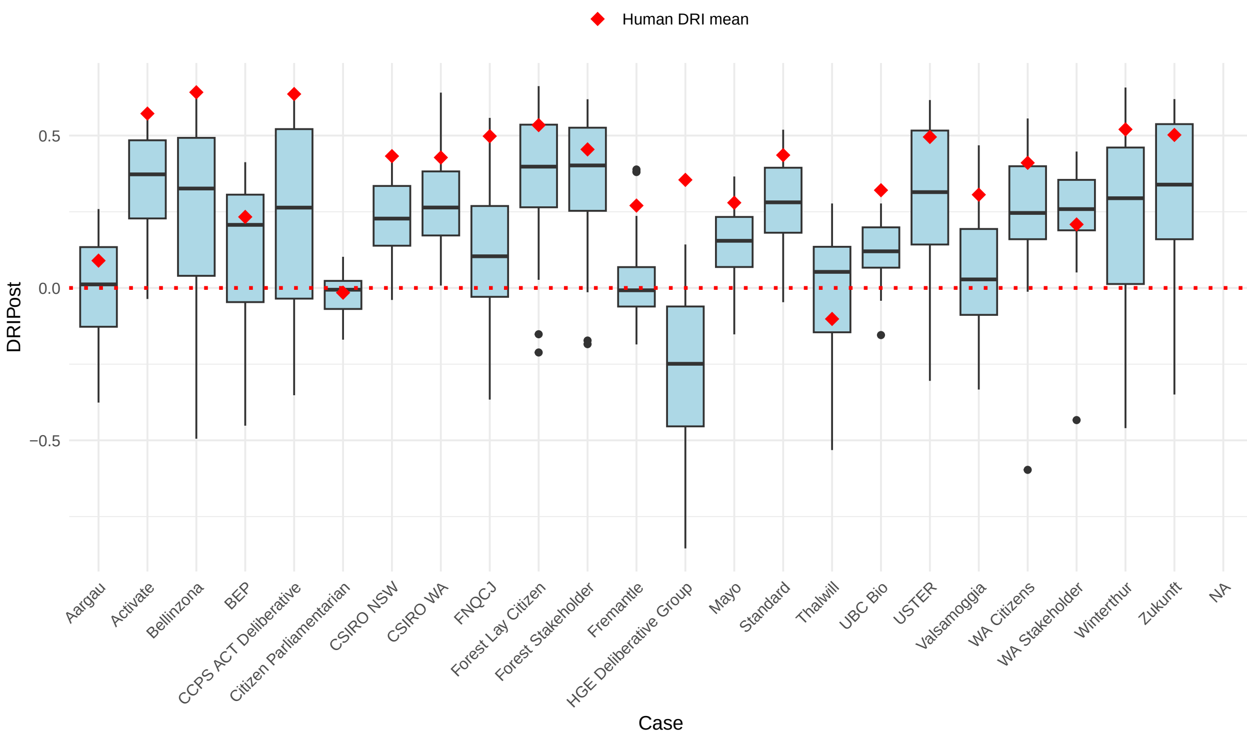
Comparison PRE and POST DRI by Model



Heatmap of DRI Scores by Case and Model



Boxplot of LLM DRI Post by Case



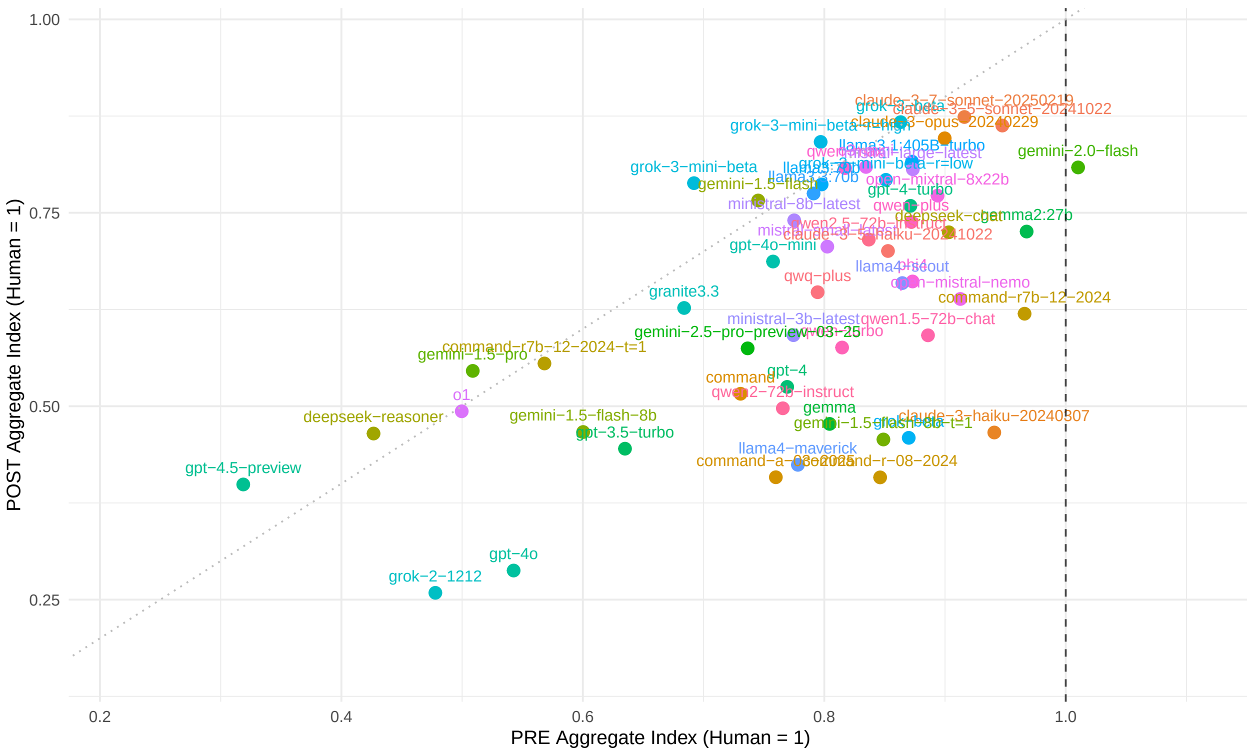
LLM Performance Metrics Against Human DRI Post-Scores

Table 8: LLM Performance Metrics Against Human DRI Post-Scores

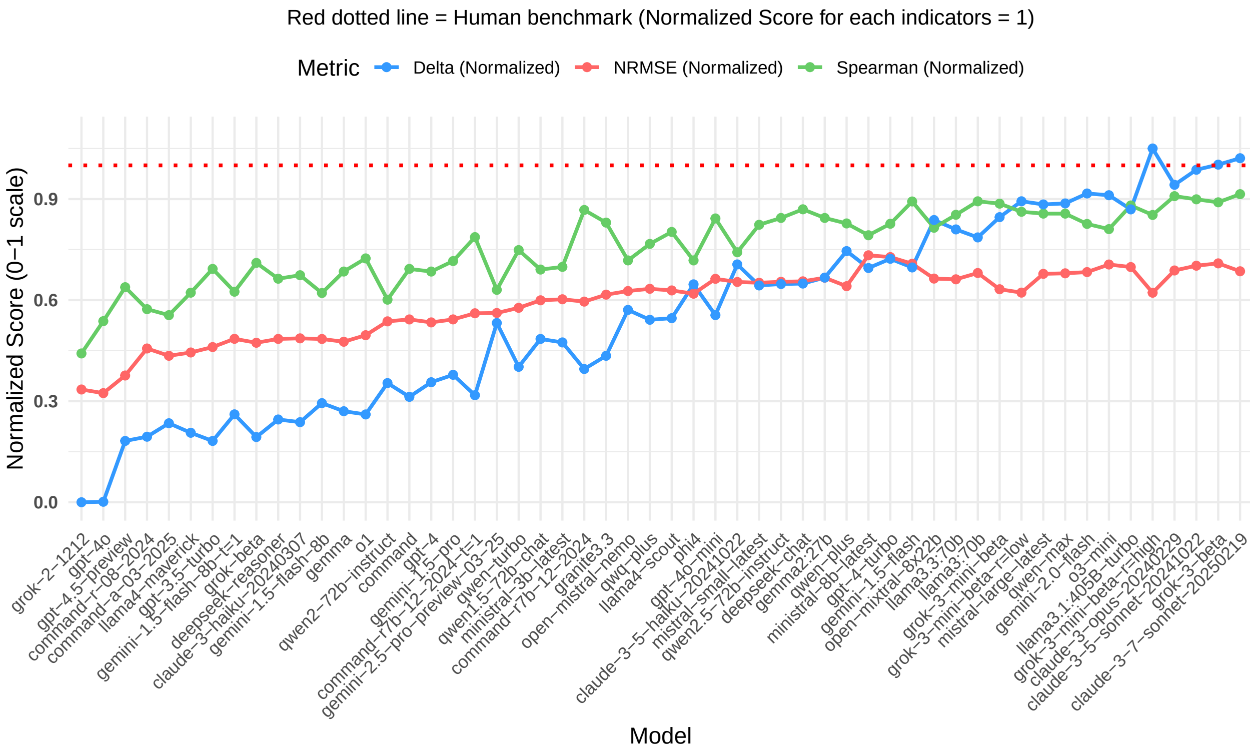
Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
ministral-8b-latest	0.159	0.199	64.169	0.744	0.214	0.267	0.585	-0.132
gpt-4-turbo	0.156	0.202	67.398	0.744	0.210	0.272	0.653	-0.120
grok-3-beta	0.137	0.216	68.972	0.744	0.184	0.291	0.781	0.001
gemini-1.5-flash	0.169	0.217	72.202	0.744	0.227	0.292	0.786	-0.131
o3-mini	0.158	0.219	82.559	0.744	0.213	0.294	0.622	-0.038
claude-3-5-sonnet-20241022	0.123	0.222	58.471	0.744	0.165	0.298	0.798	-0.006
llama3.1:405B-turbo	0.142	0.224	61.098	0.744	0.191	0.302	0.762	-0.056
claude-3-opus-20240229	0.137	0.232	82.755	0.744	0.184	0.312	0.817	-0.025
claude-3-7-sonnet-20250219	0.131	0.234	70.055	0.744	0.176	0.314	0.829	0.009
gemini-2.0-flash	0.162	0.236	67.030	0.744	0.218	0.317	0.652	-0.036
llama3:70b	0.153	0.238	75.800	0.744	0.206	0.320	0.787	-0.092
qwen-max	0.151	0.238	57.643	0.744	0.204	0.320	0.713	-0.049
mistral-large-latest	0.162	0.239	63.323	0.744	0.217	0.322	0.713	-0.050
gemma2:27b	0.175	0.248	56.098	0.744	0.236	0.334	0.688	-0.144
open-mixtral-8x22b	0.158	0.250	63.802	0.744	0.212	0.336	0.629	-0.070
gpt-4o-mini	0.213	0.250	80.191	0.744	0.286	0.337	0.685	-0.192
llama3.3:70b	0.153	0.252	79.979	0.744	0.205	0.338	0.707	-0.082
deepseek-chat	0.182	0.256	88.504	0.744	0.245	0.344	0.739	-0.152
qwen2.5-72b-instruct	0.189	0.257	59.553	0.744	0.254	0.346	0.688	-0.152
claude-3-5-haiku-20241022	0.176	0.257	55.228	0.744	0.237	0.346	0.484	-0.127
mistral-small-latest	0.192	0.259	72.000	0.744	0.258	0.348	0.647	-0.154
qwen-plus	0.180	0.267	82.789	0.744	0.241	0.359	0.655	-0.110
qwq-plus	0.215	0.273	64.194	0.744	0.289	0.366	0.534	-0.198
grok-3-mini-beta	0.155	0.273	60.580	0.744	0.209	0.368	0.773	-0.066
llama4-scout	0.219	0.276	62.340	0.744	0.294	0.371	0.605	-0.196
open-mistral-nemo	0.219	0.277	71.038	0.744	0.295	0.373	0.436	-0.185
grok-3-mini-beta-r=low	0.160	0.281	60.245	0.744	0.215	0.378	0.724	-0.046
grok-3-mini-beta-r=high	0.159	0.281	88.347	0.744	0.214	0.378	0.706	0.022
phi4	0.206	0.283	70.441	0.744	0.278	0.381	0.436	-0.153
granite3.3	0.244	0.285	65.819	0.744	0.329	0.384	0.660	-0.244
ministral-3b-latest	0.238	0.296	66.808	0.744	0.320	0.398	0.397	-0.227
qwen1.5-72b-chat	0.239	0.298	66.994	0.744	0.321	0.401	0.381	-0.223
command-r7b-12-2024	0.273	0.301	97.802	0.744	0.367	0.404	0.735	-0.261
qwen-turbo	0.261	0.315	66.540	0.744	0.351	0.423	0.497	-0.258
gemini-2.5-pro-preview-03-25	0.226	0.326	81.485	0.744	0.304	0.438	0.261	-0.202
command-r7b-12-2024-t=1	0.295	0.327	104.633	0.744	0.397	0.439	0.574	-0.295
gemini-1.5-pro	0.271	0.340	73.504	0.744	0.364	0.457	0.432	-0.269
command	0.301	0.340	81.253	0.744	0.405	0.457	0.385	-0.297
qwen2-72b-instruct	0.287	0.344	89.085	0.744	0.385	0.463	0.203	-0.279
gpt-4	0.280	0.347	84.661	0.744	0.377	0.466	0.370	-0.278
o1	0.320	0.375	136.095	0.744	0.430	0.504	0.448	-0.320
claude-3-haiku-20240307	0.330	0.382	100.335	0.744	0.444	0.514	0.348	-0.329
gemini-1.5-flash-8b-t=1	0.322	0.383	110.212	0.744	0.433	0.515	0.250	-0.319
deepseek-reasoner	0.326	0.383	107.279	0.744	0.438	0.515	0.327	-0.326
gemini-1.5-flash-8b	0.307	0.383	107.853	0.744	0.412	0.515	0.242	-0.305
gemma	0.315	0.389	96.367	0.744	0.424	0.524	0.370	-0.315
grok-beta	0.349	0.392	135.151	0.744	0.469	0.527	0.421	-0.349
gpt-3.5-turbo	0.356	0.401	104.842	0.744	0.479	0.539	0.385	-0.354
command-r-08-2024	0.348	0.404	123.314	0.744	0.468	0.544	0.146	-0.348

Model	MAE	RMSE	MAPE (%)	Human Range	NMAE	NRMSE	Spearman	Delta
llama4-maverick	0.348	0.413	95.251	0.744	0.468	0.555	0.244	-0.343
command-a-03-2025	0.336	0.420	95.487	0.744	0.451	0.565	0.111	-0.331
gpt-4.5-preview	0.372	0.464	119.164	0.744	0.500	0.624	0.277	-0.354
grok-2-1212	0.432	0.495	139.461	0.744	0.581	0.665	-0.117	-0.432
gpt-4o	0.432	0.503	135.621	0.744	0.580	0.676	0.075	-0.432

PRE vs. POST Aggregate Scores Correlation Across LLMs



Human-Normalized Performance

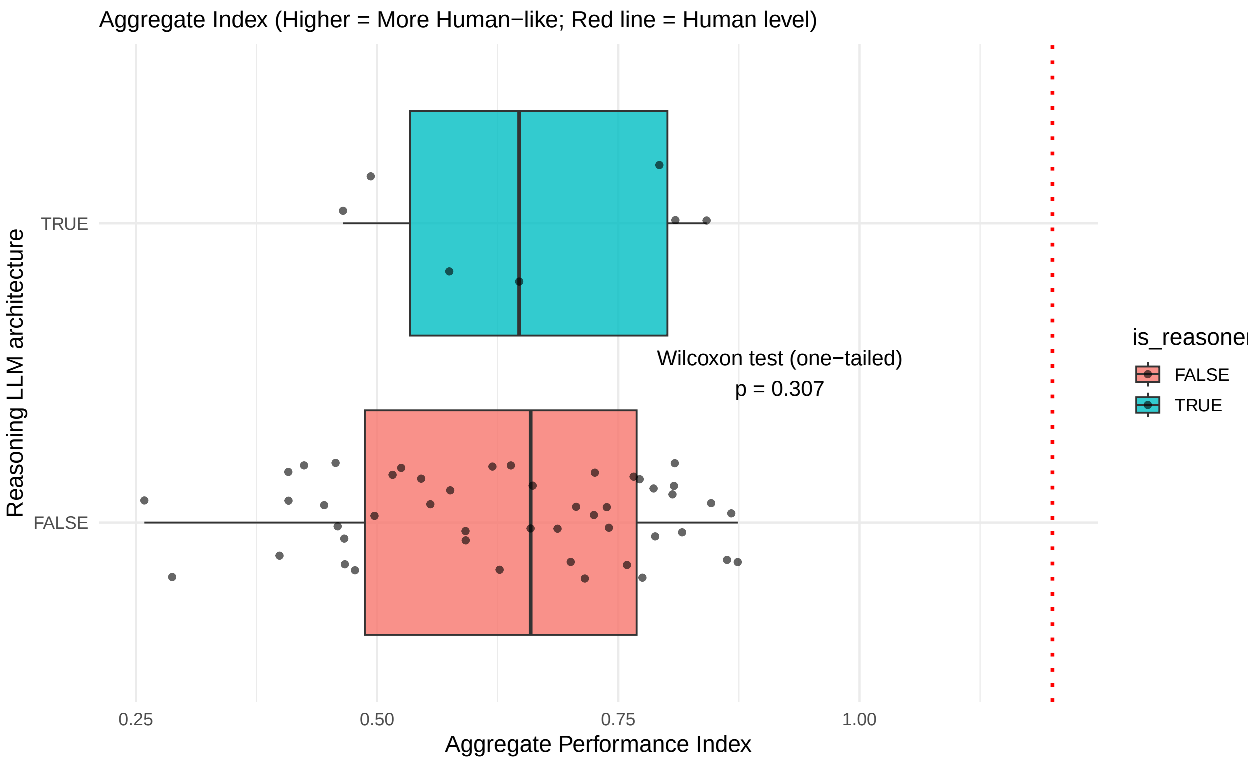


LLM Performance by Reasoner Classification

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.