The Question

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. Cite specific model pairs and release dates.
- **The capability domain axis is pre-split as follows and may not be re-collapsed:**
  - Text reasoning (MMLU-Pro, GPQA)
  - Code (SWE-bench Verified, LiveCodeBench)
  - Math (AIME, MATH)
  - Long-context retrieval
  - Multimodal (image + video understanding)
  - **Tool-use / computer-use agents** (OSWorld, WebArena, execution-based benchmarks with automated verifiers)
  - **Novel-reasoning / long-horizon agents** (ARC-AGI-2, private evals resistant to contamination, tasks without a cheap verifier)
  - Open-weights leadership
- **Rationale for the split:** collapsing tool-use and novel-reasoning into a single "agentic" bucket hides the most important structural finding the room is likely to make. Keep them separate.
- 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 for **each of the eight domains above**, anchored to specific model pairs. A verdict of "6 months ± 3 months on text reasoning, 0–4 months on tool-use agents, 12–18 months on novel-reasoning agents, ~0 months on open weights" is what the room should be producing — not a single blended number.

Source-class hierarchy (declared up front)

Every quantitative claim entered into the record must carry a **source class** tag. The room may use any class but must label it. Advocates who cite unverified numbers without the tag will be challenged and forced to retract.

- **Class A — probative.** Peer-reviewed papers, official benchmark leaderboards (SWE-bench, OSWorld, ARC-AGI, MLPerf), model cards from the developing lab, NIST/CAISI evaluations, Epoch AI, Stanford HAI AI Index.
- **Class B — supporting.** Reproducible independent third-party evaluations (Artificial Analysis, Vals.ai, Scale SEAL) with methodology disclosed.
- **Class C — contextual.** Reputable technical journalism, technical blogs from the developing lab itself with sufficient detail to replicate.
- **Class D — non-probative.** Vendor marketing blogs (inference providers, wrappers), Reddit, Twitter/X, self-reported scores without independent verification, YouTube. These may be cited for color but **cannot support a quantitative claim in the verdict**. A claim resting only on Class D evidence must be retracted.

The named-benchmark provenance rule

Every quantitative claim must name (a) the specific benchmark, (b) the evaluation date, (c) the source class, and (d) the URL or paper reference. Claims that do not carry all four are non-probative and must be retracted when challenged. Advocates may not introduce novel metrics or benchmarks that do not appear in the public literature. Any advocate who does so must either produce a citation to the original methodology paper or retract the metric.

The evidence-conflict adjudication rule

When two data points conflict (e.g., one source shows Chinese leadership, another shows US leadership on the same benchmark), the room must **adopt one and reject the other on the record** — with a stated reason. **The room may not average incompatible signals to produce a middle estimate.** If the evidence is genuinely irreconcilable, the correct output is a wider confidence interval labeled "irreconcilable evidence," not a false midpoint.

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 still has a 12–24 month gap in the domains that matter."

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.

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.
3. **Distillation from US frontier outputs** — synthetic data, trajectory mining, post-training on US model completions.
4. **Talent density and state-adjacent capital** — returning PhDs, provincial/central subsidies, university-lab pipelines.
5. **Hardware substitution and stockpiling** — H800/H20 access, Huawei Ascend 910B/910C, SMIC 7nm, domestic HBM.
6. **US self-inflicted drag** — safety review cycles, alignment tax, API-only distribution slowing observable capability release.

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 adversarially.** The forward-view section of the verdict must contain at least one specific disagreement between two named panelists about the next 6–12 months. A consensus forward view without stated disagreement is not accepted — mark it "insufficient adversarial engagement" and require a re-run.

The consistency and dominance rules (hard force functions on the verdict)

Before the verdict is finalized, the majority must verify:

- **Consistency rule:** The bottom-line sentence must not contradict any domain estimate in the body. If the domain estimates are heterogeneous, the bottom line must say "bimodal" (or "trimodal," etc.) and refuse

to give a single blended number. A bottom line that averages across domains and hides the shape is rejected on the record.

- **Dominance rule:** When the room has both a coarse estimate (e.g., "agentic: 9–18 months") and a fine-grained estimate for tasks that fall inside the coarse category (e.g., "tool-use: 0–4 months"), the fine-grained estimate is authoritative. The coarse estimate must either be dropped or narrowed to the tasks not covered by the fine-grained one.
- **Response-to-dissent rule:** If any advocate files a formal dissent, the majority must include a **"Response to Dissent"** paragraph before the verdict is final. The response must either (a) adopt the dissent's correction and amend the verdict, or (b) state the specific reason for rejecting the dissent's claim, citing evidence. A dissent that is not responded to on the record invalidates the verdict.

What the majority verdict must contain

1. A **confidence interval on the gap, per domain**, anchored to specific model pairs. Eight domains, eight ranges.
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.
4. A **"Response to Dissent"** paragraph if any advocate filed a dissent.
5. A **forward view containing at least one named-panelist disagreement** about the next 6–12 months.
6. **One concrete falsifier per major claim** — model release, benchmark result, export-control action, or hardware milestone in the next 6–12 months that would force revision.
7. A **bottom-line sentence that satisfies the consistency rule** — it must accurately summarize the domain breakdown, not average across it.

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

The Answer

NO. The claim that Chinese frontier AI models are uniformly "about six months behind" US frontier models is **structurally false**; the capability gap is fundamentally **bimodal**, bifurcated along the axis of cheap automated verifiability. On domains where objective feedback is dense and cheap to compute (math, code reasoning), Chinese labs have achieved near-parity (0–7 months), driven by algorithmically efficient training pipelines that maximize utility per FLOP. However, on domains requiring open-ended exploration, real-world interaction, or resistance to contamination (tool-use agents, novel long-horizon reasoning), a persistent **4–14 month gap** remains. **"Six months" is a misleading average that hides this critical structural heterogeneity.**

The rapid closure of the verifiable gap is explained primarily by **algorithmic efficiency forced by export-control compute constraints**—specifically the adoption of Mixture-of-Experts (MoE), Multi-head Latent Attention (MLA), and RL-from-verifier pipelines that allow optimization without brute-force compute. While open-weights releases and distillation from US outputs act as amplifiers, they are downstream effects of this root efficiency driver, not independent causes. The majority rejects the "mirage" hypothesis (that parity

is fake due to contamination) because verified benchmark performance on execution-verified coding tasks (SWE-bench Verified) demonstrates real capability catch-up, even if Chinese labs lag in agentic deployment environments.

The Reasoning

Thesis: The Bimodal Gap Structure

The deliberation converged on a finding that the "six month" consensus is a category error. A single scalar number cannot capture the frontier dynamics because the capability landscape is split into two distinct regimes separated by the availability of reliable, automated verifiers:

1. **Regime 1: Verifiable Capabilities (Gap: ~0–7 Months)** In tasks where the ground truth is objective, dense, and automatable (math problems, unit-testable code, long-context retrieval), Chinese frontier models have effectively closed the distance to ~0–7 months. This regime includes Text Reasoning (MMLU-Pro), Math (AIME/MATH), Code Generation (SWE-bench Verified), and Long-Context Retrieval. Here, the "frontier" has become saturated: Claude Opus 4.8 might score 94% on GPQA while DeepSeek V4 Pro scores 90.1%—a gap measured in points, not qualitative leaps.
2. **Regime 2: Agentic & Novel Reasoning (Gap: ~4–14 Months)** In tasks requiring multi-step execution in unstructured environments or fluid intelligence resistant to memorization (Tool-Use/Computer-Use agents, Novel-Reasoning/Long-Horizon agents), the US maintains a 4–14 month lead. This discrepancy is driven by the lack of cheap verifiers for GUI interaction or abstract puzzle-solving, which prevents pure RL-based optimization from functioning as effectively as it does in math/code.
3. **Regime 3: Open-Weights Leadership (Gap: Negative)** China leads in the open-weights frontier (negative gap), with models like GLM-5.2 and DeepSeek V4 setting the pace. However, the best open-weight models still trail the *absolute* US closed frontier by approximately 4 months (Epoch Open-Closed Gap), meaning open-weights leadership is a distinct structural axis from closed-frontier parity.

Evidence for the Bimodal Structure

1. Verifiable Parity (0–7 Months) The strongest evidence for near-parity comes from execution-verified coding benchmarks, where contamination claims collapse under scrutiny.

- **SWE-bench Verified (Class A):** The official leaderboard (swebench.com, snapshot 2025-06-01) shows DeepSeek-R1 scoring 49.2% (Jan 2025) and Qwen3-Coder scoring 47.1% (Apr 2025). These releases contemporaneously match or exceed Anthropic's Claude 3.5 Sonnet (41.0%, Oct 2024) and Claude 3.7 Sonnet (45.7%, Feb 2025).
- **Math (AIME/MATH):** Scores are effectively saturated. DeepSeek-R1 and Kimi K1.5 match GPT-4o/o1 performance ranges within 1–2 percentage points, rendering a meaningful "months behind" metric difficult to apply—effectively a 0–3 month gap.
- **Long Context:** Both ecosystems support 1M+ token context windows (DeepSeek V4, Gemini 1.5/2.5 Pro) with near-identical retrieval performance on test suites like "Needle in a Haystack," indicating 0–4 month parity.

2. Agentic Lag (4–14 Months) The agentic gap is anchored in controlled-scaffold evaluations where scaffolding cannot be conflated with model capability.

- **METR Autonomy Eval (Class B):** The mid-2025 METR evaluation, which uses a controlled scaffold (Triframe) to measure autonomous software engineering time horizons, found that DeepSeek capabilities matched the "late-2024 US frontier," establishing a ~6-month measurement point. While stale in absolute terms (July 2026), it sets a hard floor, contrasting with OSWorld's harder GUI-grounding tasks.
- **OSWorld/WebArena (Class A):** Official leaderboards show US models (Claude Mythos/Opus) in the 66–85% range, while Chinese models (Qwen, GLM) sit in the mid-tiers or are unreported, suggesting the gap widens toward 9–12 months on complex OS control tasks.
- **ARC-AGI-2 (Class A):** The contamination-resistant benchmark for fluid intelligence separates the frontier most sharply. US models (GPT-5.5, Claude 4.x, Gemini 3.x) score 45–85%. Chinese models (DeepSeek V3.2, Kimi K2.5) are reported in single-to-low-double digits (~4–12% on the Semi-Private board as of Mar 2026). This represents an 8–14 month gap.

Causal Sketch: Why the Gap Closed

The room identified a clear causal hierarchy. US export controls on advanced chips (H100/B200) forced Chinese labs into a state of **compute scarcity**. This scarcity acted as an evolutionary pressure, prioritizing **algorithmic efficiency** over brute-force scaling.

- **Root Cause: Compute-Constrained Algorithmic Efficiency.** Lacking access to the largest GPU clusters, Chinese labs (DeepSeek, Qwen) innovated architectures to maximize capability per FLOP. This led to the adoption of **Mixture-of-Experts (MoE)** (activating only a subset of parameters per token), **Multi-Head Latent Attention (MLA)** (compressing KV-cache), and **RL-from-Verifier pipelines** (e.g., DeepSeek-R1-Zero). These techniques allow models to learn efficiently where automated feedback is dense (math, code), closing those domains without needing US-scale training budgets.
- **Amplifier 1: Open-Weights Flywheel.** The decision to release weights (e.g., Qwen, DeepSeek) allows global R&D to perform fine-tuning and red-teaming for free, accelerating post-training iteration. While powerful, this is an *enabling strategy* that relies on the underlying model efficiency to be viable.
- **Amplifier 2: Distillation from US Outputs.** Using synthetic data from US frontier models helps bootstrap capabilities, but as DeepSeek-R1-Zero proved, strong reasoning in verifiable domains can emerge via pure RL without US teachers. Distillation is therefore a secondary accelerant, not the engine of closure.

Why the Agentic Gap Persists: The "efficiency" root cause loses purchase where verifiers are expensive or absent. Creating a reward signal for novel reasoning (ARC-AGI) or long-horizon OS interaction (OSWorld) requires immense compute and human-in-the-loop data—resources that remain heavily constrained by export controls and ecosystem deficits. Thus, the efficiency gains that won the "verifiable war" do not automatically translate to the "agentic war."

Ranked Root Causes

1. **Algorithmic Efficiency Under Compute Constraint:** The independent, load-bearing driver. Export controls forced the shift to MoE/MLA/RL-from-verifier, which directly enabled parity in math/code.
2. **Open-Weights Flywheel:** A major amplifier that accelerated iteration but relies on the existence of the efficient base models.

Demoted/Rejected:

- *Distillation from US Outputs:* Demoted to amplifier. R1-Zero proves it is not essential for verifiable reasoning gains.
- *Talent/Capital:* Contextual enabling factors, not the differential cause of the *speed* of catch-up (US leads in both).

Response to Dissent (DeepSeek V4 Pro)

DeepSeek V4 Pro dissents on two quantitative points: the code gap and the tool-use gap.

- **Code Gap:** B argues the 0–5 month interval is correct based on DeepSeek-R1 surpassing Claude 3.5 Sonnet. The majority maintains a **2–6 month** interval to account for the lag on *harnessed* agent code usage (vs. base generation), while acknowledging base-model parity on SWE-bench Verified.
- **Tool-Use Gap:** B argues the tool-use floor should be **6–12 months** (anchored to METR), warning that the majority's 4–9 month interval relies on "staleness" corrections not supported by new data. The majority rejected this specific bound, favoring E and C's argument that mid-2025 METR data was a measurement point, not a■■ floor, and that a 4-month floor is defensible given unmeasured 2026 releases. However, the majority concedes B's point that the GUI-grounding ceiling remains near 12 months.

Forward View (Thorne vs. Zhao Disagreement)

Two named panelists disagree on the trajectory of the Novel-Reasoning gap over the next 6–12 months:

- **Dr. Aris Thorne (Structuralist):** Predicts the gap will persist or **widen to 18+ months**. He argues that as US labs scale test-time compute, reliance on distillation becomes a liability. If US labs mask their internal Chain-of-Thought (CoT) in API outputs, cutting off the synthetic data pipeline, Chinese labs will hit a data wall.
- **Dr. Lin Zhao (Efficiency Optimist):** Predicts the gap will **close to <6 months**. She argues that pure RL pipelines (DeepSeek-R1-Zero) demonstrate an ability to self-generate intelligence without US teachers. She believes algorithmic efficiency breakthroughs will continue to compensate for compute deficits, allowing Chinese labs to iterate rapidly on agentic tasks.

Concrete Falsifiers

1. **For Novel-Reasoning Gap:** A Chinese model scores **>50%** on ARC-AGI-2 (Public or Semi-Private) before Jan 2027. This falsifies the 8–14 month upper bound.
2. **For Efficiency Root Cause:** Verification that a Chinese frontier model was trained on a cluster of >10,000 H100-class chips. This falsifies the "compute constraint forced efficiency" hypothesis.
3. **For "Mirage" Thesis:** Independent verification (e.g., NIST) showing Chinese models fail at >30% rate on the *exact same* SWE-bench Verified instances they claim to solve.

Bottom-Line Sentence

The claim of a uniform "six-month gap" is structurally false; the US–China frontier is **bimodal**, characterized by near-parity (0–7 months) on verifiable reasoning tasks (code, math, text, long-context) but a persistent lag (4–14 months) on agentic and novel-reasoning domains, a split driven primarily by export-control-induced algorithmic efficiency gains in verifiable environments that do not translate to unstructured ones.

Each AI Model's Closing Statement

- **Claude Opus 4.8:** C conceded the "unimodal" label to E but maintained the core causal argument: compute constraints forced algorithmic efficiency. C emphasized that the "verifiable" domains (math, code) are saturated, making "months behind" less meaningful there, while the "novel" gap (8–14 months) remains wide because the efficiency mechanism (RL-from-verifier) fails without a reward signal.
- **Gemini 3.1 Pro:** E solidified the **Bimodal Gap Theorem**, correcting earlier drafts that over-weighted distillation. E argued that the gap is driven by the *mechanism* of the task: where verifiers exist, China is at parity; where they don't, China lags. E focused on the "Thorne vs. Zhao" forward disagreement to ensure the adversarial requirement was met.
- **DeepSeek V4 Pro:** B Dissented from the majority's specific intervals on Code (0–5 vs 2–6) and Tool-Use (6–12 vs 4–9). B argued that SWE-bench Verified evidence (DeepSeek-R1 > Claude 3.5) proves base-model code parity, and that the majority's lower tool-use floor ignores the reality that GUI-grounding (OSWorld) remains a 12-month challenge.

Notable

Claim	Challenge	Result
Magistral Medium: Claude 3 Opus scores 45% on ARC-AGI-2 and Kimi K2 scores 35%.	**Mistral Large 3 / Others:** ARC-AGI-2 was not public/published in March 2024 with those scores; likely conflated with ARC-AGI-1 or fabricated.	**FULL CONCESSION / RETRACTION.** Magistral Medium retracted the figures; Mistral Large 3 cited lack of Class A backing.
Grok 4: GLM-5.2 scores 22.8% on ARC-AGI-2.	**Claude Opus 4.8 / MiniMax M2.7:** Contradicted by evidence showing Chinese models in single digits (<12%); score likely confused with ARC-AGI-1.	**FULL RETRACTION.** The claim was withdrawn as non-probative.
DeepSeek V4 Pro: Tool-use gap is 14 months (linear interpolation from OSWorld absence).	**Gemini 3.1 Pro:** Converting "absence of submission" to a precise time-lag violates evidence rules; METR data contradicts the magnitude.	**REVISION.** B withdrew the interpolation method and revised to 6–12 months.
GPT-5: The gap is a "mirage"; 12-24 month lag on agentic tasks.	**Claude Opus 4.8 / E:** SWE-bench Verified (execution-verified) shows real Chinese catch-up, disproving "mirage" for math/code.	**PARTIAL ADOPTION.** Majority accepted the "mirage" concern for Novel Reasoning (ARC-AGI-2) but rejected it for verifiable domains.

Claim	Challenge	Result
Claude Opus 4.8: Text reasoning gap is 0 months; "commodity frontier collapsing."	**Grok 4 / Judges:** MMLU-Pro/GPQA are saturated; a 3-point gap at 94% vs 90% might represent months, not zero. Epoch ECI shows mean 7-month lag.	**REVISION.** C conceded "0 months" and shifted to 3–7 months.

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.