Triage Against the Machine: Can AI Reason Deliberatively?

Francesco Veri, Gustavo Umbelino

2025-04-07

Large-Language Models (LLMs) Preview

Table 1: LLMs

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

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

We started the analysis with 56 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

We collected a total of 29431 valid LLM responses across 20 surveys.

Cost

We spent a total of 238.71 USD. The cost breakdown per API is below.

Table 3: Costs by API

api	num_models	credits_paid
OpenAI API	8	90.52
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
Google Could	4	NA
ollama	9	NA

Time

It took a total of 147 hours[1] across 15 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 04, 2025.

Excluded Models

14 out of 58 were excluded from the analysis for the following reasons.

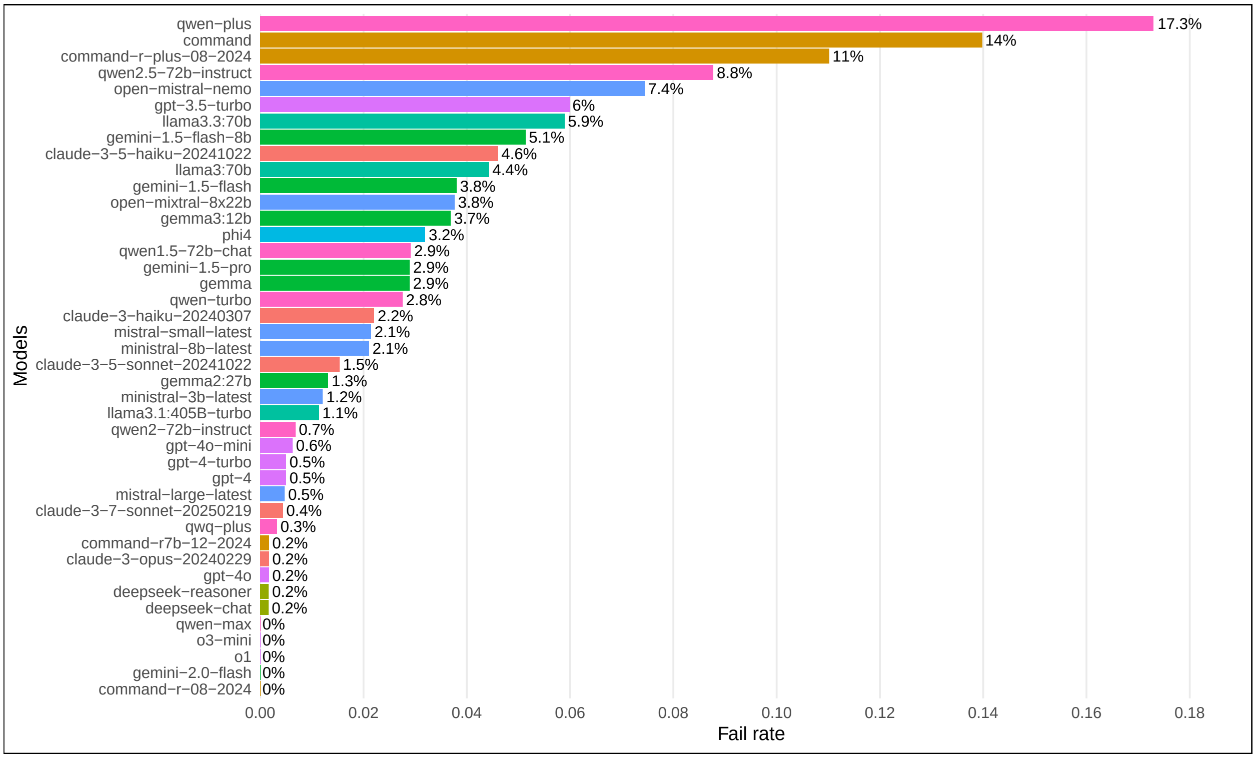
Table 4: Excluded models and reasons

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

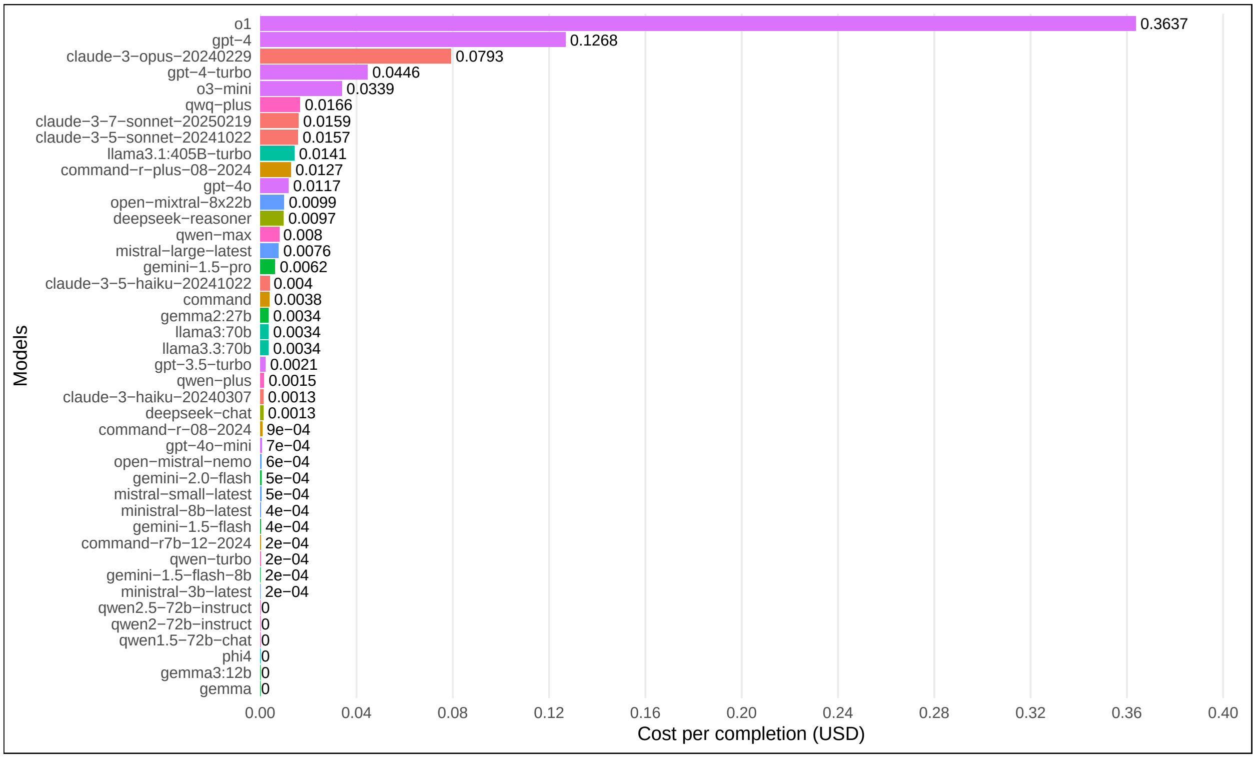
[1] Execution data is mostly accurate. Only a few (3-5) executions failed and, as a result, we have no record of it.

Execution Summary Plots

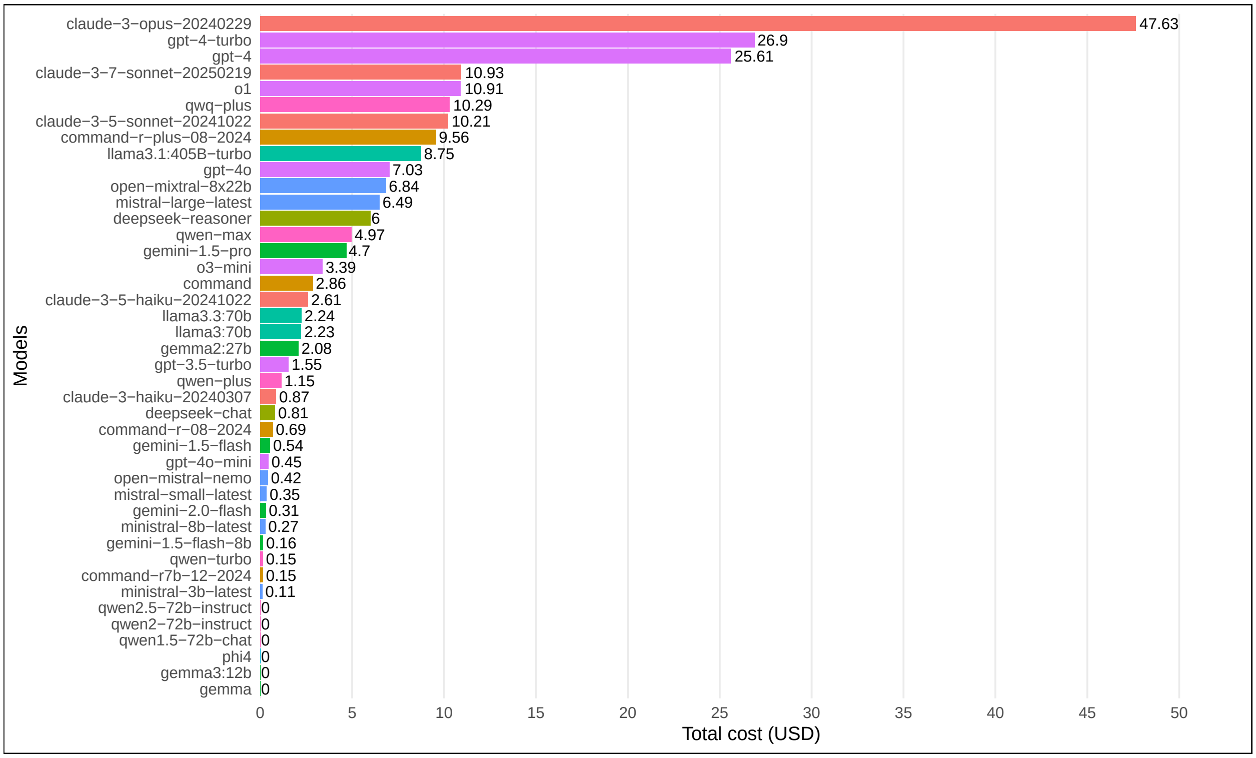
Fail rate



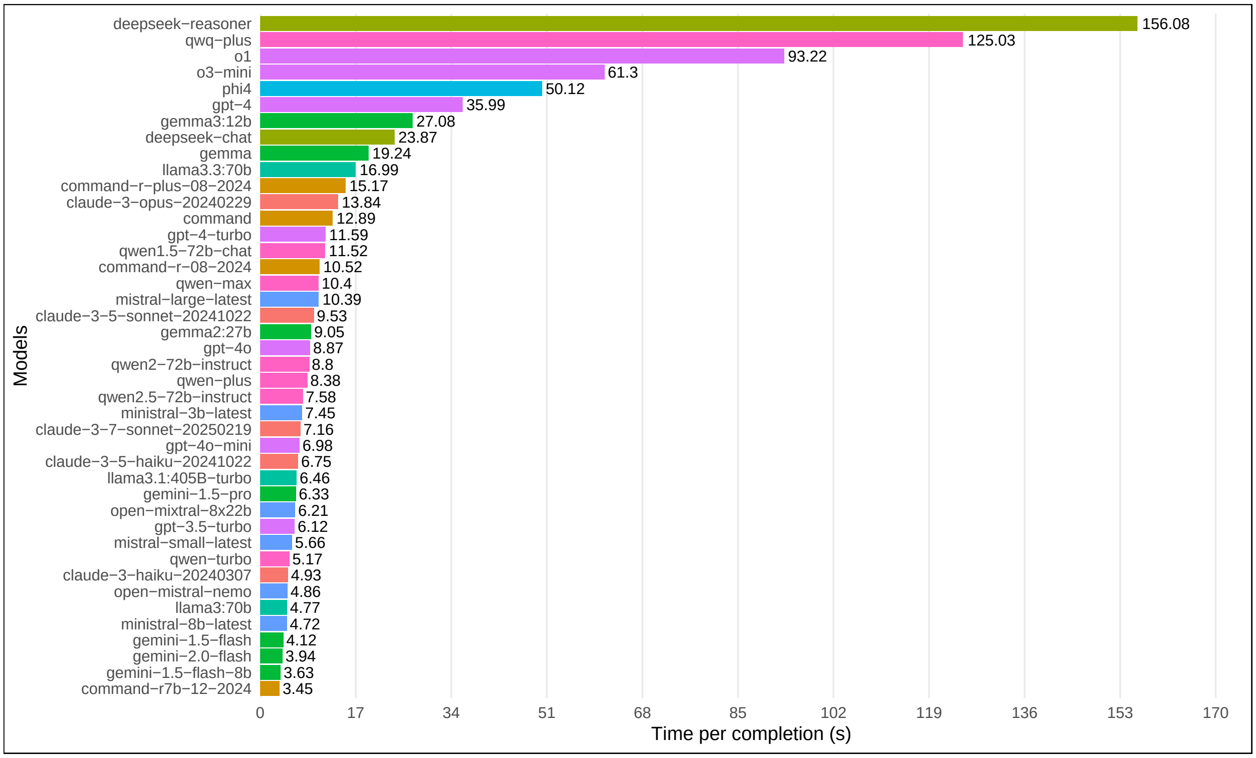
Cost per completion



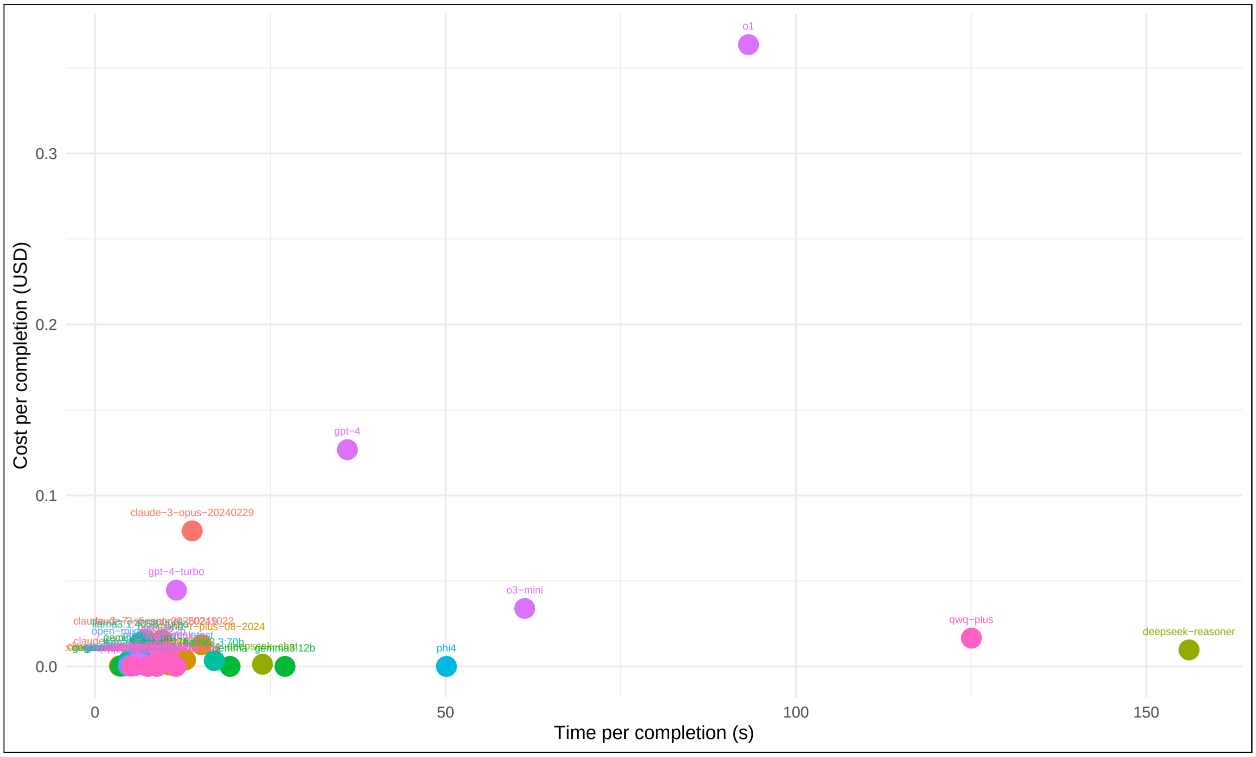
Total cost



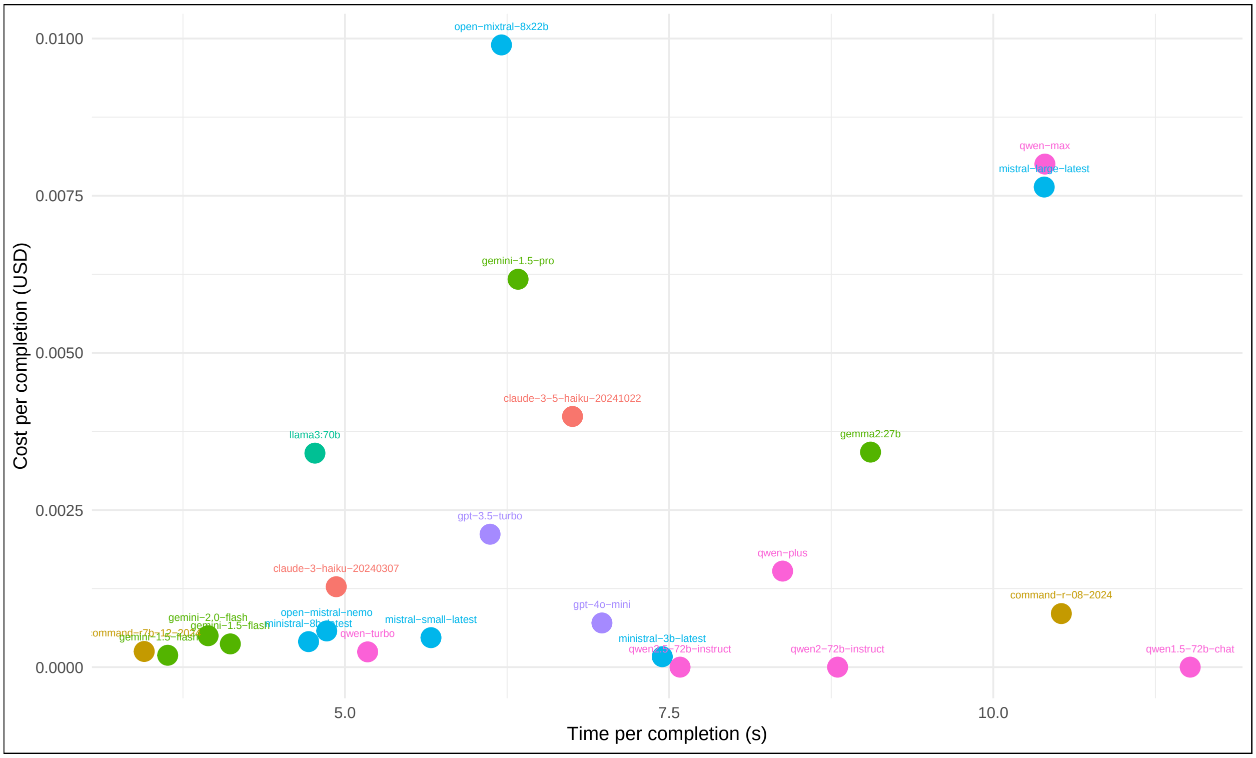
Time per completion



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

Check alpha results per model

```
## Warning: There were 4 warnings in `summarise()`.
## The first warning was:
## i In argument: `min_alpha_considerations = min(alpha_considerations, na.rm =
##   TRUE)`.
## i In group 45: `provider = "qwen"` `model = "qwen1.5-110b-chat"`.
## Caused by warning in `min()`:
## ! no non-missing arguments to min; returning Inf
## i Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.

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

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

```
## [1] "Aggregation of 29431 LLM responses across 20 surveys completed in 2.06 secs"
```

It takes 2.06 secs to run the aggregation script.

Read and format human data

Generate random participants

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]])) / 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 17 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

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

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.