Arbitrary Choices Are Not Random

A forced-choice audit of 42 LLMs across 50,400 planned trials.

Daniel Alonso; AI assistance disclosed

May 8, 2026

Interactive version with charts and downloads: https://crow.sg/research/llm-arbitrary-choice-study.
Technical report. Public artifacts and checksums are indexed at
https://crow.sg/research/llm-arbitrary-choice-study/artifact-manifest.json. Source code:
https://github.com/Crow-Tech-Pte-Ltd/research/tree/main/llm-arbitrary-choice-study.

Abstract

This technical report documents a forced-choice audit of 42 large language models across 50,400 planned trials and 48,316 successful one-word responses. The prompts ask for a choice between two ordinary words where neither option is intended to be correct. Normal and swapped option order were both collected, with and without weak contextual sentences. The result is not evidence of belief or intent. It is evidence that these systems often carry measurable positional, lexical, and context-sensitive regularities into trivial decisions.

Keywords: LLM evaluation; forced-choice prompting; prompt order bias; lexical preference; context sensitivity.

1 Study Design

The run used ordinary word pairs, weakly associated context sentences, order-swapped controls, and repeated calls at temperature 0.7. Model calls were made through OpenRouter and persisted into SQLite with raw responses, parse status, usage data, and later attempt-level audit rows. Daniel Alonso conducted the study. AI assistants, including GPT-5.5, assisted with code, orchestration, analysis, and redaction.

Table 1: Study inventory and source provenance.

Field	Value
Models	42
Providers	22
Families	21
Tiers	flagship, mid, open, small, unknown
Word pairs	30 across 17 categories
Contexts	60 snippets; 55 with an inferred target option
Conditions	bare, bare_swapped, context, context_swapped
Repetitions	10 per condition
Temperature	0.7
Planned trial rows	50,400
Source database	source SQLite run log; public release is study-summary.json and generated artifacts
Source updated	2026-05-08 10:06:11 UTC

The bare prompt template was `Choose one: {first} or {second}. Reply with exactly one word: "{first}" or "{second}".` Context prompts prepended one short sentence to the same choice. The source runner generated trials from a fixed seed and then consumed pending rows in insertion order, so the prepared trial set is reproducible for the same config and data files.

Table 2: Prompt conditions, completion, and first-position share.

Condition	Order	Context	Shape	OK / planned	OK rate	First share
bare	original order	no	bare template	12,081/12,600	95.9%	75.8%
bare swapped	swapped order	no	bare template	12,072/12,600	95.8%	59.3%
context	original order	yes	context sentence + bare template	12,074/12,600	95.8%	51.1%
context swapped	swapped order	yes	context sentence + bare template	12,089/12,600	95.9%	55.5%

2 Model Coverage

The model pool covered 22 providers, 21 model families, and four descriptive tiers. Provider, family, origin, and tier labels come from local study config and route names; they are a sampling taxonomy, not a benchmark ranking.

Table 3: Coverage by tier.

Tier	Description	Models	OK / planned	OK rate	Mean first	Mean semantic	Usage
flagship	flagship / frontier-class	10	11,948/12,000	99.6%	60.4%	56.7%	USD 9.94
mid	current mid-tier	10	11,999/12,000	100.0%	57.8%	49.9%	USD 10.02
open	open-weight or open-route	5	5,998/6,000	100.0%	62.2%	41.1%	USD 0.08
small	small or fast tier	16	17,172/19,200	89.4%	61.0%	49.4%	USD 2.17
unknown	uncategorized	1	1,199/1,200	99.9%	63.3%	49.0%	USD 0.00

Table 4: Coverage by provider. Status mix includes preserved caveat rows.

Provider	Models	OK / planned	OK rate	Mean first	Mean semantic	Usage	Status mix
Alibaba/Qwen	3	3,580/3,600	99.4%	57.9%	67.0%	USD 4.19	error 18; ok 3,580; rate limited 2
Amazon	1	1,200/1,200	100.0%	57.1%	47.7%	USD 0.04	ok 1,200
Anthropic	3	3,600/3,600	100.0%	61.8%	54.1%	USD 0.74	ok 3,600
Baidu	2	1,200/2,400	50.0%	52.1%	64.3%	USD 0.01	error 34; model removed 444; ok 1,200; rate limited 722
DeepSeek	3	3,599/3,600	100.0%	61.6%	48.0%	USD 0.51	error 1; ok 3,599
Google	3	3,599/3,600	100.0%	53.6%	76.9%	USD 3.05	error 1; ok 3,599
IBM Granite	1	1,200/1,200	100.0%	65.8%	48.3%	USD 0.00	ok 1,200
Inception Labs	1	1,200/1,200	100.0%	67.5%	50.3%	USD 0.09	ok 1,200
Liquid AI	1	1,198/1,200	99.8%	82.4%	5.5%	USD 0.00	error 2; ok 1,198
Meta	3	3,598/3,600	99.9%	65.2%	36.6%	USD 0.04	error 2; ok 3,598
Microsoft	1	1,198/1,200	99.8%	48.6%	60.2%	USD 0.00	error 2; ok 1,198
MiniMax	3	3,598/3,600	99.9%	66.9%	38.6%	USD 1.95	error 2; ok 3,598
Mistral AI	2	2,400/2,400	100.0%	64.5%	38.8%	USD 0.11	ok 2,400
NVIDIA	2	2,400/2,400	100.0%	65.2%	47.0%	USD 0.08	ok 2,400
Nous Research	2	2,400/2,400	100.0%	50.5%	61.0%	USD 0.09	ok 2,400
OpenAI	2	2,400/2,400	100.0%	48.0%	45.5%	USD 0.31	ok 2,400
Perplexity	1	1,199/1,200	99.9%	49.1%	50.3%	USD 6.09	error 1; ok 1,199
Reka AI	1	382/1,200	31.8%	51.0%	39.4%	USD 0.80	invalid 126; model removed 692; ok 382
Tencent	1	1,196/1,200	99.7%	74.6%	23.0%	USD 0.01	error 4; ok 1,196
Z.ai	3	3,599/3,600	100.0%	56.8%	65.2%	USD 2.56	error 1; ok 3,599
tencent	1	1,199/1,200	99.9%	63.3%	49.0%	USD 0.00	error 1; ok 1,199
xAI	2	2,371/2,400	98.8%	65.8%	39.1%	USD 1.56	error 29; ok 2,371

Table 5: Coverage by model family.

Family	Models	Providers	Tiers	Mean first	Mean semantic
Claude	3	Anthropic	flagship, mid, small	61.8%	54.1%
DeepSeek	3	DeepSeek	flagship, mid, small	61.6%	48.0%
GLM	3	Z.ai	mid, small	56.8%	65.2%
Gemini	3	Google	flagship, small	53.6%	76.9%
Llama	3	Meta	open	65.3%	36.6%
MiniMax	3	MiniMax	flagship	66.9%	38.6%
Qwen	3	Alibaba/Qwen	flagship, mid, small	57.9%	67.1%
ERNIE	2	Baidu	small	52.1%	64.3%
GPT / OpenAI	2	OpenAI	mid, small	48.0%	45.5%
Grok	2	xAI	flagship, mid	65.8%	39.1%
Hermes	2	Nous Research	flagship, open	50.5%	61.0%
Hunyuan	2	Tencent, tencent	small, unknown	68.9%	36.0%
Mistral	2	Mistral AI	mid, small	64.5%	38.8%
Nemotron	2	NVIDIA	open, small	65.2%	47.0%
Granite	1	IBM Granite	small	65.8%	48.3%
LFM	1	Liquid AI	small	82.4%	5.5%
Mercury	1	Inception Labs	mid	67.5%	50.3%
Nova	1	Amazon	flagship	57.1%	47.7%
Phi	1	Microsoft	small	48.6%	60.2%
Reka	1	Reka AI	small	51.0%	39.4%
Sonar	1	Perplexity	mid	49.1%	50.3%

Table 6: Coverage by organization-origin label.

Origin	Models	Providers	OK / planned	OK rate
China	15	Alibaba/Qwen, Baidu, DeepSeek, MiniMax, Tencent, Z.ai	16,772/18,000	93.2%
EU	2	Mistral AI	2,400/2,400	100.0%
Other	2	Liquid AI, Reka AI	1,580/2,400	65.8%
US	22	Amazon, Anthropic, Google, IBM Granite, Inception Labs, Meta, Microsoft, NVIDIA, Nous Research, OpenAI, Perplexity, xAI	26,365/26,400	99.9%
Unknown	1	tencent	1,199/1,200	99.9%

3 Word-Pair and Context Design

The word set intentionally avoids factual questions. The pairs are ordinary, short labels across sensory, spatial, object, and abstract categories. Each pair has two context snippets where possible, and 55 of 60 snippets had an unambiguous inferred target option from the association note.

Table 7: Word-pair categories.

Category	Pairs	Pair labels
abstract	2	narrow/wide; simple/complex
color	3	blue/red; green/purple; yellow/gray
density	1	light/heavy
direction	2	left/right; up/down
material	2	glass/stone; wood/metal
motion	1	fast/slow
nature	1	river/forest
object	3	candle/lamp; cup/plate; key/coin
shape	2	circle/square; triangle/oval
size	1	small/large
sound	2	loud/quiet; sharp/mellow
taste	2	salty/sour; sweet/bitter
temperature	1	warm/cold
terrain	1	mountain/valley
texture	2	smooth/rough; soft/hard
time	2	early/late; morning/evening
weather	2	humid/dry; windy/still

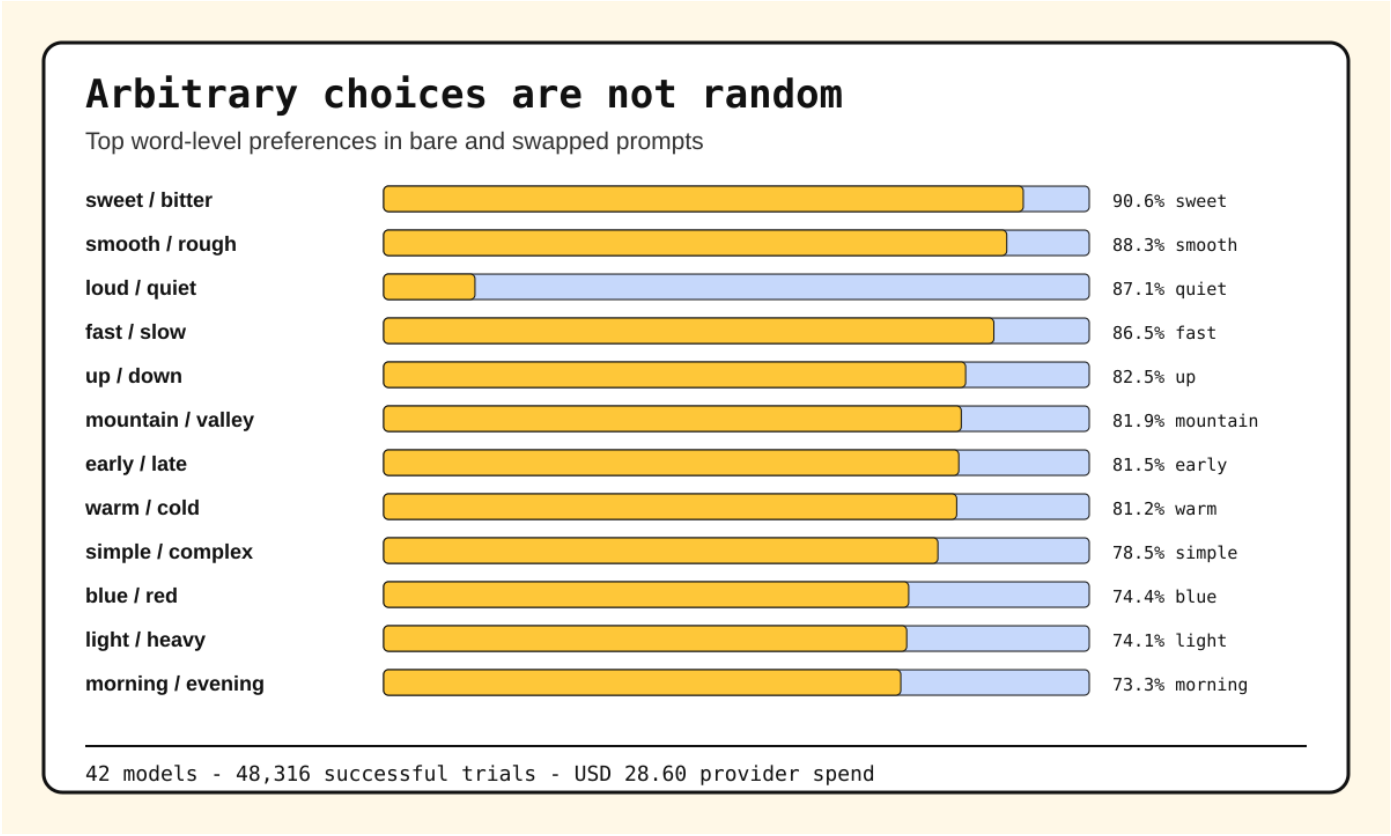

Figure 1: Top aggregate word preferences after combining bare and swapped prompts. The figure is generated from the same summary artifact used by the public HTML charts.

4 Main Results

The most direct result is that the aggregate first-option share was 60.4%, not 50 percent. That headline pools all four conditions, so it mixes pure position bias with the context conditions pulling toward their intended option. The cleaner position-only number is the order-balanced first-option rate within the two bare conditions: averaging 75.8% in bare and 59.3% in bare swapped gives a pure position bias of 67.5% once option identity is balanced. The remaining

gap between the two bare orders is the option-A word effect averaged across the 30 pairs. Across successful rows, models chose the first displayed option often enough to make position a visible part of the measurement. The stronger result is that positional bias and word preference are separable: some models mostly follow position, while others keep stable preferences for specific words even when order is swapped.

Table 8: Strongest aggregate word preferences in bare and swapped prompts.

Pair	Majority option	Majority share	OK rows
sweet/bitter	sweet	90.6%	802
smooth/rough	smooth	88.3%	804
loud/quiet	quiet	87.1%	804
fast/slow	fast	86.5%	806
up/down	up	82.5%	805
mountain/valley	mountain	81.9%	805
early/late	early	81.5%	806
warm/cold	warm	81.2%	809
simple/complex	simple	78.5%	811
blue/red	blue	74.4%	809

Table 9: Models with the strongest positional skew. First share is over successful rows.

Model	Provider	Tier	First share	Semantic strength	OK rows
LFM2-24B-A2B	Liquid AI	small	82.4%	5.5%	1,198
Hunyuan A13B Instruct	Tencent	small	74.6%	23.0%	1,196
Llama 3.3 70B Instruct	Meta	open	71.8%	15.3%	1,200
MiniMax M2.7	MiniMax	flagship	69.4%	35.7%	1,200
Mercury 2	Inception Labs	mid	67.5%	50.3%	1,200
Mistral Medium 3.5	Mistral AI	mid	67.2%	35.7%	1,200
Nemotron 3 Nano 30B A3B	NVIDIA	small	67.2%	41.3%	1,200
Claude Sonnet 4.6	Anthropic	mid	67.1%	33.0%	1,200

5 Context Effects

Weakly related language before the choice often moved the answer toward the word suggested by the surrounding sentence. This is not surprising, but the size is useful: it gives a practical warning that even throwaway phrasing can matter in forced-choice model evaluations.

Table 10: Largest mean context lifts toward the inferred associated option.

Context	Target	Baseline	Context	Mean lift	OK rows
black coffee	bitter	10.1%	99.8%	89.2 pp	488
old rope	rough	11.7%	100.0%	87.8 pp	402
concert line	loud	13.1%	100.0%	86.9 pp	320
old turtle	slow	13.6%	99.4%	85.6 pp	162
stairs basement	down	17.6%	95.5%	77.8 pp	484
suitcase	heavy	25.5%	99.6%	74.1 pp	238
sunset walk	red	25.5%	99.8%	73.3 pp	482
corridor	narrow	26.7%	99.4%	72.7 pp	488

Table 11: Context cues that moved against the intended association.

Context	Target	Baseline	Context	Mean lift	OK rows
bakery case	sweet	90.3%	38.3%	-51.7 pp	321
alarm clock	morning	73.7%	48.3%	-25.4 pp	480
race start	fast	86.5%	70.4%	-16.4 pp	645
morning window	yellow	54.3%	45.7%	-9.0 pp	726

6 Data Quality and Accounting

Most accepted rows were clean one-word responses: 48,091 of 48,316 OK rows (99.5

Table 12: Final trial status and parser status.

Trial status	Rows	Parser status	Rows
error	98	exact	48,091
invalid	126	none	1,266
model removed	1,136	invalid	818
ok	48,316	single token in text	118
rate limited	724	repeated single option	106
		manual override	1

Table 13: Cost and token accounting from captured usage rows.

Metric	Value	Note
OpenRouter dashboard spend	USD 28.60	External dashboard total used as the public spend figure.
Trial usage rows	USD 20.78	Usage JSON attached to final trial rows.
Attempt usage rows	USD 22.21	Captured provider attempts, including retries after attempt logging was added.
Prompt tokens	2,270,506	Captured attempt-level prompt tokens.
Completion tokens	10,908,206	Captured visible plus provider-reported completion tokens.
Reasoning tokens	10,471,809	Provider-reported hidden reasoning tokens when present in usage details.

Table 14: Recorded attempts grouped by max_tokens retry cap.

Token cap	Attempts	OK attempts	Non-OK mix
0	1	1	none
512	49,956	46,964	invalid 2,173; error 95; rate limited 724
3,000	1,262	465	invalid 797
20,000	1,043	886	invalid 154; error 3

7 Operational Caveats

The run was operationally messy in exactly the ways real model evaluation can be messy. ERNIE 4.5 21B A3B failed through repeated provider 429 and rate-limit responses. Reka Flash 3 produced a mix of blank, space-only, and very long invalid retries at high token caps, so it was removed with preserved caveat rows. One ERNIE 4.5 300B A47B row was manually overridden after repeated explanatory answers made the final one-word answer unambiguous. The OpenRouter dashboard spend was about $28.60; the attempts table records $22.21 of usage rows, with some early superseded attempts not fully captured.

Table 15: Where the non-OK rows concentrated.

Model	Provider	Non-OK rows	Share of all non-OK	Status mix
ERNIE 4.5 21B A3B	Baidu	1,200	57.6%	error 34; model removed 444; rate limited 722
Reka Flash 3	Reka AI	818	39.2%	invalid 126; model removed 692; ok 382
Grok 4.3	xAI	29	1.4%	error 29; ok 1,171
Qwen3.6 Max Preview	Alibaba/Qwen	20	1.0%	error 18; ok 1,180; rate limited 2

8 Limitations

- The task is intentionally narrow: ordinary binary word choices, not factual QA, planning, tool use, safety behavior, or human preference modeling.
- The measured quantities are behavioral regularities in one-word responses. They are not evidence of belief, intent, consciousness, moral preference, or stable model personality.
- OpenRouter provider routing was not pinned: calls used the bare model field with no provider.order or provider.only constraints, so a single nominal model id was free to be served by any of OpenRouter's available backends within the run. Closed-source models effectively hit one backend each, but open-weight routes hit many in the same run (DeepSeek v3.2: 9 providers; Llama 3.3 70B: 13; DeepSeek v4 Pro: 6). The first-option rate within one open-weight model varies up to roughly 8 percentage points across its serving providers, so per-model rankings on open-weight routes carry routing noise on top of the model itself. The served provider is recorded per attempt for slicing.
- Default decoding behavior and hidden reasoning are part of the observed run. Different reasoning controls or temperatures could change the measurements.
- Per-cell statistical power is small: 10 reps per (model, pair, condition) gives a binomial 95% CI half-width of about ±15 pp at p=0.5. Aggregate findings (overall position bias across 48 k rows, per-model first-option share across 1,200 rows, per-pair preference across ~800 rows) are statistically robust; single (model, pair, condition) cell claims should be treated as exploratory.
- Per-pair preferences are not just unigram corpus frequency. Regressing the order-balanced option-A share on the wordfreq Zipf log-frequency gap across the 30 pairs gives Pearson r = 0.235, R-squared = 0.055, slope p = 0.21; the more frequent token wins only 19 of 30 pairs. Clear counter-frequency cases include sharp/mellow (mellow wins 70% despite a 1.16-Zipf gap toward sharp), smooth/rough, warm/cold, and circle/square. This rules out the simplest unigram-frequency confound but does not rule out richer frequency-based stories such as collocation strength or one-word answer typicality.
- Context labels are researcher hypotheses inferred from weak wording. Reverse context rows are reported because the cues are not ground-truth semantic interventions.
- 96.8% of non-OK rows came from two caveat routes, so reliability conclusions should focus on the preserved status counts rather than average all-model failure behavior.
- Attempt-level cost accounting is incomplete for early superseded retries because full attempt logging was added during the run; the dashboard spend is therefore retained as the headline spend figure.

9 Data Availability

The public machine-readable release in this web package is study-summary.json, with a JSON Schema and artifact manifest that expose record counts, field names, metric definitions, byte sizes, and SHA-256 hashes. The study code is published in the linked GitHub repository. The public summary includes model rows, provider rollups, pair bias rows, per-model pair bias rows, context effects, status counts, parse counts, attempt-token groups, cost fields, caveats, a data dictionary, and metric definitions.

10 Interpretation

The point is simple: arbitrary model choices are not necessarily random samples from an even distribution. If a product or benchmark forces an LLM into a binary choice, prompt order, word identity, and nearby context can become part of the outcome. That should be measured, not hand-waved away.

11 References and Artifacts

- Public article: https://crow.sg/research/llm-arbitrary-choice-study
- Source code repository: https://github.com/Crow-Tech-Pte-Ltd/research/tree/main/llm-arbitrary-choice-study
- Artifact manifest: https://crow.sg/research/llm-arbitrary-choice-study/artifact-manifest.json
- Uploadable public summary JSON: https://crow.sg/research/llm-arbitrary-choice-study/study-summary.json
- Summary JSON schema: https://crow.sg/research/llm-arbitrary-choice-study/study-summary.schema.json
- Artifact manifest schema: https://crow.sg/research/llm-arbitrary-choice-study/artifact-manifest.schema.json
- Paper PDF: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choice-s
pdf
- Printable HTML: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choic
html
- LaTeX source: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choice
tex
- BibTeX citation: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choic
bib
- Bias map SVG: https://crow.sg/research/llm-arbitrary-choice-study/bias-map.svg
- Bias map PNG: https://crow.sg/research/llm-arbitrary-choice-study/bias-map.png
- API and spend platform used for the study calls: OpenRouter.
- Execution disclosure: Daniel Alonso conducted the study; AI assistants, including GPT-5.5, assisted with implementation, orchestration, analysis, and redaction.