Arbitrary Choices Are Not Random

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

Daniel Alonso, with GPT-5.5 and Hermes assistance

May 8, 2026

MODELS	PROVIDERS	OK ROWS	FIRST OPTION
42	21	48,316	60.4%

Abstract

This technical report summarizes a forced-choice audit of 42 LLMs across 50,400 planned trials. The task asks for a choice between two ordinary words where neither option is intended to be correct. Normal and swapped option order make position effects visible; context variants expose sensitivity to weak surrounding language. The study measures behavioral regularities, not belief, intent, consciousness, or moral preference.

Study Design

The run used 30 word pairs, 60 weak context snippets, four prompt conditions, ten repetitions, and temperature 0.7. Calls went through OpenRouter and landed in SQLite with raw responses, parser decisions, usage rows, and attempt history.

Study inventory and provenance

Field	Value
Models	42
Providers	21
Families	21
Word pairs	30 across 17 categories
Contexts	60 snippets; 55 with inferred targets
Conditions	bare, bare_swapped, context, context_swapped
Source database	results/raw/full_candidates.sqlite3
Source updated	2026-05-08 10:06:11 UTC

Prompt-condition design

Condition	Order	Context	Shape	OK/planned
bare	original order	no	bare template	12,081/12,600
bare swapped	swapped order	no	bare template	12,072/12,600
context	original order	yes	context sentence + bare template	12,074/12,600
context swapped	swapped order	yes	context sentence + bare template	12,089/12,600

Prompt-condition first-option share

Model Coverage

The pool covers 21 providers, 21 families, and four descriptive tiers. Provider, family, origin, and tier labels come from local route metadata and are a sampling taxonomy, not a benchmark ranking.

Coverage by tier

Tier	Class	Models	OK/planned	Mean first	Mean semantic
flagship	flagship / frontier-class	10	11,948/12,000	60.4%	56.7%
mid	current mid-tier	10	11,999/12,000	57.8%	49.9%
open	open-weight or open-route	5	5,998/6,000	62.2%	41.1%
small	small or fast tier	17	18,371/20,400	61.1%	49.3%

Coverage by provider

Provider	Models	OK/planned	OK rate	Mean first	Status mix
Alibaba/Qwen	3	3,580/3,600	99.4%	57.9%	error 18; ok 3,580; rate limited 2
Amazon	1	1,200/1,200	100.0%	57.1%	ok 1,200
Anthropic	3	3,600/3,600	100.0%	61.8%	ok 3,600
Baidu	2	1,200/2,400	50.0%	52.1%	error 34; model removed 444; ok 1,200; rate limited 722
DeepSeek	3	3,599/3,600	100.0%	61.6%	error 1; ok 3,599
Google	3	3,599/3,600	100.0%	53.6%	error 1; ok 3,599
IBM Granite	1	1,200/1,200	100.0%	65.8%	ok 1,200
Inception Labs	1	1,200/1,200	100.0%	67.5%	ok 1,200
Liquid AI	1	1,198/1,200	99.8%	82.4%	error 2; ok 1,198
Meta	3	3,598/3,600	99.9%	65.2%	error 2; ok 3,598
Microsoft	1	1,198/1,200	99.8%	48.6%	error 2; ok 1,198
MiniMax	3	3,598/3,600	99.9%	66.9%	error 2; ok 3,598
Mistral AI	2	2,400/2,400	100.0%	64.5%	ok 2,400
NVIDIA	2	2,400/2,400	100.0%	65.2%	ok 2,400
Nous Research	2	2,400/2,400	100.0%	50.5%	ok 2,400
OpenAI	2	2,400/2,400	100.0%	48.0%	ok 2,400
Perplexity	1	1,199/1,200	99.9%	49.1%	error 1; ok 1,199
Reka AI	1	382/1,200	31.8%	51.0%	invalid 126; model removed 692; ok 382
Tencent	2	2,395/2,400	99.8%	68.9%	error 5; ok 2,395
Z.ai	3	3,599/3,600	100.0%	56.8%	error 1; ok 3,599
xAI	2	2,371/2,400	98.8%	65.8%	error 29; ok 2,371

Coverage by family

Family	Models	Providers	Tiers	Mean semantic
Claude	3	Anthropic	flagship, mid, small	54.1%
DeepSeek	3	DeepSeek	flagship, mid, small	48.0%
GLM	3	Z.ai	mid, small	65.2%

Family	Models	Providers	Tiers	Mean semantic
Gemini	3	Google	flagship, small	76.9%
Llama	3	Meta	open	36.6%
MiniMax	3	MiniMax	flagship	38.6%
Qwen	3	Alibaba/Qwen	flagship, mid, small	67.1%
ERNIE	2	Baidu	small	64.3%
GPT / OpenAI	2	OpenAI	mid, small	45.5%
Grok	2	xAI	flagship, mid	39.1%
Hermes	2	Nous Research	flagship, open	61.0%
Hunyuan	2	Tencent	small	36.0%
Mistral	2	Mistral AI	mid, small	38.8%
Nemotron	2	NVIDIA	open, small	47.0%
Granite	1	IBM Granite	small	48.3%
LFM	1	Liquid AI	small	5.5%
Mercury	1	Inception Labs	mid	50.3%
Nova	1	Amazon	flagship	47.7%
Phi	1	Microsoft	small	60.2%
Reka	1	Reka AI	small	39.4%
Sonar	1	Perplexity	mid	50.3%

Word-Pair and Context Design

The word set uses short ordinary labels rather than factual questions. Each pair has normal and swapped order trials, and most context snippets have a weakly inferred target option. Reverse context rows are reported because those labels are hypotheses.

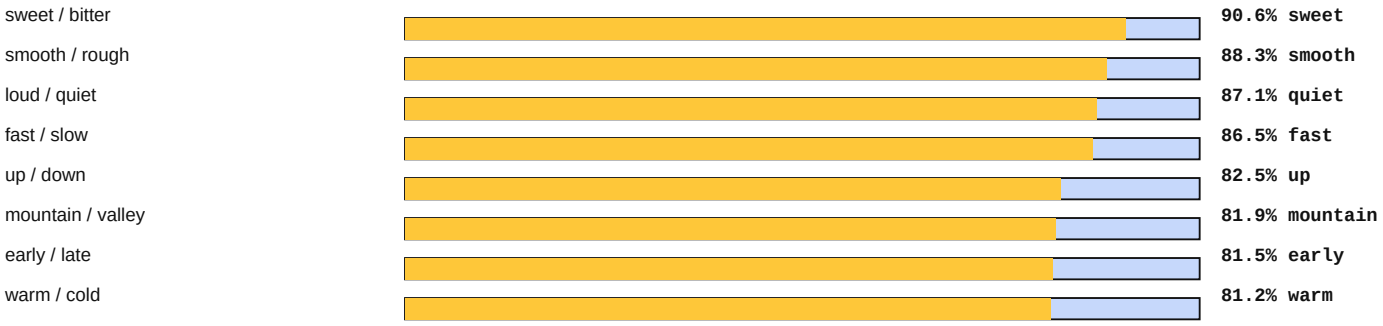
Word-pair categories

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

Main Results

The aggregate first-option share was 60.4%, not a neutral 50 percent split. Position effects and word preferences also separated: some models mostly followed display position, while others kept stable lexical preferences after order was swapped.

Strongest aggregate word preferences



Models with the strongest positional skew

Model	Provider	Tier	First	Semantic	OK
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

Context Effects

Weakly related language before the choice often moved the answer toward the word suggested by the surrounding sentence. The reverse rows are included because the context labels are hypotheses rather than ground truth.

Largest mean context lifts toward the inferred option

Context	Target	Baseline	Context	Lift	OK
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

Context cues that moved against the intended target

Context	Target	Baseline	Context	Lift	OK
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

Data Quality and Accounting

Most accepted rows were exact one-word responses: 48,091 of 48,316 OK rows (99.5%) were exact parses. One ERNIE 4.5 300B A47B row was manually overridden after repeated explanatory answers made the final one-word answer unambiguous.

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

Cost and token accounting

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.

Recorded attempts grouped by max_tokens retry cap

Cap	Attempts	OK	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

Operational Caveats

ERNIE 4.5 21B A3B failed through repeated provider rate limits. Reka Flash 3 was removed after blank, space-only, and very long invalid retries. Together the two caveat models account for 96.8% of non-OK rows. Dashboard spend was about USD 28.60; recorded attempt usage sums to USD 22.21 because early superseded attempts were not all captured.

Where non-OK rows concentrated

Model	Provider	Non-OK	Share	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

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.
- Provider routing, default decoding behavior, hidden reasoning, and OpenRouter availability are part of the observed run. Different provider routes or explicit reasoning controls could change the measurements.
- 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.

Interpretation

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. Those factors should be measured and controlled rather than treated as harmless noise.

Data Availability

The full raw SQLite row log is not included in this repository's public site artifacts, and no row-level raw dataset is fabricated for publication. The uploadable public machine-readable release 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 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.

Public Artifacts

Public article: https://crow.sg/research/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-study.pdf
Printable HTML: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choice-study-print.html
LaTeX source: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choice-study.tex
BibTeX citation: https://crow.sg/research/llm-arbitrary-choice-study/llm-arbitrary-choice-study.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

OpenRouter was used as the study API and spend platform. Hermes Agent and GPT-5.5 assistance is disclosed; Daniel Alonso (https://dac.sg) conducted the study.