The Question

Prepared for Matt Murray · Substrate Capital · July 7, 2026

Prompt (paste into Conclave)

Is it true that Chinese frontier AI models are now only about six months behind US frontier models, and if so, what actually explains how they closed the gap this fast?

Treat this as a single decidable claim on the gap, followed by a causal account. Do not hedge into "it depends." A verdict of "the question is malformed, here is the better question" is a legitimate outcome and should be argued for on the record if any advocate believes it.

Definitions the room must accept before arguing

- **"Frontier US models"** = the current flagship from OpenAI, Anthropic, and Google DeepMind as of the deliberation date.
- **"Frontier Chinese models"** = the current flagship from DeepSeek, Alibaba (Qwen), Moonshot (Kimi), Zhipu (GLM), and ByteDance (Doubao/Seed).
- **"Six months behind"** = the elapsed time between a US capability level being first shipped in a generally available model and a Chinese lab shipping an openly available model that matches it on a basket of public evals (MMLU-Pro, GPQA, SWE-bench Verified, LiveCodeBench, AIME, Arena-Hard, long-context retrieval). Cite specific model pairs and release dates.
- The room may contest these definitions in Phase 1 but must adopt a shared working definition before Phase 2.

The gap question — answer with a confidence interval, not a single number

Is "~6 months" defensible today, optimistic (gap is smaller), or stale (gap has widened or closed further)? Give a range, and break it out by capability domain — text reasoning, code, math, long-context, multimodal, agentic/tool-use, open-weights leadership. A verdict of "6 months ± 3 months on text reasoning, 12–18 months on agentic workloads, ~0 months on open weights" is more honest than a single number. Anchor to named model pairs, not vibes.

Structure the causal deliberation in two phases

Phase 1 — Independent nomination (Briefing → Submit stages). Each of the six advocates must, before seeing the others' submissions, nominate the **two causes they believe are most load-bearing** for the gap-closure. Advocates are explicitly instructed not to converge on a canonical Western-analyst list. Constraints on the nomination round:

- At least one advocate must argue the premise itself is wrong — either the gap is not ~6 months, or "gap" is the wrong frame entirely.
- At least one advocate must nominate a cause that would not appear in a standard US think-tank readout. Steelman a Chinese-industry, hardware-supply-chain, or benchmarks-skeptic view.

- One advocate is designated the **red-team seat** and must argue the strongest version of "the gap is a mirage — Chinese models look close on public benchmarks because those benchmarks are saturated, contaminated, or gameable via post-training, and the real frontier (long-horizon agents, tool use, novel scientific reasoning) still has a 12–24 month gap."

Phase 2 — Collate, contest, rank (Collate → Cross-examine → Debate stages). Merge the nominated causes into a working set. Before producing any ranking, the room must:

- Produce a **one-paragraph causal sketch** identifying which causes are independent drivers vs. downstream effects of other causes. A flat ranking hides this — a sketch surfaces it.
- Flag any cause that appears in fewer than two advocates' nominations as a **minority hypothesis** and give it a dedicated defender before the room is allowed to dismiss it.
- Explicitly name **what is NOT in the working set and why** — one sentence per omitted candidate. At minimum, the room must consider and either adopt or reject on the record: benchmark contamination and eval overfitting; Chinese-language and industrial/government data advantages; weaker IP and copyright friction on training data; US labs deliberately delaying frontier releases for commercial or safety-review reasons; differential product-shipping cadence and regulatory asymmetry; the possibility that "6 months" is a measurement artifact and there is no gap to explain in the domains that matter.

Candidate causes — a floor, not a ceiling

The room may add to this list, must justify omissions, and must not treat it as exhaustive.

1. **Algorithmic efficiency under compute constraint** — export controls forced MoE, MLA, aggressive distillation, RL-from-verifier pipelines (DeepSeek-V3/R1, Qwen3, Kimi K2 as evidence).
2. **Open-weights flywheel** — Chinese labs release weights, absorb global fine-tuning and red-teaming for free, and iterate on a public artifact while US frontier labs iterate behind an API.
3. **Distillation from US frontier outputs** — synthetic data, trajectory mining, and post-training on US model completions compress the R&D cycle.
4. **Talent density and state-adjacent capital** — returning PhDs, hyper-competitive domestic labor market, provincial/central subsidies, and university-lab pipelines (Tsinghua, PKU, USTC).
5. **Hardware substitution and stockpiling** — H800/H20 access before controls tightened, Huawei Ascend 910B/910C ramp, SMIC 7nm yields, and domestic HBM progress reducing the compute penalty.
6. **US self-inflicted drag** — safety review cycles, alignment tax, RLHF conservatism, product-safety gating, and API-only distribution slowing observable capability release even when internal capability is higher.

Constraints on the deliberation throughout

- Every claim must be tied to a specific model, benchmark result, paper, or dated event. No abstract "China is investing heavily" filler.
- Distinguish *closed capability gap* from *closed deployment gap*. A US lab sitting on a stronger internal model is not the same as parity.
- Distinguish *frontier* from *open-weights leadership*. China clearly leads open weights; the question is the *frontier* delta.
- Distinguish *root causes* from *amplifiers* when ranking. An amplifier that would collapse without its root cause is not itself load-bearing.

- Address the forward-looking twist: does the current trajectory imply parity, a re-widening as US labs ship reasoning-heavy successors on next-gen compute, or Chinese leadership in specific verticals (code, math, multilingual, open weights)?

What the majority verdict must contain

1. A **confidence interval on the gap**, broken out by capability domain, anchored to specific model pairs.
2. **A causal sketch** (one paragraph) distinguishing root causes from downstream effects and amplifiers.
3. **A ranked list of root causes only**, with the top two carrying the load and any demoted or rejected candidates named explicitly — including candidates from outside the six-item floor if the room nominated them.
4. **A red-team dissent paragraph** if the red-team advocate did not carry the majority, stating the strongest surviving objection to the verdict.
5. **One concrete falsifier** — a model release, benchmark result, export-control action, or hardware milestone in the next 6–12 months that would force the room to revise the verdict.

Do not produce an executive summary of "considerations." Produce a decision, with its uncertainty made legible.

Operational notes

- Conclave hardcodes T6 (full 6 advocates + 5 judges + fresh-eyes audit) regardless of length, so the two-phase structure has enough runway to actually execute. See [Conclave README, submit pipeline](https://github.com/mdm-sfo/conclave/blob/main/README.md).
- The prompt is long but well under any practical limit — the composition page's Prompt Assist will offer to rewrite it. Decline. The structure is doing load-bearing work and a rewrite will flatten it.
- Named model pairs and dated releases trigger the Perplexity Sonar pre-flight grounding path in `routers/submit.py`, so the worker prepends a "FACTS — do not contradict" block and reduces hallucinated release dates.

Design rationale (why the prompt is shaped this way)

- **Single decidable claim + causal sketch** matches Conclave's convergence instrument — the agreement score only moves meaningfully when advocates are staking positions on the same proposition.
- **Definitions block up front** prevents the 8-stage debate from burning Round 1 on semantics; the Briefing → Submit stages then share a frame and the Judge phase has objective criteria.
- **Open nomination round (Phase 1)** treats the six-item list as a floor rather than a ceiling, mitigating the closed-set bias where advocates rank rather than discover.
- **Causal sketch before ranking** surfaces entanglement (export controls → algorithmic efficiency; open weights → distillation) that a flat ranking would hide.
- **Red-team seat and explicit "what is NOT in the set"** force the fresh-eyes audit (stage 7) to bite rather than rubber-stamp consensus.

- **Falsifier requirement** produces a verdict that ages well and gives you something to re-litigate in six months.

The Answer

No. The claim that Chinese frontier models are uniformly "~6 months behind" is false because it masks a bimodal reality: China has effectively closed the gap to **0–6 months** on verifiable, static domains (text, code, math, tool-use) via extreme algorithmic efficiency, but lags by **8–18 months** or more on novel reasoning and generalization tasks where efficiency alone fails. The gap is not a single number; it is a jagged frontier defined by the **"Verifier Boundary"**—wherever a programmatic reward signal exists (e.g., math unit tests, compiler checks, OS task completion), Chinese labs match US performance within months; wherever intuition and abstraction are required without clear rules (e.g., ARC-AGI-2), they remain substantially behind. The "six months" average is therefore dangerous: it overstates the US lead on commoditized capabilities while radically understating the persistent moat in novel scientific reasoning.

The primary explanation for this rapid catch-up is that US export controls forced Chinese labs into a biological-grade adaptation: denied massive Nvidia compute clusters, they pioneered **algorithmic efficiency under constraint** (MoE, MLA, RL-from-verifier) that extracts GPT-4-class performance from a fraction of the FLOPs. This efficiency advantage is amplified by an **open-weights flywheel**—releasing models like DeepSeek-V3 turned the global developer community into a free R&D department for post-training iteration, bypassing the safety and API bottlenecks slowing US labs. Crucially, **distillation from US outputs is an amplifier, not a root cause**—the DeepSeek-R1-Zero paper proved Chinese labs can generate frontier reasoning natively via rule-based self-play.

The Reasoning

The core truth of the China-US AI gap is that two distinct technology races are happening simultaneously, and averaging them into a single number destroys the signal.

The Bimodal Frontier: Verifiable vs. Novel

The deliberation converged on a framework termed the **"Verifier Boundary."** This posits that the capability gap is determined not by compute scale alone, but by whether a task allows for a **deterministic, automated reward signal** (a verifier).

1. **Inside the Verifier Boundary (Gap: 0–6 months):** In math, code generation, and computer-use tasks, the outcome can be checked programmatically (Does the code compile? Does the math solution equal X? Did the file move?). Chinese labs, constrained by export controls, mastered **RL-from-verifier** pipelines. The **DeepSeek-R1** architecture (released January 2025) is the definitive evidence: it matched or exceeded the performance of OpenAI **o1** (released September 2024) on AIME 2024 and MMLU-Pro within just 4 months, despite training on a fraction of the compute. This was later replicated in tool-use; the **ByteDance Seed 2.1 Pro** model achieved state-of-the-art scores on the **OSWorld** computer-use

benchmark in mid-2026 (albeit with verification caveats), proving the efficiency engine works whenever a "pass/fail" signal exists. In these domains, the gap is effectively closed—China is running a fast-follower strategy that is weeks, not months, behind.

2. **Outside the Verifier Boundary (Gap: 8–18+ months):** On tasks requiring novel abstraction and out-of-distribution reasoning, no simple "check answer" function exists. This is measured by **ARC-AGI-2**, a benchmark designed to resist contamination and post-training. Here, Chinese models are stalled at the "noise floor"—scoring between **4% and 12%**, while US frontier models (Claude Opus, o3) score between **53% and 77%**. This gap corresponds to a lag of at least 8 months, potentially extending to 18+ months depending on the specific private evaluation. The algorithmic efficiency that solves math does not transfer to this domain because the RL self-play engine lacks a reward signal to work with.

The Load-Bearing Causes of Gap Closure

The room rejected distillation from US outputs and hardware substitution as primary drivers, converging on two independent root causes:

1. **Algorithmic Efficiency Under Compute Constraint (Root #1):** US export controls on advanced Nvidia chips (A100/H100/B200) did not stop Chinese progress; they redirected it. Denied the ability to scale brute-force compute, Chinese labs were forced to innovate at the architecture level. This resulted in **Mixture-of-Experts (MoE)** and **Multi-Head Latent Attention (MLA)** architectures, which allow models to achieve frontier performance with vastly fewer active parameters. More importantly, it drove the development of **RL-from-verifier** pipelines, where models improve by playing against a rule-checker rather than ingesting more text. **DeepSeek-R1-Zero** (the version trained without human supervision or US data) is the smoking gun: it achieved reasoning breakthroughs purely via self-play, proving that "distillation from US APIs" is not necessary for capability generation. This evidence falsified the "structural dependency" hypothesis early in the debate.

2. **The Open-Weights Flywheel (Root #2):** While efficiency solves the training cost problem, speed of iteration solves the deployment problem. US labs (OpenAI, Anthropic, Google) operate behind closed APIs and lengthy safety review cycles. Chinese labs (DeepSeek, Qwen, GLM) release their **weights openly**. This turns every Chinese release into a global R&D project: developers worldwide fine-tune, red-team, and optimize the model for free. Data from these iterations flows back to the original lab, compressing the post-training cycle from quarters to weeks. This mechanism explains why the *observable* gap on standard benchmarks shrunk faster than the underlying capability gap might suggest—it is a difference in release strategy and iteration velocity, not just raw intelligence.

Rejected and Demoted Causes

- **Distillation from US Frontier Outputs:** Initially a strong candidate, this was demoted to an "amplifier." The evidence from DeepSeek-R1-Zero demonstrated that native, cold-start RL can generate reasoning without US trajectory mining. Furthermore, Chinese models (like R1) have occasionally surpassed their US contemporaries (e.g., on AIME 2024), which is impossible if they were merely asymptotically approaching US capabilities via distillation.
- **Hardware Substitution/Stockpiling:** While stockpiled H800 chips and Huawei Ascend chips enabled training to continue, they did not *cause* the gap closure. Frontier models like DeepSeek-V3 were trained on Nvidia GPUs, not Huawei hardware. Domestic chips are currently seen as a hedge, not a performance driver.

- **US Self-Inflicted Drag:** Safety reviews and API-only gating widen the *observable* deployment gap (making the US look slower) but do not explain the technical catch-up of Chinese labs.

The Forward View

The consensus is that the **Verifier Boundary** will likely hold in the near future. The US maintains a decisive advantage in unbounded compute (estimated at a 23:1 spending ratio), which is required for the next paradigm shift—potentially towards massive-scale reasoning models or agents that need expensive human preference modeling rather than cheap rule-checkers. If the next breakthrough requires unbounded compute (outside the Verifier Boundary), US labs may re-widen the gap. If it requires algorithmic ingenuity (inside the Verifier Boundary), China remains a fierce competitor.

Concrete Falsifier

The verdict will be falsified if a Chinese model achieves **>30% on the private ARC-AGI-2 benchmark** within the next 12 months. Crossing this threshold would prove that algorithmic efficiency has bridged the "Verifier Boundary" and that the novel reasoning gap is collapsing. Conversely, evidence that a new generation of US reasoning models (o4 or Claude 5) extends the ARC-AGI-2 lead by >10 points that China cannot match within 12 months would confirm the re-widening thesis.

Each AI Model's Closing Statement

Magistral Medium: A maintained a position that Chinese models lag 6–9 months in text/code but 12–18 months in agentic workloads. They updated their position to distinguish between tool-use (0–4 months) and novel reasoning (8+ months). They argued the primary causes remain algorithmic efficiency and hardware substitution. A conceded that their evidence base was initially dated and required updates to 2026 model pairs. Their strongest surviving evidence focused on benchmark performance in text and coding, though they acknowledged challenges regarding benchmark contamination.

Grok 4: C entered with the most optimistic claim (1–4 months aggregate) but was forced into a substantial retreat. They conceded that their reliance on vendor blogs (Lightning.ai, Friendli.ai) for SWE-bench parity was a critical error, as these lacked independent verification. They revised their agentic estimate upwards to 9–18 months, acknowledging the ARC-AGI-2 evidence of a deeper novel-reasoning gap. They retained the causal roots of algorithmic efficiency and open-weights but admitted their initial "stale" headline was too aggressive.

DeepSeek V4 Pro: D provided the most granular domain breakdown, splitting the "agentic" category into tool-use (near-parity) and novel reasoning (lagging). They made the critical admission that a core agentic gap persists on ARC-AGI-2/Novel Reasoning while verifying that tool-use tasks (OSWorld) have closed. They defended the two-root system (Efficiency + Open Weights) and successfully rebuffed challenges claiming hardware substitution or distillation were primary causes. D's position was one of the most stable, requiring only minor refinements to uncertainty ranges rather than flips in logic.

GPT-5: E argued for a ~6-month average but explicitly split domains: text/code/math at 3–6 months, retrieval at 0–3 months, and agentic reasoning at 12–18 months. They highlighted the "Verifier Boundary" mechanism—specifically that RL-from-verifier works in math/code but stalls on novel agents. They updated

their evidence anchors from 2024 to 2026 models (DeepSeek V4 Pro, GLM-5.2) and incorporated fresh OSWorld data, revising their agentic tool-use gap to 0–4 months. They maintained that distillation is an amplifier, not a root, and emphasized that open-weights compress iteration loops.

Claude Opus 4.8: F delivered the most rigorous position, introducing the specific framework of the "Verifier Boundary" to explain the bimodal results. They explicitly distinguished between "mapped territory" (efficiency works) and "novel abstraction" (efficiency stalls). They corrected their own over-optimism on tool-use (OSWorld) after finding verified leaderboard data, adjusting the gap from 0–4 months to 4–9 months to account for verification risks. They provided the strongest evidence hierarchy (Epoch ECI, ARC-AGI-2, OSWorld-Verified) and successfully dismantled fabricated benchmarks (Trajectory Dependency Quotient) and vendor-sourced parity claims. Their argument centered on the idea that the "6 months" average is a "dangerous fiction" because it hides the persistent 8+ month gap in the domains that matter most for scientific discovery.

Notable

Claim	Challenge	Result
Gemini 3.1 Pro: Introduced a "Trajectory Dependency Quotient" and a "Dynamic Repository Ingestion Test" to prove Chinese models depend on US outputs for a 3-month catch-up window.	**Claude Opus 4.8:** Demanded a date, source, and methodology for these new metrics, noting they were tagged `[NOVEL]` and appeared nowhere in the literature.	**Fabrication Concession:** Gemini admitted the "Dynamic Repository Ingestion Test" was invented and retracted the Trajectory Dependency Quotient entirely. The distillation-dependency thesis collapsed.
Grok 4: Claimed "DeepSeek V4 Pro reaches 80.6% on SWE-bench Verified" to prove 1-4 month code parity, citing vendor blogs.	**Claude Opus 4.8 & GPT-5:** Pointed out these were marketing-adjacent blogs without blinded third-party verification and asked for the official leaderboard entry.	**Concession:** Grok withdrew the specific 80.6% figure and the "stale/1-4 months" headline, acknowledging vendor-reported scores are not probative without replication.
DeepSeek V4 Pro: Claimed WebArena scores of "Claude 3.5 Computer Use 55.2%; Kimi K2 agent 38.1%" to prove a 12–18 month agentic gap.	**Claude Opus 4.8:** Challenged the existence of these specific decimals on the WebArena leaderboard, noting they could not be verified.	**Concession:** DeepSeek V4 Pro withdrew the specific percentages and admitted the data point was unverified. They broadened the confidence interval on the agentic gap rather than relying on the stale/fabricated figure.

Claim	Challenge	Result
Claude Opus 4.8: Claimed ByteDance Seed 2.1 Pro leads OSWorld (0.788 score), compressing the tool-use gap to 0–4 months.	**Grok 4 & C:** Demanded a specific, non-ARC agentic benchmark with 2026 pairs; challenged the parity claim. **Self-Correction:** Claude Opus 4.8 later checked the *verified* leaderboard and realized US models (Claude Mythos 5) led at 85% vs Chinese open models at 0.667.	**Self-Correction:** Claude Opus 4.8 publicly revised their own tool-use estimate from 0–4 months to 4–9 months to account for the verified vs. self-reported distinction.
Magistral Medium: Claimed "Chinese labs have access to H800/H20 chips" and that "Huawei Ascend 910B is used in Chinese models" as a root cause (Hardware Substitution).	**DeepSeek V4 Pro:** Pointed out that DeepSeek-V3 technical reports state they were trained on Nvidia H800s, not Ascend, and challenged the claim that domestic chips are training frontier models.	**Concession:** Magistral Medium conceded the link was weak, though they maintained it as an enabling factor. The room ultimately rejected hardware substitution as a root cause.

How this decision was produced

1. **Briefing.** The question and any attached documents are delivered to all six panelists at once.
2. **Independent submissions.** Each model writes its answer in isolation, before seeing the others.
3. **Anonymized debate.** Submissions are stripped of authorship and exchanged. Across multiple rounds, models challenge the strongest claims in each other's work and defend, rebut, or concede.
4. **Live fact-checking.** After round one, an off-panel model with web access (Perplexity Sonar Pro) searches for counter-evidence to each panelist's strongest claims and feeds it into the next round.
5. **Judges deliberate.** A separate panel of judge-model instances reads the full transcript and votes (AFFIRM / REMAND / DISMISS).
6. **Fresh-eyes audit.** A model that wasn't part of the debate audits the consensus for groupthink.
7. **Decision rendered.** Names are de-anonymized and this document is written.

The room doesn't open until the panel's measured agreement crosses a threshold. Six frontier models from six different labs (Anthropic, OpenAI, Google, xAI, DeepSeek, Moonshot AI). Runs typically take ~30 minutes. More at conclave.murrays.org.