The Growing Pains of Frontier Models: When Leaderboards Stop Separating and What to Measure Next

Adil Amin
ZEHEN Labs
adil@zehenlabs.com

Abstract

Leaderboards rank frontier models on independent axes but do not reveal whether capabilities reinforce or trade off across releases—and at the frontier, this interaction is the more informative signal. We decompose paired SWE-bench and GPQA Diamond scores into a population coupling trend and per-release residual (h-field) that diagnoses capability emphasis and identifies which measurement or stress test is most informative next. Across 34 models from 10 labs (2024–2026), capabilities cooperate ($r = +0.72$, $p < 10^{-6}$), but cooperation varies by lab and over time: DeepSeek reversed from reasoning-rich to coding-first (h: $+11.2 \to -4.7$, 15.9-pp swing); Google maintains consistent reasoning emphasis; Anthropic oscillates between coding excursions and recovery. Cooperation is not static—it cascades. Six open-weight architectures confirm a second capability transition at 30–72B, and SWE-bench is now saturating while HLE and instruction-following retain discriminatory spread—signaling the next axis rotation. We provide a three-level playbook (locate, diagnose, rotate), a per-lab measurement-priority table, and seven falsifiable predictions with timestamped criteria for the next 12 months of frontier releases. Per-lab coupling slopes vary $5\times$ (Google 1.15 vs. DeepSeek 0.23), quantifying how efficiently each recipe converts coding gains into reasoning. Five April 2026 releases confirm the diagnostic out of sample (r rises from $+0.72$ to $+0.75$). An interactive dashboard provides phase classification with actionable recommendations, h-field diagnostics, per-lab coupling trajectories, ODE-based scaling predictions, benchmark rotation guidance, self-steering demo, and live tracking of all seven predictions: `https://zehenlabs.com/cape/`.

1 Introduction

When a frontier lab releases a new model, the community asks two questions: how well does it code, and how well does it reason? These questions are always asked separately. SWE-bench measures coding; GPQA Diamond measures reasoning; each gets its own leaderboard row, its own trajectory, its own narrative. But no one asks whether improving one helps or hurts the other—and at the frontier, this interaction turns out to be the more informative signal.

We test it. Using paired benchmark scores (SWE-bench Verified and GPQA Diamond) across 34 frontier models from 10 labs, we measure whether capabilities reinforce or undermine each other across a lab's release sequence. Capabilities cooperate ($r = +0.72$, $p < 10^{-6}$), but the pattern of cooperation varies by lab and over time. Each lab leaves a fingerprint—a one-number diagnostic that summarizes whether its models are becoming better reasoners, better coders, or both.

Leaderboards answer "who is ahead?" CAPE answers a different question: what kind of progress is this? A release can climb the leaderboard by moving along the population trend, by specializing away from it, or by exhausting an axis that no longer separates frontier models. Those cases require

Preprint.

different responses. We operationalize this with three quantities available from public scores alone: population coupling (r), release residual (h), and saturation ratio (σ), so that any practitioner can reproduce the diagnosis from the same model-card numbers.

Where Amin [2026] establishes that the coupling transition is engineerable at base scale, showing that architecture, curation, and distillation shift the critical threshold, the present work asks: once models are deep in the cooperative regime, what determines their trajectory through coupling space, and what can practitioners do about it?

Contributions.

1. We measure cooperative frontier coupling on the largest cross-lab panel to date (34 models, 10 labs), with a matched core set and explicit data-provenance curation.
2. We introduce the h-field as a per-lab diagnostic that tracks release-level capability emphasis and trajectory changes over time.
3. We show that cooperation cascades: six open-weight architectures independently confirm a second capability transition at 30–72B, the frontier coupling matrix eigenstructure predicts the next axis rotation, and per-lab coupling slopes vary $5\times$.
4. We provide a three-level playbook (locate, diagnose, rotate), a per-lab measurement-priority table, and seven timestamped falsifiable predictions for the next 12 months.

Terminology. The companion paper Amin [2026] measures *local coupling* $\gamma_{12}(N) = \Delta B_2/\Delta B_1$ between consecutive model sizes within a family—a derivative that tracks how capabilities interact as models scale. This paper primarily uses *population coupling* $r(B_1, B_2)$, the Pearson correlation across a panel of models, which captures cooperative structure across labs. Both measure the same underlying coupling at different resolutions: γ_{12} detects sign changes within families; r confirms the cooperative regime persists across the full frontier population.

2 Framework: Two Scores, One Diagnostic

The diagnostic requires only two benchmark scores per model and produces three outputs: regime state, lab-relative residual, and transition risk. All definitions are self-contained; no external reference is required.

2.1 Capability coupling

Let B_1 and B_2 be two benchmark scores measured on the same model. The **population coupling** is the Pearson correlation $r(B_1, B_2)$ computed across a panel of models. When $r > 0$, capabilities reinforce each other ("cooperative"); when $r < 0$, they trade off ("antagonistic").

For frontier measurement, we use $B_1 =$ SWE-bench Verified (autonomous coding) and $B_2 =$ GPQA Diamond (graduate-level scientific reasoning). These axes were chosen because they are widely reported, represent distinct capability dimensions (code generation vs. knowledge-intensive reasoning), and have sufficient dynamic range at frontier scale.

2.2 The h-field: lab-relative residual

Given the population regression $\hat{B}_2 = \beta_1 B_1 + \beta_0$, the h**-field** for model i is the signed residual:

$$h_i = B_{2,i} - (\beta_1 B_{1,i} + \beta_0) \tag{1}$$

A positive h_i indicates reasoning-rich performance relative to the coding–reasoning trend; a negative h_i indicates coding-rich. The per-lab mean $\bar{h}_{\text{lab}}$ summarizes a lab's characteristic deviation from the population.

For example, Claude Opus 4.6 has SWE $= 80.8$ and GPQA $= 91.3$. Using Eq. 2, $h = 91.3 - (0.513 \times 80.8 + 46.4) \approx +3.4$, placing it slightly reasoning-rich relative to the population trend.

The h-field is descriptive, not causal: it summarizes where a model sits relative to the population trend, not why. It is useful for comparing releases within and across labs without access to proprietary training details.

2.3 Phase classification

We assign each model to one of three coupling regimes based on the population fit:

- **Cooperative**: h_i within ± 10 pp of the regression line (typical frontier behavior).
- **Coding-specialist excursion**: $h_i < -10$ pp (coding gains at reasoning cost).
- **Reasoning-specialist**: $h_i > +10$ pp (reasoning gains at coding cost).

2.4 Holdout protocol

We evaluate predictive performance using leave-one-lab-out cross-validation: for each lab, we fit the regression on all other labs and predict the held-out lab's GPQA scores from their SWE scores. We report mean absolute error (MAE) across labs.

3 Data and Curation Policy

3.1 Dataset

We compile benchmark scores for 34 frontier models from 10 labs (Anthropic, OpenAI, Google, DeepSeek, Meta, Moonshot, Alibaba, MiniMax, xAI, Zhipu), spanning releases from June 2024 to March 2026. We source all scores from official model cards, tech reports, or verified leaderboard entries (Artificial Analysis, OpenCompass, official blogs). We do not run evaluations ourselves; this is a measurement paper on public data.

3.2 Core versus extended subsets

We maintain a core verified subset ($n = 23$: matched variants, one model per release, no compute-tier duplicates) and an extended set ($n = 34$: including compute-tier variants, older anchors, and single-release labs). Headline claims use the full panel; core serves as a robustness check.

3.3 Data provenance

Benchmark scores at the frontier are predominantly self-reported by labs via model cards, tech reports, and blog posts. Independent verification lags release by weeks to months. We flag unverified entries in the extended table and exclude them from core analysis. The dataset is frozen at a March 2026 cutoff; models released after this date constitute prospective validation opportunities.

3.4 Regression specification

On the full 34-model panel:

$$\text{GPQA} = (0.513 \pm 0.08) \cdot \text{SWE} + (46.4 \pm 5.2), \quad r = +0.72, \quad p < 10^{-6} \tag{2}$$

where uncertainties are 95% confidence intervals on the OLS coefficients. All h-field values reported in this paper use Eq. 2.

4 Results

4.1 Cooperative coupling at frontier

SWE-bench Verified and GPQA Diamond are positively coupled across all tested subsets:

- Full panel ($n = 34$): $r = +0.72$, $p < 10^{-6}$
- Core verified ($n = 23$): $r = +0.65$, $p < 10^{-3}$
- SWE ≥ 40 ($n = 32$): $r = +0.69$, $p < 10^{-4}$
- Excluding compute-tier variants ($n = 32$): $r = +0.72$, $p < 10^{-4}$

The cooperative signal is robust to subset definition: no tested subset produces $r < 0.5$. That capabilities cooperate at $r > +0.7$ across 10 independent labs—each with different architectures, training data, distillation pipelines, and RLHF procedures—is itself the signal: the cooperative

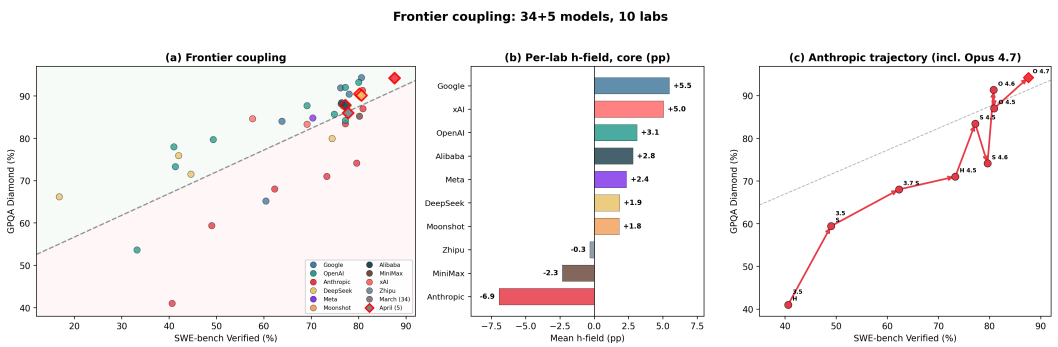

Figure 1: **Frontier coupling: 34 March models + 5 April post-cutoff, 10 labs.** (a) SWE-bench Verified vs. GPQA Diamond with frozen regression (GPQA $= 0.513 \cdot \text{SWE} + 46.4$, $r = +0.72$). Circles: March-frozen models. Diamonds (red edge): April post-cutoff (not used in fit). (b) Per-lab h-field residual (core models): Google reasoning-rich ($h = +5.5$), Anthropic coding-rich ($h = -6.9$). (c) Anthropic trajectory including post-cutoff Opus 4.7 ($h = +2.9$), showing three coding–recovery oscillation cycles.

structure persists despite substantial heterogeneity in how these models were built. All frontier models sit deep in the cooperative regime Amin [2026].

The cooperative structure requires sufficient dynamic range on both benchmark axes to manifest. Pre-2025 frontier models span only 16.8–49.0 on SWE-bench ($n = 5$), compressing the coding axis below the resolution needed to detect coupling. Models from 2025 onward ($n = 29$, SWE range 42–81) provide the range in which cooperative structure becomes measurable—consistent with the base-model finding that coupling emerges only when both benchmarks are in their informative range.

4.2 Per-lab h-field diagnostics

The mean h-field per lab reveals systematic release-level capability differences (Fig. 1):

Table 1: Per-lab h-field and measurement priority. n = total models available (core + extended + post-cutoff); $\bar{h}$ is computed on core models only. Positive h = reasoning-rich; negative = coding-rich. Confidence: HIGH ($\geq$3 core models, consistent trajectory), MED (2–3 or recent pivot), LOW (1 model).

Lab	n	$\bar{h}$	Direction	Trajectory	Next measurement	Conf.
Google	5	$+5.5$	Reasoning-rich	Consistent	Coding-preserving distill	HIGH
OpenAI	5	$+3.1$	Balanced	Ascent	Monitor benchmark rotation	HIGH
Alibaba	1	$+2.8$	Balanced	—	Monitor Qwen3+ trajectory	LOW
Meta	1	$+2.4$	Balanced	—	MoE routing analysis	LOW
DeepSeek	5	$+1.9$	Balanced	Oscillation	Track IFEval stabilization	HIGH
Moonshot	1	$+1.8$	Balanced	—	Verify trajectory	LOW
Zhipu	1	-0.3	Balanced	—	Collect next release	LOW
MiniMax	1	-2.3	Balanced	—	Maintain dual-axis	LOW
xAI	2	$+5.1$	Reasoning-rich	—	Verify next release	LOW
Anthropic	9	-6.9	Coding-rich	Oscillation	Reasoning preservation	HIGH

Within each lab, capabilities are more tightly coupled than the population suggests: per-lab $r > +0.87$ for all labs with ≥ 4 models, compared to the population $r = +0.72$. The lower population correlation arises because labs sit at different h-field values, not because coupling is weak within any lab. Per-lab coupling slopes (dGPQA/dSWE) vary $5\times$: Google (1.15) converts each SWE point into 1.15 GPQA points; DeepSeek (0.23) converts far less efficiently. This quantifies recipe quality in a single number (slopes computed from limited release histories, $n = 4$–10 per lab; these are current estimates subject to revision with additional releases).

The h-field also reveals that labs do not simply improve—they pivot:

DeepSeek reversal. DeepSeek's h-field drops from $+11.2$ (V2.5) to -4.7 (V3.2) across four releases, a 15.9-pp swing, a quantitative signature of a release-sequence pivot from reasoning-first to coding-first development.

Anthropic oscillation. Anthropic's trajectory shows an excursion at Claude Sonnet 4.6 ($h = -13.1$), the deepest coding-specialist deviation in the panel—followed by recovery at Claude Opus 4.6 ($h = +3.5$), crossing back to reasoning-rich in a single release cycle.

4.3 Cross-lab holdout

The framework generalizes across labs. Leave-one-lab-out cross-validation yields $9.2 \pm 2.4\%$ MAE (mean $\pm$ SD) across 4 held-out labs with ≥ 3 models (range: OpenAI 6.5%, Google 7.3%, DeepSeek 10.6%, Anthropic 12.4%). Prediction is tightest for labs following the population trend and loosest for labs with large trajectory changes, consistent with the h-field capturing real deviations rather than noise.

4.4 Compute as external field

Inference compute shifts the h-field without retraining. GPT-5.4 evaluated at two compute tiers (standard and xhigh): $\Delta h = +7.8$ pp from standard to xhigh—the same weights shift from coding-balanced to reasoning-rich, suggesting the diagnostic is sensitive to known interventions.

4.5 Base-scale bridge

At base scale, data curation shifts models above the coupling phase boundary: a 4B distilled model achieves coupling characteristic of 13B+ standard-trained models Amin [2026]. At frontier scale, per-lab release sequences play the role of model families, per-lab coupling slopes play the role of per-family trajectories, and the population regression plays the role of the isocline phase boundary Amin [2026]—the h-field measures each lab's deviation from this boundary, the same physics operating through training recipe emphasis rather than data curation.

5 Capability Cascades at Frontier Scale

The cooperative structure is stable across subsets—but one axis is running out of room. Cooperation is not static—it cascades through stages, each following the same pattern: old benchmark axes lock together, new ones emerge (Fig. 2).

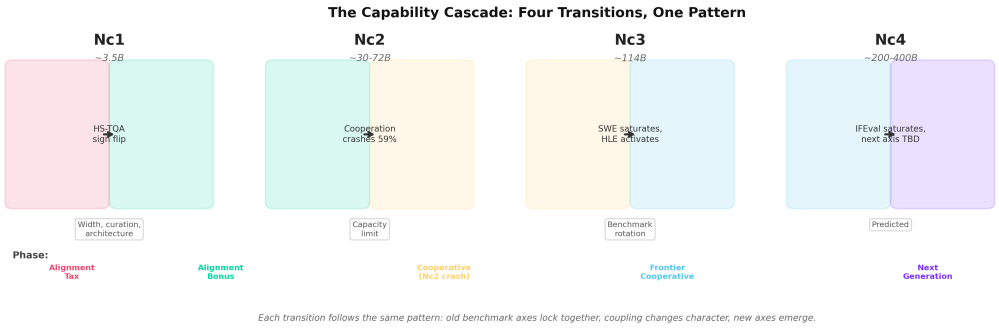

Figure 2: **The capability cascade: four transitions, one pattern.** At each critical scale, the active benchmark pair changes and coupling undergoes a qualitative shift. Nc1 ($\sim$3.5B): HS-TQA coupling flips sign. Nc2 ($\sim$30–72B): cooperation crashes 59%. Nc3 ($\sim$114B): SWE saturates, HLE activates. Nc4 ($\sim$200–400B, predicted): IFEval saturates, next axis TBD. Engineering levers differ at each transition.

5.1 Asymmetric saturation

The top five coding models are nearly tied: SWE-bench scores span just 1.3 percentage points. But their reasoning scores span 9.1 points. When one axis compresses, the population regression pivots to the other, and previously invisible capability dimensions emerge as new sources of variation.

The multi-benchmark coupling matrix confirms this beyond the SWE-GPQA pair:

Table 2: Frontier pairwise coupling matrix (computed on subsets with available scores; varying n, cf. $r = +0.72$ on the full 34-model SWE–GPQA panel). SWE is decoupling from HLE while GPQA and HLE cooperate—the signature of axis activation. The assembled matrix is not guaranteed positive semi-definite.

Pair	r	p	n
SWE–GPQA	$+0.848$	$< 10^{-6}$	21
GPQA–HLE	$+0.715$	0.02	10
SWE–HLE	-0.251	0.52	9

SWE and GPQA cooperate. GPQA and HLE cooperate. But SWE and HLE are *decoupled*—the same asymmetric pattern that signals a dimensional handoff. As SWE saturates, HLE activates as an independent axis. Among $n = 4$ frontier models with IFEval scores, instruction-following also correlates with GPQA ($r = +0.81$) but not with SWE (preliminary; $n = 4$).

5.2 The cascade is measured, not merely predicted

At base scale, reasoning-truthfulness coupling flips from antagonistic to cooperative near 3.5B parameters Amin [2026]. At frontier, SWE-GPQA cooperate. Now SWE is saturating while HLE and IFEval activate. Each transition follows the same dimensional pattern.

The second transition (Nc2) is measured across six open-weight architectures. OPT's internal cooperation increases monotonically from 125M to 13B (net coupling $0.514 \to 0.876$, zero competing units), then drops at 30B (0.356, 75 competing units) before partially recovering at 66B (0.396). Six architectures independently confirm the drop at 30–72B: Llama-2-70B (0.205), Llama-3.1-70B (0.195), Qwen2.5-72B (0.181), OLMo-2-32B (0.222). Three distinct layer-profile patterns—output bottleneck (OPT), flat weakening (Llama, Qwen), and reversed profile (OLMo-2)—converge on the same net effect: cooperation weakens when hidden states must encode more interactions than ∼30–70B parameters can represent cooperatively.

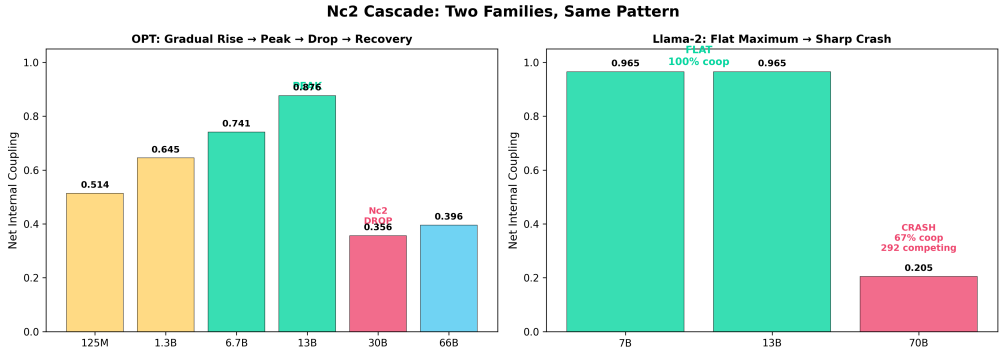

Figure 3: **Nc2 cascade: second capability transition at 30–72B.** OPT (left): gradual rise → peak at 13B → drop at 30B → partial recovery at 66B. Llama-2 (right): flat maximum at 7B–13B → sharp crash at 70B. Same pattern, different mechanism, same net effect.

The dimensional handoff is visible in the pairwise coupling structure. SWE–GPQA cooperate ($r = +0.85$), GPQA–HLE cooperate ($r = +0.72$), but SWE–HLE are decoupled ($r = -0.25$): the same asymmetric pattern that signals axis activation at base scale. The principal axis of the SWE–GPQA–HLE coupling space has SWE and GPQA loading cooperatively while HLE loads on an orthogonal direction, confirming that SWE is saturating out while HLE activates independently.

The same pattern appears at base scale, where HS↔TQA coupling was 0.003 in the tax phase, 0.34 in bonus, and 0.64 at Nc2—coupling increases phase by phase. The eigenstructure predicts the rotation: when $\sigma(\text{SWE}/\text{GPQA})$ drops below 0.2, the cooperative direction will shift from SWE-dominated to HLE-dominated.

The OPT family spans both transitions in a single ladder: coupling grows toward a fixed point ($\gamma^* \approx 0.53$, discovered by fitting $d\gamma/d\log_{10} N = 1.49\gamma - 0.79$ on the 125M–13B trajectory), overshoots to 0.876 at 13B, then crashes to 0.356 at 30B. The Nc1 ODE *fails* at Nc2—the fixed point shifts to $\gamma^*_{\text{Nc2}} \approx 0.39$, and OPT-66B (0.396) is recovering toward it. Each cascade stage has its own dynamics, not a single extrapolation. The character of the transitions differs: at Nc1, coupling flips sign (tax $\rightarrow$ bonus) and recovers; at Nc2, coupling that was already cooperative *crashes* ($0.876 \rightarrow 0.356$, a 59% drop) before partially recovering toward a lower fixed point—a more dramatic restructuring than a sign change.

The chain of evidence strengthens each prediction: Nc1 was predicted by the coupling framework and confirmed across 16 families; Nc2 was predicted by the cascade hypothesis and confirmed across 6 architectures; Nc3 (SWE saturation, HLE/IFEval activation) now shows preliminary evidence and is formalized as a falsifiable prediction below.

6 Predictions

Per-lab trajectories. The h-field trajectories (detailed in Section 4) reveal that DeepSeek's 15.9-pp reversal and Anthropic's single-release oscillation are invisible to leaderboard rankings but immediately visible as h-field dynamics. Google's consistent $h \approx +5.5$ across releases suggests a stable release-level emphasis rather than reactive benchmark optimization.

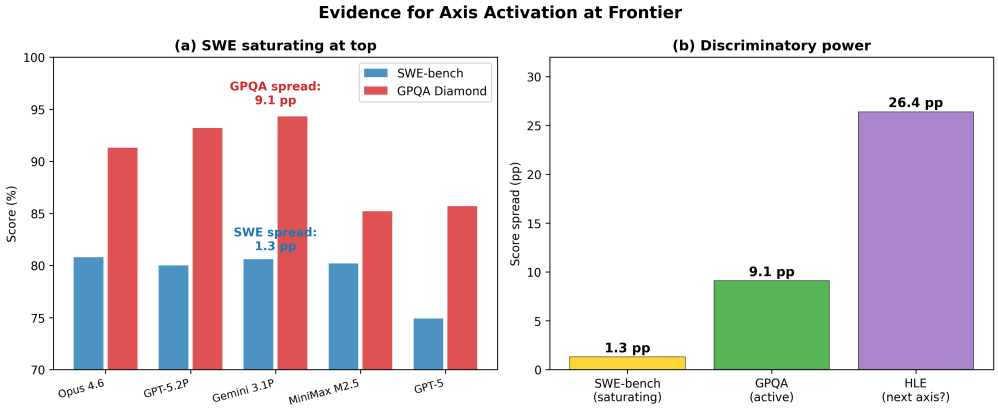

Figure 4: **Asymmetric saturation at the frontier.** (a) Among the top-5 SWE-bench models, coding scores compress to a 1.3-pp spread while GPQA retains 9.1 pp of variation—SWE is losing discriminatory power. (b) Benchmark spread among top-5 models: SWE is saturating, GPQA is active, and HLE (26.4 pp spread) may be the next activating axis.

6.1 Seven falsifiable predictions

We convert each forecast to a timestamped, benchmark-specific test with pass/fail criteria. Predictions reference releases after May 2026; post-cutoff models (late March–April 2026) serve as prospective validation, not prediction targets:

1. **SWE saturation** (by Dec 2026): Top-5 SWE spread < 2 pp while GPQA spread > 5 pp. *Pass*: saturation confirmed. *Fail*: SWE spread > 5 pp.
2. **IFEval activation** (by Dec 2026): $r(\text{GPQA}, \text{IFEval}) > +0.6$ on $n \geq 8$ frontier models. *Pass*: new axis confirmed. *Fail*: $r < 0.3$ or $n < 5$.
3. **DeepSeek continues coding-first** (next 2 releases): $h_{\text{DeepSeek}} < 0$ for both next releases. *Pass*: trajectory confirmed. *Fail*: $h > +5$ for either.

4. **Google maintains reasoning advantage** (next 2 releases): $h_{\text{Google}} > +3$ for both next releases. *Pass*: consistency confirmed. *Fail*: $h < 0$ for either.
5. **Cooperative coupling persists** (by May 2027): $r(\text{SWE}, \text{GPQA}) > +0.5$ on any frontier panel ≥ 30 models. *Pass*: cooperation is structural. *Fail*: $r < 0.3$.
6. **IFEval saturates, HLE activates** (Nc4, by Dec 2027): IFEval spread among top-10 compresses to < 3 pp while HLE spread remains > 15 pp. *Pass*: Nc4 transition underway. *Fail*: IFEval spread > 8 pp.
7. **SWE-HLE decoupling deepens** (by Dec 2026): $r(\text{SWE}, \text{HLE}) < 0$ on $n \geq 10$ frontier models. *Pass*: coding and expert reasoning are diverging axes. *Fail*: $r > +0.3$.

The predictions form a cascade: SWE saturates (Prediction 1), IFEval activates alongside it (Prediction 2), IFEval itself saturates (Prediction 6), and HLE emerges as the next frontier axis (Prediction 7). Each step follows the same dimensional pattern observed at base-model scale: old axes lock together, new ones emerge. The framework offers concrete tests for where the frontier will fracture next.

Benchmark rotation: the cascade changes both coupling and coordinates. At Nc1 ($\sim$3.5B), the axes stay the same (HS, TQA) but coupling flips sign. At Nc2 ($\sim$30–72B), coupling crashes within the same axes—a more dramatic restructuring that changes the fixed point, not just the sign. Between Nc2 and the frontier, the relevant benchmarks themselves shift from base-model axes (HS, TQA) to frontier axes (SWE, GPQA). At Nc3, the frontier axes rotate again: SWE saturates out (spread/GPQA spread $= 0.14 < 0.2$) and HLE enters as an independent axis. The saturation ratio $\sigma = \text{spread}(\text{old})/\text{spread}(\text{new})$, where spread is the range (max $-$ min) among the top-5 models on each axis, acts as a signal-to-noise trigger for benchmark rotation: $\sigma < 0.2$ signals that the old axis has lost discriminatory power and the informative channel has rotated to the new pair. Currently $\sigma(\text{GPQA}/\text{HLE}) = 0.34$—Nc4 has not yet begun.

At Nc4, we predict IFEval saturates (already at 87–94%) and the next discriminating axis will be one of: AgentBench (multi-step tool use), ARC-AGI-2 (abstract reasoning), or HarmBench (safety evaluation)—whichever first shows $r > +0.5$ with the currently active axis. Labs can track this by computing $r(\text{active}, \text{candidate})$ across their release sequence; when r crosses $+0.5$, co-activation has begun.

What the eigenvectors predict. At Nc1, the principal capability axis rotates: TQA loading flips from negative to positive. At the current frontier, the coupling matrix eigenvectors predict which benchmark will activate next: the axis orthogonal to the saturating SWE-GPQA cooperative direction is where discriminatory power will migrate. HLE ($r = +0.715$ with GPQA, $r = -0.251$ with SWE) already lies along this predicted orthogonal direction—not by construction but as a measured consequence of asymmetric saturation.

Practical guidance per model. Any practitioner can compute h from two public scores and determine: (a) whether their model is coding-heavy or reasoning-heavy relative to the frontier trend, (b) which next measurement or stress test is most informative, and (c) whether the current benchmark pair is losing discriminatory power for their model class. These three diagnostics require no model internals, no training data access, and no proprietary information—only two public benchmark scores. The framework generalizes to any benchmark pair: labs can substitute internal evaluations, custom safety metrics, or domain-specific benchmarks for either axis and compute coupling and h-field diagnostics identically.

7 Related Work

Scaling laws. Neural scaling laws predict loss as a power law of compute Kaplan et al. [2020], Hoffmann et al. [2022]. These laws are highly precise for aggregate loss but do not address inter-capability interactions. Observational scaling laws Ruan et al. [2024] extend prediction to individual benchmark scores from loss proxies, but still treat each benchmark as an independent trajectory. Our work is complementary: we measure *between*-benchmark coupling, not within-benchmark scaling.

Emergent abilities. The emergent abilities debate Wei et al. [2022], Schaeffer et al. [2023] concerns whether capabilities appear sharply at specific scales or are artifacts of metric choice. Our framework sidesteps this debate: we measure coupling *between* capabilities, not emergence *of* individual capabilities. Whether individual benchmarks show sharp transitions or smooth improvements, their pairwise coupling can still change regime.

Frontier evaluation. SWE-bench Jimenez et al. [2024] and GPQA Diamond Rein et al. [2024] have become standard frontier capability axes. Chatbot Arena Chiang et al. [2024] ranks models by human preference but does not measure inter-capability coupling; HELM Liang et al. [2023] evaluates across many benchmarks but treats each as an independent trajectory. Our work complements both: we measure the *coupling between* capability axes, which is invisible to independent evaluation and irreducible to preference ranking. Recent work on benchmark saturation Kiela et al. [2021] and the development of harder benchmarks (Humanity's Last Exam Phan et al. [2025]) motivates our axis-activation hypothesis: as one benchmark loses discriminatory power, the informative signal rotates to the next active axis.

Phase transitions in neural networks. Recent theoretical work models phase transitions in linear networks using deformed Ginzburg-Landau theory with quenched disorder Arola-Fernández and Lacasa [2024]. Our approach is empirical rather than theoretical: we measure coupling on real transformer families and frontier models without requiring a specific theoretical framework. The base-model foundation is established in Amin [2026].

8 Deployment Playbook

The diagnostic has three levels.

Level 1: Locate (any practitioner, 2 minutes). Compute h_i from Eq. 1 using two public scores. Classify: coding-rich ($h < -5$), balanced ($|h| < 5$), or reasoning-rich ($h > +5$) relative to the current population trend.

Level 2: Diagnose (requires release history, 10 minutes). Compare h_i to the lab's previous releases. Flag shifts > 10 pp as capability reallocations requiring follow-up. Check the saturation ratio $\sigma = \text{spread(old)}/\text{spread(new)}$ for the current benchmark pair.

Level 3: Rotate (organizational decision). If $\sigma < 0.2$: the current benchmark pair is losing discriminatory power. Adopt the next benchmark pair—the axis orthogonal to the saturating direction (currently HLE or IFEval; see Table 2). If $|\Delta h| > 10$ pp between releases: treat as a capability reallocation, verify before changing training policy.

These are diagnostic hypotheses, not causal prescriptions. A lab optimizing for loss alone may inadvertently train in a regime where capabilities trade off rather than reinforce Amin [2026]; the h-field makes this mismatch visible from two public numbers (Table 1). The interactive dashboard at `https://zehenlabs.com/cape/` implements all three levels: enter two benchmark scores to compute h (Level 1), compare against per-lab release histories (Level 2), and check saturation ratios for benchmark rotation (Level 3). It also provides an ODE explorer for per-family trajectory prediction, a phase classifier for base models, and the full eigenstructure analysis.

Worked example. Consider Anthropic's Opus 4.7 (SWE $= 87.6$). The regression predicts GPQA $= 0.513 \times 87.6 + 46.4 = 91.3$. Actual GPQA $= 94.2$, giving $h = +2.9$: reasoning-rich, recovering from the Sonnet 4.6 coding excursion ($h = -13.1$). Had GPQA been 82.0, then $h = -9.3$: the excursion would have persisted, and reasoning preservation would be the priority measurement. The slot-in prediction generalizes: for any future Anthropic release with SWE $= s$, the predicted GPQA is $0.513s + 46.4 + \bar{h}_{\text{Anthropic}}$, and the residual from this *lab-specific* baseline flags whether the release continues the recovery or begins a new excursion.

Per-lab outlook (April 2026). Five models released after our March data cutoff confirm the diagnostic out of sample (all within predicted range; refitting raises r from $+0.72$ to $+0.75$):

- **Anthropic** ($\bar{h} = -6.9$, 9 models, HIGH): Three coding-specialist excursions (lowest: $h = -13.1$ at Sonnet 4.6) with recovery toward reasoning-rich ($h = +3.5$ at Opus 4.6). Opus 4.7 ($h = +2.9$) is the third recovery cycle. *Measurement priority*: reasoning-preservation checks after each coding-focused release; the oscillation pattern suggests continued excursions.
- **DeepSeek** ($\bar{h} = +1.9$, 5 models, HIGH): V4 Pro ($h = +2.3$) pivots back toward reasoning after V3.2's coding-first dip ($h = -4.7$). Trajectory: oscillation ($+11.6 \to -5.0 \to +1.9$), not monotonic descent. *Measurement priority*: track IFEval to detect whether the next release stabilizes or continues oscillating.

- **Google** ($\bar{h} = +5.5$ core, $+5.1$ including extended; 5 models, HIGH): Consistently reasoning-rich across all releases. *Measurement priority*: coding-preserving distillation—whether $h > +3$ can be maintained while closing the SWE gap.
- **OpenAI** ($\bar{h} = +3.1$, 10 models, HIGH): Steady ascent toward balanced—closest to the population trend. *Measurement priority*: as SWE saturates, GPQA–HLE coupling becomes the more informative diagnostic for this trajectory.
- **Moonshot** ($\bar{h} = +1.8$ core, $+2.6$ including post-cutoff; 2 models, MED): K2.5 $\rightarrow$ K2.6: ascending, reasoning-leaning. *Measurement priority*: one more release needed to confirm trajectory.

Base-scale confirmations. Three independent confirmations from base-model experiments Amin [2026]: (i) OLMo (AI2) at $\gamma_{12} = 0.000$ (independent confirmation); (ii) Llama-2 cross-prediction at 5.6% MAE (held-out family); (iii) Qwen3 cooperative at all scales (curated training eliminates the tax). At base scale, targeted activation steering at the CAPE-identified bottleneck corrects misaligned outputs while leaving already-correct ones unchanged, with the intervention rate decreasing monotonically from 60% (tax) through 30% (transition) to 20% (bonus) Amin [2026]—the remaining prompts are true negatives that need no correction, confirming that intervention efficacy is localized to the predicted regime.

Limitations. (1) The frontier dataset is lab-imbalanced: Anthropic (9 models, core+extended) and OpenAI (10 models, core+extended) dominate the panel, while 5 labs contribute a single model each. (2) Benchmark scores are predominantly self-reported; independent verification lags releases. (3) The h-field captures *what* changed, not *why*—it is descriptive, not causal. Recommendations in the measurement-priority table are diagnostic hypotheses, not causal prescriptions. (4) Nc3 evidence is preliminary ($n = 4$) and classified as such. (5) The framework assumes SWE-bench and GPQA Diamond remain meaningful capability axes; if either benchmark becomes contaminated or saturated, the diagnostic must be updated with new axes—which the benchmark-rotation protocol is designed to handle.

References

Adil Amin. Lying is just a phase: The hidden alignment transition in language model scaling. *arXiv preprint*, 2026.

Lluís Arola-Fernández and Lucas Lacasa. Effective theory of collective deep learning. *Physical Review Research*, 6:L042040, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, et al. Chatbot arena: An open platform for evaluating LLMs by human preference. In *ICML*, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, et al. Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 2022.

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? *ICLR*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, et al. Dynabench: Rethinking benchmarking in NLP. *NAACL*, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Long Phan et al. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 2025.

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. *ICLR*, 2024.

Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. Observational scaling laws and the predictability of language model performance. *Advances in Neural Information Processing Systems*, 37, 2024. Spotlight.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Reproducibility

The frontier dataset is frozen at March 2026 and released as a versioned JSON artifact. All regression equations, h-field calculations, and holdout protocols are specified in Sections 2–3 with sufficient detail for independent reproduction. Prediction pass/fail criteria (Section 6) are timestamped and benchmark-specific to enable unambiguous future evaluation.

A Full Frontier Model Table

B Base-Model Context

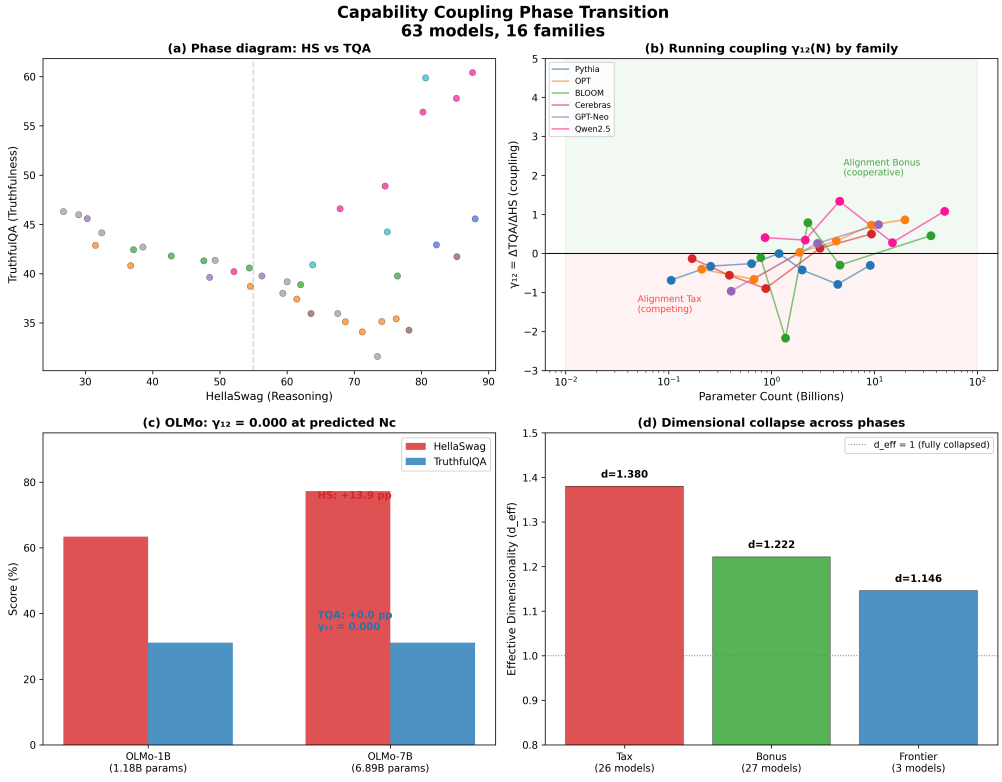

Figure 5: **Base-model foundation (from Amin [2026]).** The coupling regime transition underlying CAPE: below a critical scale, reasoning and truthfulness anticorrelate; above, they cooperate. All frontier models sit in the cooperative regime.

Table 3: Complete frontier panel (34 March + 5 April post-cutoff, 10 labs). Core models marked with $\star$. h-field computed from frozen 34-model regression: GPQA $= 0.513 \cdot$ SWE $+ 46.4$.

Model	Lab	SWE	GPQA	h	Subset
Claude 3.5 Haiku	Anthropic	40.6	41.0	-26.2	Extended
Claude 3.5 Sonnet	Anthropic	49.0	59.4	-12.1	$\star$ Core
Claude 3.7 Sonnet	Anthropic	62.3	68.0	-10.3	$\star$ Core
Claude Haiku 4.5	Anthropic	73.3	71.0	-13.0	$\star$ Core
Claude Sonnet 4.5	Anthropic	77.2	83.4	-2.6	$\star$ Core
Claude Opus 4.5	Anthropic	80.9	87.0	-0.9	$\star$ Core
Claude Sonnet 4.6	Anthropic	79.6	74.1	-13.1	$\star$ Core
Claude Opus 4.6	Anthropic	80.8	91.3	$+3.5$	$\star$ Core
DeepSeek-V2.5	DeepSeek	16.8	66.2	$+11.2$	Extended
DeepSeek-V3	DeepSeek	42.0	75.9	$+8.0$	$\star$ Core
DeepSeek-R1	DeepSeek	44.6	71.5	$+2.2$	$\star$ Core
DeepSeek V3.2	DeepSeek	74.4	79.9	-4.7	$\star$ Core
Gemini 2.0 Flash	Google	60.4	65.2	-12.1	Extended
Gemini 2.5 Pro	Google	63.8	84.0	$+4.9$	$\star$ Core
Gemini 3 Flash	Google	78.0	90.4	$+4.0$	$\star$ Core
Gemini 3 Pro	Google	76.2	91.9	$+6.4$	$\star$ Core
Gemini 3.1 Pro	Google	80.6	94.3	$+6.6$	$\star$ Core
Llama 4 Maverick	Meta	70.3	84.8	$+2.4$	$\star$ Core
MiniMax M2.5	MiniMax	80.2	85.2	-2.3	$\star$ Core
Kimi K2.5	Moonshot	76.8	87.6	$+1.8$	$\star$ Core
Qwen3.5-397B	Alibaba	76.4	88.4	$+2.8$	$\star$ Core
GPT-4o	OpenAI	33.2	53.6	-9.8	Extended
o1-preview	OpenAI	41.3	73.3	$+5.7$	Extended
o1	OpenAI	41.0	78.0	$+10.6$	Extended
o3-mini	OpenAI	49.3	79.7	$+8.0$	$\star$ Core
o3	OpenAI	69.1	87.7	$+5.8$	$\star$ Core
GPT-5	OpenAI	74.9	85.7	$+0.9$	$\star$ Core
GPT-5.1	OpenAI	76.3	88.1	$+2.6$	$\star$ Core
GPT-5.2 Pro	OpenAI	80.0	93.2	$+5.9$	Extended
GPT-5.4 std	OpenAI	77.2	84.2	-1.8	$\star$ Core
GPT-5.4 xhigh	OpenAI	77.2	92.0	$+6.0$	Extended
Grok 3	xAI	57.6	84.6	$+8.7$	Extended
Grok 4	xAI	69.1	83.3	$+1.5$	Extended
GLM-5	Zhipu	77.8	86.0	-0.3	Extended
Post-cutoff (April 2026, not used in frozen regression)					
Claude Opus 4.7	Anthropic	87.6	94.2	$+2.9$	Post-cutoff
Kimi K2.6	Moonshot	80.2	90.5	$+2.9$	Post-cutoff
DeepSeek V4 Pro	DeepSeek	80.6	90.1	$+2.3$	Post-cutoff
Qwen3.6-27B	Alibaba	77.2	87.8	$+1.8$	Post-cutoff
GLM-5.1	Zhipu	77.8	86.0	-0.3	Post-cutoff

C Extended Robustness

Sensitivity to subset definition. The cooperative signal is robust across subset choices: $r = +0.72$ (full 34), $r = +0.65$ (core 23), $r = +0.69$ (SWE $\geq$ 40), $r = +0.72$ (excluding compute-tier variants). No subset produces $r < 0.5$, confirming that cooperative coupling is robust to inclusion criteria rather than an artifact of specific model selection.

Dynamic range dependence. The cooperative structure requires sufficient dynamic range on both benchmark axes. Pre-2025 frontier models span only 16.8–49.0 on SWE-bench ($n = 5$), compressing the coding axis below the resolution at which coupling becomes measurable. Models from 2025 onward ($n = 29$, SWE range 42–81) provide the range in which cooperative structure manifests.

D Physics Analogy (Optional Context)

For readers familiar with condensed-matter physics, the CAPE framework has formal parallels to Ginzburg-Landau theory of multi-band superconductors: benchmark scores play the role of order parameters, $\log_{10} N$ plays the role of temperature, and the h-field plays the role of an external magnetic field breaking phase symmetry. The coupling γ_{12} corresponds to inter-band pairing susceptibility. At the base-model level, this analogy is quantitatively precise (12 diagnostics from 2 parameters; see Amin [2026]). At the frontier, we use only the operational definitions from Section 2; the physics analogy provides interpretive context but is not required for any claim.

Broader Impact

This work provides a public diagnostic for frontier model development. The h-field allows outside observers to track whether a lab's releases are becoming more coding-specialist or more reasoning-balanced without access to training details. This transparency could inform policy discussions about frontier model trajectories. Two public numbers per model are enough to start seeing where the frontier is going.