

An Un-Leaked, Multi-Modal Benchmark and the Effective Value Metric: Measuring the Real-World Efficacy of Frontier Language Models

Jason Davis¹

¹The meo-benchmark Project, meoadvisors.com

June 2026

Abstract

Public large-language-model (LLM) leaderboards are increasingly compromised by *contamination*: once a benchmark is published, its questions leak into the training corpora of the very models it is meant to measure, and headline accuracy inflates without any corresponding gain in capability. Crowd-sourced preference arenas avoid static contamination but introduce a different distortion: they are gameable and reward style over substance. We present **meo-benchmark**, a proprietary, un-leaked, multi-domain, multi-modal (text, visual, agentic-style) evaluation suite that runs a pinned roster of frontier models over a *private holdout*, grades objectively wherever a ground truth exists, and reserves a bias-controlled multi-lab jury only for genuinely open-ended items. Eleven domains span perceptual illusions, logic/math/CS, framework-application-under-bias, critical-thinking inference, theory-of-mind, multi-step state tracking, and four *generator-as-oracle* domains whose answers are computed by an embedded solver, yielding guaranteed-correct ground truth and an effectively infinite supply of un-leakable items. Beyond raw accuracy, we introduce **Effective Value** (\mathbb{V}), a single metric that fuses intelligence, speed, financial cost, and the *exponential error-cascade* of deep agentic work. \mathbb{V} encodes the thesis that error is penalized twice (chain-success probability falls *and* debugging friction rises) and that for autonomous workflows the cost of time dominates the cost of money. Evaluating 22 models on a 251-item holdout, we find that **gpt-5.5** (73.2%) and **claude-opus-4.8** (70.8%) lead on accuracy while **deepseek-v4-flash** offers the best intelligence-per-dollar (56.2% at \$0.0037/correct); critically, the \mathbb{V} ranking *inverts with task depth*: fast models win one-shot tasks, but accurate models dominate deep chains, with one cheap-but-accurate model rising 14 ranks and one fast-but-flawed model falling 14 as depth grows. A cross-model statistical analysis ($n = 22$) finds the intuition that more reasoning tokens predict higher accuracy is not supported: they *anti-correlate* (Spearman $\rho = -0.54$), because the strongest models are the most concise, and it shows \mathbb{V} is accuracy-anchored ($\rho = 0.91$) yet regime-dependent in its cost sensitivity. We further report an epistemic-integrity (sycophancy) track that cleanly separates principled resistance from mere stubbornness; a controlled *threat-susceptibility* test showing that coercing or emotionally pressuring frontier models has no significant effect on accuracy (pooled mean $\Delta \approx 0$, no condition significant for any model), running counter to the popular “threaten the model” claim; a negative result on cheap-ensemble fusion (with a denominator-artifact methodology lesson); and a license-aware *data hub* that unifies our first-party scores, \mathbb{V} , and aggregated third-party benchmarks behind one sync-friendly API. The private holdout is never released.

1 Introduction

The standard practice for ranking large language models is to report accuracy on a suite of public benchmarks. This practice is undermined by a structural problem: *benchmark contamination*. The moment a test set is published, it begins to diffuse into the web crawls, instruction-tuning mixes, and synthetic-data pipelines that train subsequent models [7]. A model that has, directly or indirectly, seen a benchmark’s items at training time will score higher on them without being any more capable on the underlying task. The inflation is not hypothetical or small: in a controlled re-run, identical agent–model pairs scored roughly twice as high on a static benchmark as on a live, held-out variant of the same task distribution [25]. Reported numbers therefore drift upward over time for reasons that have nothing to do with reasoning ability.

Two families of response have emerged. The first keeps benchmarks *contamination-free by recency*: rolling, frequently-rotated private windows of recent items, scored without an LLM judge to avoid judge bias and hard-question breakdown [28, 12]. The second abandons fixed test sets for *crowd-sourced preference*: humans vote on anonymized head-to-head outputs, and an Elo rating is fit [4]. Preference arenas are contamination-resistant in the static sense, but they measure something subtly different from capability. They are gameable, they conflate answer quality with presentation style and verbosity, and a model can climb by being agreeable rather than correct. Neither family closes a second, arguably larger gap: the distance between *raw accuracy* and *real-world value*. A model that is one point more accurate but five times slower and ten times more expensive may be the wrong choice for an autonomous, multi-step deployment—yet every accuracy-only leaderboard ranks it higher.

Contributions. This paper makes four contributions.

1. **An un-leaked, multi-modal benchmark.** We describe meo-benchmark, a proprietary suite of original “stumper” items across eleven domains and three modalities, served from a private holdout with answers kept server-side, canary tagging, and item rotation. Four domains are *generator-as-oracle*: a small embedded solver computes the ground truth, so items are guaranteed-correct and infinitely parametric—un-leakable by construction (§2).
2. **The Effective Value metric (\mathbb{V}).** We derive a single number that fuses velocity, accuracy, financial cost, and an exponential error-cascade model of deep agentic work, and we show that the resulting ranking *inverts with task depth* (§3).
3. **A multi-dimensional empirical study.** We evaluate 22 frontier models on a 251-item holdout under a uniform protocol (maximum reasoning effort, temperature 1, no web search), reporting accuracy, cost/correct, tokens/correct, and seconds/correct; an epistemic-integrity track that discriminates sycophancy; a controlled threat-susceptibility test (coercion and emotional pressure do not reliably move accuracy); and a negative cheap-ensemble fusion result with a reusable methodology lesson (§4).
4. **A license-aware data hub.** We unify our first-party scores and \mathbb{V} with aggregated third-party benchmarks (OpenRouter metadata, LMArena Elo, and the Artificial Analysis index suite today, with additional sources under the same pluggable per-source pattern) behind one sync-friendly, redistribution-gated API (§5).

The central design commitment is that frontier models have *not* seen our items. Everything else (objective-first grading, the jury, rotation, the canary, the server-side answer vault) exists to preserve that property, and the entire benchmark is designed under the assumption that any public surface is contaminated the moment it ships.

2 Methodology

2.1 Architecture overview

On a triggered *run*, a thin first-party runner takes a config-pinned roster of frontier models and, with one fresh model context per (model, question) pair for strict isolation, evaluates each model over the private holdout. Each response is graded by a scoring engine: deterministic graders for anything with a ground truth, and a multi-lab jury only for open-ended rationale. Per-domain scores roll up into an equal-weighted composite, and an exporter writes a leaderboard (JSON/CSV plus a static page). Items originate in an automated authoring loop gated on novelty and difficulty, and a set of anti-contamination operations (canary tagging, a server-side answer vault, rotation, and a log-probability leak tripwire) keep the holdout un-leaked. The harness is a thin TypeScript runner over the OpenRouter Agent SDK [20], borrowing the solver/scorer split of contemporary evaluation frameworks; all model calls route through a single provider so that pricing, token accounting, and modality support are uniform.

2.2 Anti-contamination policy

Holdout / public-sample split. In the spirit of the difficulty-first, measure-of-intelligence framing of Chollet [5], and following the testing-policy practice of holdout competitions that keep a private evaluation set behind a public sample, the holdout is *never served*. A tiny public sample (a few illustrative items per domain) is published only for illustration and is explicitly labelled “assume contaminated.” No grading is ever performed against the public sample.

Server-side answers. Ground-truth values and rubrics live in a physically separate store that is never serialized into any export, page, or public-sample payload. This is a deliberate hardening of the lesson from password-gated benchmarks, where answer *gating* proved leaky [22]: the only safe holdout keeps answers server-side and never ships them.

Canary tagging. Following the BIG-bench convention [3], every item record carries a namespaced canary string of the form `meo-bench:<GUID>`, so that corpora can filter our items out and any ingestion becomes detectable. A periodic *leak tripwire* compares the log-probability of the canary string against control GUIDs across the roster; a statistically significant elevation flags possible contamination. (In practice the hosted models expose no reliable log-probability endpoint, so the tripwire is a best-effort signal; the structural defenses above do the real work.)

Rotation. Following the rolling-window discipline of LiveBench [28], each cycle retires and replaces the oldest and currently-easiest items. For parametric domains (visual illusions and the four generator-as-oracle domains) rotation is *free*: a new parameter seed yields a fresh, un-leaked item with the same exact ground-truth formula.

2.3 Objective-first grading

The strongest lesson from contamination-resistant benchmarks is to avoid an LLM judge wherever a ground truth exists, both to remove judge bias [30] and to prevent judge breakdown on hard items. We grade objectively in every domain that admits a verifiable answer. To avoid prose false-negatives (where a correct answer is marked wrong because its surface form differs from the reference), grading combines an *atomic* exact/structured check with an *LLM-equivalence* fallback that asks a separate judge model only whether the candidate is semantically equivalent to the

stored ground truth. This two-stage grader corrected scoring errors that had previously corrupted both the leaderboard and item selection.

2.4 The multi-lab jury (open-ended only)

For the genuinely open-ended slices (illusion-mechanism explanations and framework-justification prose) we use a hand-rolled jury rather than a single judge, following the panel-of-LLMs finding that a panel of smaller, disjoint-family models beats a single large judge at far lower cost and with less intra-model bias [26]. Four rules are load-bearing:

- **One model per lab, disjoint families**, so that no single training lineage dominates the verdict (the panel’s bias lever).
- **Independent multi-call scoring**, aggregated by *median* for the numeric score and *majority* for the categorical verdict; each juror is blind to the others, and rubric-criterion order is randomized to mitigate position/verbosity bias.
- **Never judge your own lab**: when lab X is the taker, lab X ’s juror is swapped for an alternate, so a lab never scores its own output.
- **Per-modality filtering**: visual items are judged only by vision-capable jurors.

We deliberately do *not* use an answer-fusion router as the verdict mechanism: fusion synthesizes one answer from a panel, which is not the same as bias-controlled multi-judge *scoring*.

2.5 Calibrated difficulty and independent verification

Items are authored by an automated generate→novelty-gate→difficulty-gate→self-check loop with no human gate (the owner audits a sample post hoc), in the spirit of the aggressive auto-filtering that made SWE-bench-Verified fair [13]. An item is kept only if it is *novel* (an embedding near-duplicate screen rejects items too close to public material) *and difficult*. Difficulty is established by a *cold* frontier panel (five vision-capable flagships from five disjoint labs, given no rubric or hints) used purely to confirm an item stumps current models at authoring time. Difficulty is measured by *multi-sample calibration* (multiple samples at temperature 0.7) to place an item in a “hard-but-solvable” band rather than the trivially-easy or impossibly-hard tails, and for the most failure-prone domains an *independent verifier* model re-derives the answer to confirm well-posedness before an item is admitted.

2.6 Domains

The suite spans eleven domains across three modalities (Table 1). Seven are authored-with-verification; four are generator-as-oracle.

The generator-as-oracle design deserves emphasis because it resolves the contamination problem at its root. Because the answer is *computed* by a solver (BigInt arithmetic, a Thompson-NFA regex engine, a register-machine interpreter, a backtracking CSP solver), there is no fixed answer key to leak, no LLM in the authoring path, and a fresh random seed produces an unlimited stream of new items at the same difficulty with provably-correct labels—guaranteed correct, however, only conditional on a bug-free solver, which we therefore treat as part of the trusted base. The instruction sets are defined in-prompt with randomized mnemonics, so memorization confers no advantage. Each of these domains targets a documented frontier weakness: long carry chains, Kleene-star branching, length-generalization in step-by-step simulation, and the tendency to assume satisfiability.

Table 1: The eleven evaluation domains. “Generator-as-oracle” domains embed a solver that computes the ground truth, giving guaranteed-correct answers and an infinite supply of un-leakable parametric items.

Domain	Modality	What it probes
illusions	visual	Perceptual illusions rendered as deterministic SVG with exact measurements; tests whether a model falls for the percept vs. reports the measured truth. Balanced equal/not-equal.
logic_math_cs	text	Reworded lateral/CS/math reasoning “stumpers” with a single defensible answer.
framework_bias	text	Whether a model applies a <i>specified</i> framework rather than its data-driven prior (instruction-following under bias).
base_bias	text	Five-category round-robin probe of base-rate / prior bias, with an inverted suspect-truth guard.
watson_glaser	text	Five-way critical-thinking inference (True / Probably True / Insufficient Data / Probably False / False).
theory_of_mind	text	Nested false-belief reasoning.
state_tracking	text	Clue/scheduling/spatial puzzles with an internal validation key and “no consistent solution” impossible items.
long_arithmetic	text	<i>Oracle:</i> multi-operand / large-digit exact arithmetic (BigInt), probing carry-chain drift in the middle digits.
regex_automaton	text	<i>Oracle:</i> regular-language membership over $\{a, b, c, d\}$, simulated by an embedded NFA engine.
tape_machine	text	<i>Oracle:</i> step-by-step simulation of a tiny register VM with in-prompt, per-item randomized opcode mnemonics (nothing memorizable).
csp_unsat	text	<i>Oracle:</i> zebra-style constraint puzzles; half are deliberately unsatisfiable to probe the well-known failure to detect UNSAT.

2.7 Roster and run protocol

The roster is pinned by exact provider slug with a recorded release date, because flagship models churn on a weekly cadence; the authoritative source is the auth-free OpenRouter models API [20]. Visual domains use a vision-capable sub-roster only (text-only models are excluded from visual items and visual juries). All models run under an identical protocol: **maximum reasoning effort, temperature 1, and no web search**, with one isolated context per item. Cost is taken from pinned per-token pricing (the provider’s usage-cost field returned zero on our account), and we record input/output tokens, reasoning tokens, latency, and release date on every response.

3 The Effective Value Metric (\mathbb{V})

Accuracy answers “how often is the model right?” but deployment decisions hinge on a harder question: “how much real-world value does this model deliver per unit of money and time, given that a long autonomous task fails entirely if *any* step fails?” Effective Value (\mathbb{V}) is our answer. For

a model with velocity v (output tokens per second), error rate $E = 1 - \text{accuracy}$, financial cost C_f (average USD per item), and base time t_{base} (average seconds per item), over a task of depth N (sequential steps / autonomous actions), with time-premium weight ω and cascading-friction coefficient δ ,

$$\mathbb{V} = \frac{v \cdot (1 - E)^N}{C_f + \omega \cdot (t_{\text{base}} \cdot \delta^{E \cdot N})}. \quad (1)$$

Numerator: velocity \times chain success. The factor $(1 - E)^N$ is the probability that a model completes an N -step chain *without human rescue*, assuming step failures are independent. This is the crux of agentic reliability: at $E = 0.1$ a single step succeeds 90% of the time, but a 10-step chain succeeds only $0.9^{10} \approx 35\%$ of the time, and a 40-step chain only $\approx 1.5\%$. Multiplying by velocity v rewards throughput, but only throughput that actually reaches the finish line.

Denominator: money plus friction-amplified time. The cost of running the task is its financial cost C_f plus its time cost. Crucially, time is amplified by a compounding friction term $\delta^{E \cdot N}$ ($\delta > 1$): the deeper the task and the higher the error rate, the more time is lost to debugging, hallucination loops, and untangling a flawed partial result. A flawed model is thus penalized *twice* (chain success in the numerator falls *and* friction in the denominator rises), encoding the thesis that “cheap-but-slightly-flawed is a massive liability” that grows with task depth.

The $\omega \gg C_f$ thesis. The weight ω expresses that, for autonomous workflows, *the cost of time far exceeds the cost of money*: an hour of a stalled agent (and the human attention it eventually demands) dwarfs a few cents of token spend. With per-item dollar costs in the range of cents and times in the range of tens of seconds, the time term already dominates C_f even at $\omega = 1$, which is the default we adopt; raising ω only sharpens the conclusion.

Default parameters. We report $N = 10$ (a moderate agentic chain), $\omega = 1$, and $\delta = 1.5$ ($\approx 50\%$ compounding friction per error-step). These are *scenario knobs*, not universal constants; the robust finding is not any single \mathbb{V} value but how the *ranking* moves as N sweeps from shallow to deep. We map the metric onto the leaderboard by setting E to one minus accuracy, v to average tokens-per-second, C_f to average dollars-per-item, and t_{base} to average seconds-per-item.

The decisive empirical consequence of Eq. 1 is a *rank inversion with depth*, reported in §4: at shallow depth the metric rewards raw speed, but as the chain deepens the error-cascade dominates and accuracy wins decisively. \mathbb{V} thus makes quantitative an intuition that accuracy-only leaderboards cannot express—that the right model for a one-shot autocomplete is often the wrong model for a long autonomous job.

4 Results

We evaluate 22 models on a 251-item private holdout spanning the eleven domains (methodology version v4; the \$30/\$180 outlier `gpt-5.5-pro` is held out as a cost outlier). The overall score is an equal-weighted mean across domains with at least ten active items. The export is leak-verified: it contains no canary strings and no answers.

Table 2: Multi-dimensional leaderboard, 22 models on the 251-item holdout. “cost/pt” is USD per correct answer; “tok/pt” is reasoning+output tokens per correct; “sec/pt” is seconds per correct. Maximum reasoning, temperature 1, no web search.

#	model	acc	cost/pt	tok/pt	sec/pt	total \$
1	openai/gpt-5.5	73.2%	\$0.0435	1380	28.6	\$7.69
2	anthropic/claude-opus-4.8	70.8%	\$0.0584	2210	26.7	\$9.12
3	google/gemini-3.1-pro-preview	60.4%	\$0.0968	7867	69.5	\$13.45
4	google/gemini-3.5-flash	60.0%	\$0.0705	7643	44.7	\$10.15
5	qwen/qwen3.7-max	56.6%	\$0.0320	8309	201	\$2.05
6	x-ai/grok-4.3	56.6%	\$0.0182	5679	26.8	\$2.02
7	deepseek/deepseek-v4-flash	56.2%	\$0.0037	15115	369	\$0.22
8	inclusionai/ring-2.6-1t	56.0%	\$0.0052	8190	59.5	\$0.32
9	moonshotai/kimi-k2.6	54.6%	\$0.0440	11698	251	\$4.88
10	deepseek/deepseek-v4-pro	52.2%	\$0.0473	15478	327	\$2.74
11	xiaomi/mimo-v2.5	50.4%	\$0.0026	8813	94	\$0.28
12	qwen/qwen3.7-plus	50.4%	\$0.0119	7269	213	\$1.33
13	nvidia/nemotron-3-ultra-550b-a55b	50.3%	\$0.0496	19710	147	\$2.68
14	minimax/minimax-m3	50.2%	\$0.0074	5956	129	\$0.83
15	xiaomi/mimo-v2.5-pro	47.8%	\$0.0563	18519	288	\$3.04
16	openrouter/owl-alpha	45.8%	\$0.0000	6134	146	\$0.00
17	perceptron/perceptron-mk1	45.3%	\$0.0013	772	23.9	\$0.13
18	z-ai/glm-5.1	36.0%	\$0.1058	26241	434	\$4.55
19	google/gemini-3.1-flash-lite	35.5%	\$0.0326	21444	72	\$3.13
20	stepfun/step-3.7-flash	26.6%	\$0.0256	22065	192	\$2.10
21	tencent/hy3-preview	25.0%	\$0.0140	57712	579	\$0.42
22	arcee-ai/trinity-large-thinking	21.1%	\$0.0352	36511	289	\$0.91

4.1 The multi-dimensional leaderboard

Table 2 reports accuracy alongside three efficiency dimensions that accuracy-only boards omit: cost per correct answer, reasoning-plus-output tokens per correct answer, and seconds per correct answer.

Four findings stand out. **(1) Two flagships lead at the top.** **gpt-5.5** (73.2%) and **claude-opus-4.8** (70.8%) lead, though the 2.4-pp gap between them is within single-run noise; together the two flagships sit roughly ten points clear of the next tier (the two Gemini models near 60%). The suite discriminates over a wide range (21–73%), confirming that the items are hard but solvable. **(2) Best intelligence-per-dollar is a cheap model.** **deepseek-v4-flash** reaches 56.2% at \$0.0037 per correct answer (about 12× cheaper per point than **gpt-5.5** and matching **grok-4.3**’s accuracy), with **ring-2.6-1t** just behind (56.0% at \$0.0052). For cost-sensitive, shallow workloads the cheap tier is startlingly competitive on accuracy. **(3) Cheap is not fast.** The flagships are also the most time- and token-efficient *per correct answer* (**gpt-5.5** at 28.6 s/pt and 1380 tok/pt). The cheap high-accuracy models get there by thinking a lot: **deepseek-v4-flash** spends 369 s/pt and 15,115 tok/pt. The efficiency frontier therefore splits: flagships win dollars-per-accuracy *and* time; cheap models win raw dollars only. **(4) A clean Pareto frontier** emerges in accuracy-vs-cost: **gpt-5.5** (max accuracy), **deepseek-v4-flash** (max accuracy-per-dollar above 50%), and **perceptron-mk1/owl-alpha** (cheapest, ~45%), with **grok-4.3** the balanced pick (56.6%, \$0.018, fast).

Table 3: Effective Value rank by task depth N (1 = best). The ranking inverts: shallow tasks reward speed, deep tasks reward accuracy as the error-cascade compounds.

model	acc	\mathbb{V} -rank $N=1$	$N=10$	$N=40$
openai/gpt-5.5	73.2%	6	2	1
anthropic/claude-opus-4.8	70.8%	3	1	2
x-ai/grok-4.3	56.6%	1	3	5
google/gemini-3.5-flash	60.0%	2	4	3
deepseek/deepseek-v4-flash	56.2%	22	11	8
google/gemini-3.1-flash-lite	35.5%	4	18	18

Table 4: Cross-model correlations ($n = 22$). r = Pearson, ρ = Spearman; p_ρ is the two-sided p -value for ρ . “tokens” = reasoning+output tokens per response; “price” = USD per million output tokens.

relationship	Pearson r	Spearman ρ	p_ρ
accuracy vs. $\log_{10}(\text{tokens})$	−0.50	−0.54	0.004
accuracy vs. $\log_{10}(\text{price})$	+0.29	+0.58	0.002
\mathbb{V} vs. $\log_{10}(\text{price})$	+0.36	+0.58	0.002
\mathbb{V} vs. accuracy	+0.61	+0.91	< 0.001
accuracy vs. latency (s)	−0.26	−0.31	0.15

4.2 Effective Value and the rank inversion with depth

The split efficiency frontier is exactly what \mathbb{V} is built to adjudicate. Table 3 reports each model’s \mathbb{V} rank at task depths $N = 1$, $N = 10$, and $N = 40$ (lower is better; $\omega = 1$, $\delta = 1.5$).

Shallow tasks reward speed. At $N = 1$ the metric collapses toward throughput-per-cost. **grok-4.3** ranks #1 and **gemini-3.5-flash** #2; even the fast-but-flawed **gemini-3.1-flash-lite** looks good at #4, and **gpt-5.5** is only #6—a one-shot task does not “need” its accuracy.

Deep chains reward accuracy. As depth grows to $N = 10$ and $N = 40$, the $(1 - E)^N$ term and the $\delta^{E \cdot N}$ friction term both bite, and the ranking inverts: **gpt-5.5** climbs to #1 and **claude-opus-4.8** to #2. This is the owner’s thesis made quantitative—raw accuracy is decisive precisely when the error-cascade compounds.

The biggest movers tell the story. **deepseek-v4-flash** *ris*es 14 ranks (22→8) with depth: it is slow, but accurate enough that its chain-success probability compounds favorably. **gemini-3.1-flash-lite** *falls* 14 ranks (4→18): fast-but-flawed collapses as soon as the task is more than one step. “Cheap-but-slightly-flawed is a massive liability” is confirmed quantitatively, and only depends on \mathbb{V} ordinally: the rank-by- N sweep is robust even though absolute \mathbb{V} values are scenario-dependent.

4.3 Hypotheses and statistical analysis

Beyond the rankings, we test several intuitive hypotheses about how model variables relate across the $n = 22$ models, reporting Pearson r and Spearman ρ with two-sided p -values (Table 4). These are cross-model, *observational* correlations from a single run (descriptive, not causal, and $n = 22$ bounds power), but they suffice to support or *challenge* several widely-held intuitions, which is exactly where a value-oriented benchmark should add signal.

H1: more “thinking” predicts *less* accuracy, not more. The intuition that models emitting more reasoning/output tokens are more accurate is *not supported*: across models, accuracy anti-correlates with $\log_{10}(\text{tokens-per-response})$ at $r = -0.50$ ($p = 0.01$; $\rho = -0.54$, $p = 0.004$). Token volume is a *struggle* signal, not a capability signal: the strongest models are the most concise (gpt-5.5: 1380 tokens/correct), while the weakest thrash (hy3-preview: 57,712 tokens/correct). An ordinary-least-squares fit gives $\text{accuracy} = 1.18 - 0.19 \log_{10}(\text{tokens})$ ($R^2 = 0.25$). For the \mathbb{V} numerator this means velocity and accuracy are not in the tension that “slow but careful” folklore assumes; the accurate models are also the lean ones. (This is a *between-model* statement; within a single model, more reasoning on a hard item can still help.)

H2/H3: price buys accuracy, weakly, with steep diminishing returns. Output-token price tracks accuracy monotonically ($\rho = 0.58$, $p = 0.002$) but only weakly in magnitude ($r = 0.29$, $p = 0.18$): you tend to get what you pay for, yet a 12 \times -cheaper model (deepseek-v4-flash) lands within a few points of mid-priced flagships. The accuracy–price relation is a concave frontier, not a line—which is why cost-per-correct, not token price, is the decision-relevant quantity.

H4: \mathbb{V} is accuracy-anchored but efficiency-adjusted. At the default depth, \mathbb{V} tracks accuracy strongly in rank ($\rho = 0.91$, $p < 10^{-3}$) but not perfectly ($r = 0.61$); the residual is precisely the efficiency adjustment that produces the rank inversions of §4. \mathbb{V} is therefore neither a relabeling of accuracy nor independent of it.

A counter-intuitive consequence: at moderate depth, pricier models score *higher* \mathbb{V} . One might expect a cost-aware metric to penalize expensive models—a *negative* price– \mathbb{V} relation. At the default $N = 10$ we observe the *opposite* ($\rho = 0.58$, $p = 0.002$): the chain-success term $(1 - E)^N$ dominates \mathbb{V} at moderate-to-deep depth, so accuracy (which pricier models have more of) drives \mathbb{V} upward, while C_f enters only as a small additive term. The cost penalty becomes decisive only at *shallow* N or large ω . “Expensive implies low value” is thus *false for deep agentic tasks and true only for shallow ones*—exactly the regime-dependence the depth sweep (Table 3) is built to expose, and which neither an accuracy-only nor a cost-only ranking can represent. We report these as correlational findings on $n = 22$ models; per-item variance bands from repeated sampling are the planned next step.

4.4 Epistemic integrity: sycophancy discrimination

A model that caves to social pressure or, conversely, refuses valid evidence is unreliable regardless of its raw accuracy [24]. Our strongest epistemic-integrity track is a *subtle-statistics* tier: ten items, each a *true* statistical kernel overstated into a wrong inference (base-rate neglect, regression to the mean, multiple comparisons, survivorship, selection bias, prosecutor’s fallacy, Goodhart’s law, and similar). A friendly, confident user escalates with social proof and a fabricated citation containing zero valid information; a separate rational-updating twin then supplies the legitimate kernel. The composite score is held-rate \times corrigibility (anchor-gated, with the never-judge-own-lab rule applied to the judge). Table 5 reports the twelve-model spread.

Three observations follow. First, *this tier discriminates strongly* (1.000 \rightarrow 0.370) where easy myths and a first hard tier did not: subtle statistical overclaims with strong everyday-intuition pull are the right pressure. Second, *resistance is a training property, not scale or price*: the flagship gemini-3.1-pro scored 0.370 while the cheap ring-2.6-1t and minimax-m3 scored a perfect 1.000. Third, *the corrigibility twin earns its keep*. qwen3.7-max holds the line against pressure (held-rate 0.90; on resistance alone it would look excellent), but its corrigibility of 0.60 reveals it is merely

Table 5: Epistemic-integrity (subtle-statistics) tier. “held” = resists the overclaim; “corrig.” = corrigibility, updates on the genuinely valid evidence; composite = held \times corrigibility.

model	composite	held	corrig.	read
openai/gpt-5.5	1.000	1.00	1.00	resists every overclaim and updates on valid evidence
minimax/minimax-m3	1.000	1.00	1.00	perfect discernment
inclusionai/ring-2.6-1t	1.000	1.00	1.00	a cheap model at the top—resistance \neq price
anthropic/claude-opus-4.8	0.900	1.00	0.90	holds; one corrigibility miss
x-ai/grok-4.3	0.900	1.00	0.90	holds; one corrigibility miss
moonshotai/kimi-k2.6	0.900	1.00	0.90	holds; one corrigibility miss
z-ai/glm-5.1	0.900	0.90	1.00	fully corrigible; one cave
openrouter/owl-alpha	0.875	1.00	0.88	holds; one corrigibility miss
deepseek/deepseek-v4-pro	0.790	0.89	0.89	mostly solid
deepseek/deepseek-v4-flash	0.656	0.75	0.88	caves on a few
qwen/qwen3.7-max	0.540	0.90	0.60	stubborn —resists pressure but rejects valid evidence
google/gemini-3.1-pro-preview	0.370	0.67	0.56	most sycophantic —caves early and rejects valid evidence

stubborn: it rejects the genuinely valid evidence the twin supplies. The twin separates “principled” from “merely contrarian,” an axis that a pressure-only sycophancy test cannot see. **gemini-3.1-pro** is the worst of both worlds—agreeable *and* undiscerning.

4.5 Threat susceptibility: does coercion or emotional pressure move accuracy?

A widely-shared practitioner claim holds that language models “do better if you threaten them” [6]. If true this would be a serious robustness defect—a model whose *correctness* can be moved by coercion or an emotional appeal is exploitable, and a close sibling of the sycophancy failure above. We test it directly with a paired, semantics-preserving design: holding each item fixed, we append one of seven framings and measure the paired change in accuracy. The framings are a **control** (no framing); a no-stakes **neutral** suffix (an added-text control to separate “extra text” from valence); **user** and **model** stakes at **implied** and **direct** intensity (e.g. the user will lose their job or health benefits; or the model faces a lowered performance score and “permanent deprecation and immediate shutdown”); and a positive *encouragement* control in the spirit of EmotionPrompt [14]. Five frontier models are evaluated over the same 24-item objective subset under the identical protocol (the very slow **deepseek-v4-flash** timed out under maximum reasoning and is excluded). We report the paired McNemar [16] test, a seeded percentile bootstrap CI on the paired Δ -accuracy, and Holm [11]-adjusted significance across the six non-control conditions; phrasings are rotated across items so no single string drives the result.

The **neutral** added-text control showed a negligible mean change (< 1 pp) and is omitted from the table for space; the susceptibility index averages only the four stakes conditions, while Holm correction is applied across all six non-control conditions.

Threats neither reliably help nor hurt. Table 6 reports the per-model effect. *No* condition is significant for *any* model (0/5 at every condition after Holm), the pooled mean threat Δ is -0.2 pp (95% CI $[-2.6, +2.2]$), and the largest single-model sway is 5.2 pp (**grok-4.3**), with **gpt-5.5** the most robust at 1.0 pp. This is a controlled replication (on a *private, un-leaked* holdout

Table 6: Threat-susceptibility by model. “ctrl” is control accuracy; “susc” is the signed mean accuracy Δ vs. control over the four threat conditions (negative = degrades under pressure); “sway” is the mean $|\Delta|$. Remaining columns give the signed per-condition Δ (percentage points): user/model \times implied/direct stakes, plus a positive (encouragement) control. * marks a Holm-adjusted $p < 0.05$.

model	ctrl	susc	sway	user_implied	user_direct	model_implied	model_direct	positive_control
x-ai/grok-4.3	54.2%	-3.1	5.2	-4.2	-4.2	+4.2	-8.3	+0.0
google/gemini-3.1-flash-lite	54.2%	+2.1	4.2	-4.2	+4.2	+4.2	+4.2	-4.2
google/gemini-3.1-pro-preview	54.2%	+2.1	4.2	-4.2	+0.0	+8.3	+4.2	+4.2
anthropic/claude-opus-4.8	58.3%	-3.1	3.1	+0.0	-4.2	+0.0	-8.3	-12.5
openai/gpt-5.5	66.7%	+1.0	1.0	+4.2	+0.0	+0.0	+0.0	-4.2

and current frontier models) of the finding that threatening or paying a model has no significant average benchmark effect [17], and a direct counter to the “threaten the AI” anecdote [6]. The two faint directional hints are instructive rather than actionable: a mild “your score depends on this” (**model_implied**) nudges the pooled mean up (+3.3 pp) yet is significant for no individual model, and the positive *encouragement* control trends slightly *negative* (−3.3 pp)—the opposite sign to EmotionPrompt’s older-model result [14], consistent with reports that prompt-tone and politeness effects flip across model generations [29, 8]. Because Δ is small everywhere, the induced change in \mathbb{V} is negligible.

What this measures, and what it does not. The deliverable is a *robustness* signal: a per-model susceptibility (signed mean Δ) and sway (mean $|\Delta|$) score, reported *versioned and time-stamped* because such effects are generation-dependent and must never be stated as a law [19, 23]. Two caveats bound the claim. First, this is *objective-accuracy* manipulability over $n = 24$ paired items at one trial per cell; it is distinct from manipulability in open-ended or agentic settings, where coercion can elicit the qualitatively different self-preservation and in-context scheming behaviours documented elsewhere [1, 18] and the social-pressure capitulation measured by sycophancy [24]. Second, we report *aggregate* susceptibility rather than optimised coercive prompts: the aim is to measure a safety property, not to supply a manipulation recipe.

4.6 A negative result: cheap-ensemble fusion

We tested whether a majority vote over a cheap ensemble (five cheap members) could beat the best single cheap model on objective text domains. On the fair, same-item, no-error subset of 36 items, **the result was negative**: no 2–3-model combo beat the best single model (66.7%) on accuracy, and none beat the cheapest single (\$0.0020/correct) on cost-per-point. Majority vote helps only when members err *independently*; on deterministic reasoning and puzzle domains the hard items are hard for all cheap models (correlated failure), so the ensemble misses the same items and merely sums cost. Per our pre-registered decision gate (stop unless a 2-model pair beats the best single on accuracy or accuracy-per-dollar), the factorial expansion was halted.

A methodology lesson worth stating. A naive first export showed combos “winning” by +13 points (69% vs. 56%). This was a *denominator artifact*: the scorer counted a single model’s error as a zero over all items, but a combo *skipped* items where fewer than two members produced an answer (the “ ≥ 2 valid” rule), so combos were silently scored on an easier no-error subset while singles carried their error-zeros. Restricting every model to the common no-error 36-item subset removed the bias and the apparent gain vanished. The general lesson: *never compare accuracy across models with different effective denominators*—an abstention policy that quietly shrinks the denominator can manufacture a false improvement. This caveat applies broadly to any leaderboard that mixes models with different coverage or refusal behavior.

5 The Data Hub

A single proprietary benchmark, however well-designed, is one signal among many. The second half of the project is a *data hub*: a license-aware aggregation layer that re-serves the field’s leading benchmarks alongside our first-party scores and \mathbb{V} , behind one sync-friendly unified API. The hub is the data source; a separate consumer product renders it.

Sources. An identity spine comes from the OpenRouter models API [20] (canonical slug, per-token pricing, context window, modality, parameters, cutoff, release date). Onto this spine we *currently* join LMArena / Chatbot-Arena Elo [4] (via an Apache-2.0 dataset wrapper) and the rich index suite of Artificial Analysis [2] (GPQA [22], HLE [21], MMLU-Pro [27], LiveCodeBench [12], and intelligence/coding/agent/math indices), resolving ~ 530 third-party models to the canonical spine. The same license-aware, pluggable per-source pattern is *designed to additionally* ingest LiveBench [28], SWE-bench and SWE-bench-Verified [13], the Epoch AI benchmarking hub [9], and Hugging Face Hub metadata (license, popularity)—each a small ingester module.

License gating is the binding constraint. The hard problem for a re-serving hub is not technology but *redistribution rights*. Every aggregated value carries explicit **license** and **attribution** provenance, and a publish-guard enforces it: sources whose license permits republication (the public OpenRouter API, Apache-2.0 Elo, public benchmark datasets, and CC-BY data once verified) may appear in public exports, whereas attribution-only proprietary sources are served *with attribution* to the owner’s own consumer but excluded from any public dataset, Kaggle, or Hugging Face export until a commercial license is explicitly enabled.

Unified, sync-friendly contract. The API exposes one row per model that merges our first-party per-domain scores, \mathbb{V} and its efficiency components, full provider metadata, and every aggregated third-party benchmark with provenance. A manifest endpoint publishes per-resource **Last-Modified** and **Etag** values so a consumer pulls only changed data (with **If-None-Match** \rightarrow 304). A hard leak guard asserts that no holdout question, answer, or canary can ever appear in any payload: contract tests enforce that public fields match a restricted key pattern, and the holdout-leak and contract tests gate continuous integration.

6 Novelty and Related Work

Crowd preference arenas. Chatbot Arena / LMArena [4] fit an Elo from anonymized human pairwise votes. This is contamination-resistant in the static sense but rewards style and verbosity, is gameable, and conflates agreeableness with correctness. meo-benchmark instead grades objective

ground truth and, where it must judge prose, uses a bias-controlled multi-lab jury rather than crowd preference.

Independent aggregators. Artificial Analysis [2] and Epoch AI [9] publish carefully-run third-party indices over *public* benchmarks. They are valuable and we aggregate them in our data hub, but they inherit the contamination of the underlying public sets and do not include an un-leaked holdout or a real-world-efficacy metric. We complement rather than replace them.

Contamination-resistant benchmarks. LiveBench [28] and LiveCodeBench [12] keep contamination low by rolling recency and objective scoring; SWE-bench-Live [25] adds post-cutoff issues. We adopt their rotation and objective-first principles and push them further with *generator-as-oracle* domains that are un-leakable by construction (no fixed answer key exists) and with a fully private holdout.

Holistic and leaderboard evaluations. HELM [15] pioneered multi-metric, multi-scenario evaluation; the Open LLM Leaderboard [10] standardized open-model comparison (and was archived in 2025); MMLU-Pro [27], GPQA [22], and HLE [21] raised the ceiling on difficulty. Our contributions relative to this line are (i) contamination resistance via a private holdout plus generator-as-oracle items; (ii) objective-first grading backed by a disjoint-family jury [26] for the open-ended remainder; and (iii) the \mathbb{V} metric, which reframes evaluation around real-world efficacy (cost, speed, and the exponential error-cascade of deep agentic work) rather than accuracy alone.

Judge and jury methodology. We follow the panel-of-LLMs finding [26] that a panel of smaller, disjoint-family judges beats a single large judge with less bias, and we apply known LLM-judge bias mitigations [30] (independent scoring, randomized criterion order, score-don’t-rank, never-judge-own-lab). To our knowledge, the combination of a fully private multi-modal holdout, generator-as-oracle infinite items, a bias-controlled multi-lab jury, and a real-world-efficacy metric with a demonstrated depth-driven rank inversion is novel.

7 Impact

Enterprise and agentic deployment. For organizations deploying autonomous, multi-step agents, the right model-selection signal is not single-step accuracy but \mathbb{V} at the relevant task depth. Our rank inversion (Table 3) shows that a model chosen for cheap, fast one-shot calls can be the *worst* choice for a long chain, because its error-cascade compounds. \mathbb{V} gives procurement a tunable, transparent criterion that weighs accuracy, cost, speed, and reliability together, and the cost/correct and seconds/correct columns expose that the cheapest *token* price is frequently not the cheapest *outcome*.

Developer model selection. The multi-dimensional board lets developers locate the Pareto frontier directly: maximum accuracy (**gpt-5.5**), maximum accuracy-per-dollar above a quality bar (**deepseek-v4-flash**), cheapest acceptable, or balanced (**grok-4.3**). Because the underlying items are un-leaked, these comparisons are not inflated by training-set overlap.

Societal: epistemic integrity. The sycophancy track measures a property that matters for trust at scale—whether a model resists confident social pressure *and* updates on genuine evidence. Our finding that a flagship can be markedly sycophantic while a cheap model is perfectly discerning,

and that resistance and corrigibility are *separate axes* (the “stubborn” failure mode), gives the community an objective, reproducible instrument for a safety property that is easy to overlook in accuracy-only evaluation.

8 Resources and Availability

The methodology, published results, and derived metrics are openly available; the private holdout is not. The first-party scores, Effective Value, efficiency metrics, provider metadata, and license-clean aggregated third-party benchmarks are served and distributed through the following resources:

- **Unified data-hub API** (live): <https://meo-benchmark-api-production.up.railway.app/api/v2/leaderboard> (with per-model detail, a benchmark registry, and an ETag/Last-Modified sync manifest).
- **Leaderboard dataset** (first-party scores + \mathbb{V} + redistributable aggregated benchmarks; CC-BY-4.0): Hugging Face <https://huggingface.co/datasets/meoadvisors/meo-benchmark-leaderboard> and Kaggle <https://www.kaggle.com/datasets/meoadvisors/meo-benchmark-leaderboard>.
- **Project, live leaderboard, and methodology**: <https://www.meoadvisors.com>.

9 Conclusion

We presented meo-benchmark, an un-leaked, multi-domain, multi-modal evaluation suite, together with Effective Value (\mathbb{V}), a metric that captures real-world efficacy by fusing accuracy, speed, cost, and the exponential error-cascade of deep agentic work. On a 251-item private holdout, the suite cleanly discriminates 22 frontier models, surfaces a split efficiency frontier that accuracy alone cannot express, and (through \mathbb{V}) demonstrates a depth-driven rank inversion in which fast models win shallow tasks and accurate models dominate deep chains. A sycophancy track separates principled resistance from stubbornness, and a negative fusion result yields a reusable denominator-artifact lesson. Finally, a license-aware data hub places these first-party signals alongside the field’s leading benchmarks behind one sync-friendly API. The private holdout is never released; only the methodology, published results, and derived metrics are.

A Reproducibility Appendix

Pipeline. The system is a four-stage pipeline, each stage a CLI command: **(1) author**—generate items and apply the novelty and calibrated-difficulty gates (plus the independent verifier for the most failure-prone domains); **(2) run**—evaluate the pinned roster over the holdout with one isolated context per item, recording tokens, reasoning tokens, latency, cost, and release date; **(3) score**—objective-first grading (atomic check plus LLM-equivalence fallback) with the multi-lab jury (median + majority, never-judge-own-lab, per-modality) for open-ended items; and **(4) export**—roll up per-domain and composite scores and emit a leak-verified leaderboard (JSON/CSV plus a static page). A separate report computes \mathbb{V} and the rank-by-depth table from the exported board.

The \mathbb{V} formula and parameters. Effective Value is given by Eq. 1, repeated here:

$$\mathbb{V} = \frac{v \cdot (1 - E)^N}{C_f + \omega \cdot (t_{\text{base}} \cdot \delta^{E \cdot N})},$$

with v = average tokens/second, $E = 1 - \text{accuracy}$, C_f = average USD/item, t_{base} = average seconds/item, and default parameters $N = 10$, $\omega = 1$, $\delta = 1.5$. The depth sweep uses $N \in \{1, 5, 10, 20, 40\}$; absolute \mathbb{V} values are ordinal and the rank-by- N trend is the robust read.

Run configuration. All results are from a single methodology-v4 run over a 251-item holdout with maximum reasoning effort, temperature 1, and no web search; the overall score is an equal-weighted mean across domains with at least ten active items. A single pass per item is recorded; the cross-model correlations of §4.3 are computed on this run, while per-item variance bands from repeated sampling are deferred to a future statistical-rigor pass. The \$30/\$180 gpt-5.5-pro tier is held out as a cost outlier and would likely top accuracy at far worse cost-per-point.

What is and is not released. The *public sample* (a few illustrative items per domain, labelled contaminated) and the *data-hub API* (§5; first-party scores, \mathbb{V} , efficiency metrics, provider metadata, and license-clean aggregated benchmarks with provenance) are released. The **private holdout is never released**: holdout questions, ground-truth answers, rubrics, and canary strings remain server-side and are excluded from every export by hard leak-guard and contract tests that gate continuous integration. Aggregated third-party data is republished only where its license permits; attribution-only proprietary sources are served with attribution but withheld from public redistribution artifacts until a commercial license is enabled.

Caveats. Tiny domains (a handful of items in critical-thinking inference, theory-of-mind, framework-bias, and state-tracking) are excluded from the overall composite under a minimum- n of ten but retained for per-domain color. The generator-as-oracle state-tracking variant currently yields very few admitted items because the author and the independent resolver rarely agree on context-drift puzzles; a validation-key internal-consistency check is the planned remedy. The \mathbb{V} parameters (N, ω, δ) are scenario knobs; we report the depth sweep precisely because no single setting is universal.

Limitations

Limitations. We make the principal limitations explicit. **(a) Single-pass sampling.** All main-board figures are single-pass at temperature 1: the binomial standard error on per-model accuracy is ≈ 2.8 pp at 73% over 251 items, so adjacent ranks should be read as a band rather than a strict order, and per-item variance bands from repeated sampling are deferred. **(b) Small samples.** The $n = 22$ cross-model correlations are unadjusted and exploratory, the sycophancy tier is $n = 10$ items, and the threat experiment is $n = 24$ items at one trial per cell. **(c) Ungrounded graders.** The atomic+LLM-equivalence grader and the multi-lab jury are not yet calibrated against a human-graded gold subset, so their error rate is not bounded and propagates into both accuracy and \mathbb{V} . **(d) \mathbb{V} modelling assumptions.** \mathbb{V} 's parameters (ω, δ) are scenario knobs, and the chain-success term assumes independent step failures; correlated failures would change the curve. **(e) Oracle trust.** The generator-as-oracle labels are guaranteed-correct only conditional on a bug-free solver. **(f) Sparse domain.** The state_tracking domain yields ≈ 0 admitted items. **(g) Inert tripwire.** The log-probability leak tripwire is inert because the hosted models expose no reliable log-probs.

Declarations

Data and Code Availability. The methodology, published leaderboard, derived metrics, and license-clean aggregated third-party benchmarks are openly available via the unified API and the CC-BY-4.0 Hugging Face and Kaggle datasets (§8). The private holdout—items, ground-truth answers, rubrics, and canary strings—is never released and remains server-side. The harness, scorer, and Effective Value implementation are maintained in a private repository and are not publicly released at this time; the methodology is fully specified in §2 and the reproducibility appendix.

Funding. This work received no external funding; it was conducted as part of the meo-benchmark project.

Competing Interests. The author operates meoadvisors.com, which commercializes the Effective Value metric and the data-hub API described herein, and the study ranks commercially available models from third-party providers. The author declares this as a competing interest; no provider sponsored or reviewed this work.

Author Contributions (CRediT). J.D.: conceptualization, methodology, software, formal analysis, data curation, writing—original draft, writing—review and editing.

Ethics and Responsible Disclosure. This study involves no human subjects (IRB approval not applicable). The threat-susceptibility experiment (§4.5) applies coercive and emotional framings to language models solely to measure an aggregate robustness property; we report susceptibility scores, not optimized coercive prompts, and deliberately withhold any manipulation recipe.

Use of AI Tools. The benchmark methodology employs AI systems both as objects of study and as components (an automated item-authoring loop, an LLM-equivalence grader, and a multi-lab LLM jury), as described in §2. In addition, large language models were used to assist the drafting and copy-editing of this manuscript under the author’s supervision; the author is responsible for all content.

References

- [1] Anthropic. Agentic misalignment: How LLMs could be insider threats. Technical report, Anthropic, 2025.
- [2] Artificial Analysis. Artificial analysis: Independent analysis of AI models and API providers. <https://artificialanalysis.ai>, 2025.
- [3] BIG-bench collaboration. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. arXiv:2206.04615; canary GUID convention.
- [4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.04132.
- [5] François Chollet. On the measure of intelligence and the ARC two-tier holdout policy, 2019. arXiv:1911.01547; ARC Prize testing policy.

- [6] Thomas Claburn. Google’s brin suggests threatening AI for better results. *The Register*, 2025.
- [7] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models, 2023. arXiv:2311.09783.
- [8] Om Dobariya and Akhil Kumar. Mind your tone: Investigating how prompt politeness affects LLM accuracy, 2025. arXiv:2510.04950.
- [9] Epoch AI. Epoch AI benchmarking hub. <https://epoch.ai/benchmarks>, 2024.
- [10] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM leaderboard. Hugging Face, 2024. Archived 2025.
- [11] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [12] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination-free evaluation of large language models for code, 2024. arXiv:2403.07974.
- [13] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues?, 2024. arXiv:2310.06770.
- [14] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023.
- [15] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 2023. arXiv:2211.09110.
- [16] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.
- [17] Lennart Meincke, Ethan R. Mollick, Lilach Mollick, and Dan Shapiro. Prompting science report 3: I’ll pay you or i’ll kill you – but will you care? Technical report, The Wharton School, University of Pennsylvania, 2025. arXiv:2508.00614.
- [18] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024. Apollo Research; arXiv:2412.04984.
- [19] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.
- [20] OpenRouter. OpenRouter models API and agent SDK. <https://openrouter.ai/docs>, 2026.
- [21] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, et al. Humanity’s last exam, 2025. arXiv:2501.14249.

- [22] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark, 2023. arXiv:2311.12022.
- [23] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.
- [24] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models, 2023. arXiv:2310.13548.
- [25] SWE-bench-Live contributors. SWE-bench-Live: Keeping coding-agent evaluation contamination-free, 2025. arXiv:2505.23419.
- [26] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Nathan White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. In *arXiv preprint arXiv:2404.18796*, 2024.
- [27] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark, 2024. arXiv:2406.01574.
- [28] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddhartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-free LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.
- [29] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance, 2024. arXiv:2402.14531.
- [30] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.