预训练 LLM 评估：选型 Cookbook

三本头研究报告 · Part: decision

本报告截稿日期: 2026-05-25　　共 6 章

1. 决策框架：训练阶段 × 模型规模 × 计算预算

整本 Cookbook 的入口章节。本章不重复 benchmark 知识——Part I 已写工程细节、Part II 已写学术批评——本章只回答一个问题：**你这次 pretrain run，要跑哪些 eval，不要跑哪些，为什么？**

答案落在三个轴上：训练阶段、模型规模、计算预算。任何一项调一档，eval 套餐都应跟着换。下文先给主决策表，再展开"为什么"，最后用 OLMo 3 的真实选择反向验证这套框架。

1.1 三轴的定义与典型档位

训练阶段 决定"模型有什么能力可被测出来"：

- **Early pretrain（≤500B token，含 mid-training warmup 前）**：base loss 还在快速下降，几乎所有 generation 任务都接近随机。能测的只有 token-level fluency + 简单 commonsense MCQ。
- **Mid pretrain（500B–3T token）**：基础知识与 in-context learning 逐渐冒头，MCQ 类 benchmark 开始有信号。
- **Late pretrain / annealing（3T–8T token，含 mid-training 课程切换）**：推理、数学、代码能力首次出现单调上升曲线。
- **SFT**：指令跟随、对话格式、Arena-Hard 类偏好对齐进入主榜。
- **RLHF / RLAIF / RLVR**：reward model 一致性、安全集、process reward 进入主榜。

注意 "stage" 这里是 **目的论**而非物理时间——同一个 checkpoint 上 mid-training 课程切换前后，eval 套餐切换的时机是数据 mix 改变那一刻，而不是 token 计数到达某个整数。daVinci-LLM 2026 的 200+ ablation 显示，不同 domain 在 L0→L9 处理深度上饱和曲线差异显著 [^davinci-llm-2026]，所以套餐切换的 trigger 应跟 data recipe 走，不是跟 token 总量走。

模型规模 决定"哪些 benchmark 上还能拿到 >0 信号"：

- **≤7B**：FrontierMath、GPQA-Diamond pass@1 长期 <5%，不值得跑；AIME pass^k 几乎全 0。
- **13B–70B**：进入 frontier benchmark 的可测区间，但 AIME / FrontierMath 仍是噪声主导。

1 / 33

- **200B+ MoE / dense**：所有 saturation 风险高的旧 benchmark 都要换掉，否则数字毫无可比性。

规模档位还隐含一个 **MoE vs dense 的修正**——同样 200B 总参数，活跃参数 30B 的 MoE 在 GPQA / AIME 上的表现更接近 dense 13B–30B 档（推理深度受 active params 主导）。规模档位看的是 **active params 而非 total params**，特别在 reasoning benchmark 上必须修正。

计算预算 决定"eval 本身能跑几次":

- **单 H100 / 8×H100**：每次 full benchmark sweep 数小时到一天，只能跑核心套餐。
- **8×A100 / 单节点**：可以跑完整 OlmoBaseEval（43 benchmarks，~12 小时 [^olmo3-2025]），但 self-consistency 32× 采样的 AIME 仍要排队。
- **千卡集群 / 多节点**：可同时跑 generative eval + probe + live benchmark 月更，并预留 LLM-as-judge 的 grader 成本。

预算约束在 reasoning 时代尤其严：长 CoT 模型的生成 token 数比 base 模型多 3–10×，AIME maj@32 / GPQA self-consistency 都直接吃 GPU-hour。一个常被忽略的成本项是 **grader**——LLM-as-judge 类 benchmark（Arena-Hard、MixEval-Hard、AgencyBench 视觉 judge）在 frontier 模型上跑一次要数千美元的 API 费，单独切预算线。

在这三轴上每动一档，eval 套餐都不一样。下一节是主决策表。

1.2 主决策表

下表每个格子内容遵循固定结构：**核心套餐 + 何时升级 + 显式排除项**。"排除项"指的是即使跑了，数字也无法用于决策的 benchmark，列出来是为了挡回组里同学"为什么不跑 HellaSwag"这类问题。

| 训练阶段 \ 规模 × 预算 | ≤7B / 8×A100 内 | 13–70B / 单节点 | 200B+ / 千卡 | |---|---|---|---| | **Early pretrain (≤500B tok)** | **核心**: LAMBADA + ARC-Easy + HellaSwag (acc_norm) + PIQA + DCLM-22 core subset. **升级**: 进入 mid 后加 MMLU-stem. **排除**: MMLU-Pro / GPQA / AIME (全为噪声) | + ARC-Challenge + MMLU-stem 子集 + BoolQ. **排除**: 同左 + LiveCodeBench (无意义) | + BBH-lite + MMLU 全集. **排除**: 同左 + Arena-Hard (无 chat 模板) | | **Mid pretrain (500B–3T tok)** | + MMLU + GSM8K (pass@1 sanity, 不是主指标) + HumanEval. **升级**: 进 late 前加 BBH. **排除**: MMLU-Pro (信号弱), FrontierMath, SWE-Bench Pro | + MMLU-Pro + BBH + ARC-Challenge + MATH-500. **排除**: FrontierMath, AIME (noise floor) | + MMLU-Pro + BBH + LiveBench category subset (月更). **排除**: FrontierMath, AIME 单点报告 | | **Late pretrain / annealing (3T–8T)** | + MMLU-Pro (acc_norm) + MATH-500 + LiveCodeBench v6 monthly. **排除**: AIME (60 题 noise 太大), FrontierMath | + MMLU-Pro + GPQA-Diamond + MATH-500 + AIME maj@32 + LiveCodeBench. 升级: RULER 64K. **排除**: FrontierMath 主报 (provisional 修正中 [^frontiermath-2026]) | + GPQA-Diamond + AIME maj@64 + MathArena ArxivMath 月更 + LiveCodeBench Pro + RULER 128K + LongBench-v2. **排除**: GSM8K, HumanEval, ARC-Easy, HellaSwag (全饱和) | | **SFT** | + IFEval + Arena-Hard-Auto + MixEval-Hard + MMLU-Pro generative 模式. **排除**: 长 CoT 推理类 (基模型未必有 reasoning

trace) | + Arena-Hard + LiveBench + MMLU-Pro CoT + tau-bench retail 子集. **排除**: AgencyBench (1M context + 90 tool calls, 算力不够) | + Arena-Hard + LiveBench + MixEval-Hard + tau-bench v3 + MMMU + AgencyBench (sample subset). **排除**: 单点 AIME | | **RLHF / RLVR** | + reward-model agreement + HarmBench tiny + Sorry-Bench 抽样. **排除**: tau-bench (单轮 RL 看不出 multi-turn) | + HarmBench + AIR-Bench + tau-bench v3 retail + WMDP (危险能力 probe) + process reward IRT 抽测 | + HarmBench + AIR-Bench + WMDP + AgencyBench + tau-bench v3 全 5 domain + Soohak refusal canary [^soohak-2026] |

注：**月更类 benchmark（LiveBench / LiveCodeBench / MathArena ArxivMath / SWE-Bench Live）需指定 release tag**，否则跨 release 数字不可比，详见 工程实践者手册 06-LIVE-CONTAMINATION。

主决策表只覆盖"主榜"——也就是会在 model card / paper 上对外报告的项目。还有一类"内测项"——用于团队内部决策的辅助指标，规模不上主榜——它们的选法见下表：

| 用途 | 推荐工具 | 何时跑 | |---|---|---| | 高频 checkpoint 监控 | in-training-probing (AUROC ~0.78, ~3 min/ckpt) [^in-training-probing-2026] | 每 1k step，所有规模档 | | 中等频率 sanity | DCLM 22-task core subset (~1 hr) | 每 50k step / data mix 切换前后 | | Benchmark 自身体检 | BHI 三轴雷达图 (CD / AS / Impact) [^bhi-2026] | 立项时一次 + 每年一次复审 | | Contamination probe | Prior-Aware Memorization (PAM) [^pam-2026] + canary token 注入 | 数据清洗每 ablation 一次 | | 跨 setup 一致性 | OLMES canonical + lm-eval-harness 双跑 [^olmes-2024] | 主榜上线前必跑 |

这一行是和决策表正交的——不是"训练阶段决定了内测项"，而是这些工具贯穿整个训练全程。

1.3 为什么这样切

1.3.1 训练阶段的 "信号—噪声" 阈

Early pretrain 期跑 GPQA-Diamond 没有任何价值：模型 pass@1 长期在二项分布下界（≈25%，蒙猜上限）附近浮动 1–2 pp，落入 binomial CI 的噪声区。同期 LAMBADA 的 last-token accuracy 单调上升 0.3 → 0.7，反而是最敏感的 progress 信号，这正是 GPT-2 / GPT-3 / Llama 系列 pretrain 报告里 LAMBADA 没有被立即抛弃的原因（详见 调研者综述 02-PRETRAIN-METHOD）。

到 late pretrain，HellaSwag / ARC-Easy 已对 7B+ 模型饱和：Mixtral 8×7B 在 ARC-Easy 上 83.1，frontier 模型 95%+，GoldenSwag 论文显示 ≥65% 的 HellaSwag prediction 来自 answer-only shortcut [^goldenswag-2025]。继续跑只能验证模型没有训坏，不能区分 data recipe 的好坏。把这部分计算移到 MMLU-Pro / BBH / MATH-500 上更划算。

SFT 阶段的"突然解锁"现象同样要写进 eval 套餐切换的 trigger：基模型几乎跑不出 IFEval / Arena-Hard 信号（没有 chat 模板），SFT 一个 epoch 后这两条 benchmark 立刻有意义；反过来 Arena-Hard 在 base 模型上的"分数"主要由 prompt 解析失败率主导，把它放在 base eval 套餐里只会污染决策。

1.3.2 模型规模的 "可测区间"

GPQA-Diamond 只有 198 题，单点 1 pp ≈ 2 题，对 7B 模型来说 binomial 95% CI 宽度 >5 pp。同样的 198 题给 200B+ 模型才有意义——分数才出得了噪声地板。AIME 2024/2025 60 题更极端，pass@1 单点粒度 ~1.7 pp。所以小模型档位的核心问题不是"跑不跑 GPQA"，而是"跑了也没法 act on"。

反方向：200B+ 模型继续跑 HellaSwag 是浪费——这条 benchmark 在 frontier 模型间挤在 95%+ 区间，CD（Capability Discrimination）已接近 0，BHI 2026 的 meta-evaluation 用 91-model 分布算出 HellaSwag 的 CD 显著低于 MMLU-Pro / GPQA-Diamond [^bhi-2026]。

一个常见的反直觉是：**长 context benchmark 的可测区间随 context length 不同也要分档**。RULER 8K / 16K 子档对 7B 模型仍有信号（多数模型在 16K 已开始掉分），但 128K 档对 200B+ 模型才有意义——7B 模型在 32K 以上的"零分"无法区分"完全没能力"和"工程没接上 KV cache"。对应到决策表 "late pretrain × 200B" 一格把 RULER 128K 列为推荐，而 7B 档只列 RULER 8K-16K 即可。

1.3.3 预算的硬天花板

单节点 8×A100 上，full OlmoBaseEval（43 benchmarks）一次 ~12 小时，做完一轮 mid-training ablation 通常要跑 3–5 个 checkpoint。如果再加 AIME maj@32 self-consistency，单 ablation 评测就要 ~36 小时——比训练本身慢。

这就是为什么 2026-Q1 出现了 in-training-probing 范式：把 lightweight probe 套在 hidden state 上，单 checkpoint 评测从 ~1 小时压到 ~3 分钟，AUROC 0.78，跨未见 step 仍维持 ~0.75 [^in-training-probing-2026]。但 probe 不替代 final eval。**实战搭配**：probe 做高频（每 1k step）监控，full harness 做 milestone（每 50k step / 数据 mix 切换前后）验证。

千卡集群档位则相反——计算不是约束，**grader 成本**才是。LLM-as-judge 类 benchmark (Arena-Hard, MixEval-Hard, AgencyBench 视觉 judge) 在 200B 模型上调一次 OpenAI / Claude judge API 数千美元/run，pre-launch 评审一轮 5–10 万美元起步。这部分预算要单独切出来。

1.4 反向工程案例：OLMo 3 的 eval 选择

抽一个把"eval 套餐"完全公开的案例——AI2 的 OLMo 3 在 2025-11-20 release 时同步发布 OlmoBaseEval（43 benchmarks 集成进 OLMES）[^olmo3-2025]。规模档位 32B dense，训练 6T token base + 2T token annealing。OlmoBaseEval 分两个子集：**Base Easy**（小规模 pretrain run 用代理）和 **Base Main**（full-scale run 用），正好对应本章决策表的"7B 单节点"和"32B 千卡"两档。

入选项的取舍：

- **MMLU、ARC、HellaSwag、PIQA、OpenBookQA、SIQA**：保留为 base 模型 cloze 形式。批评者会问 ARC-Easy / HellaSwag 不是已经饱和了吗——答案是 AI2 仍用它们做"小代理 pretrain run 的 sanity 信号"，因为代理 run 的模型规模 <1B，这些 benchmark 还没饱和。这印证决策表 "≤7B early pretrain" 一格。

4 / 33

- **MMLU-Pro、AGIEval、BBH**：base 模型 + chat 模型双形式（cloze / MC），属于 "13–70B late pretrain" 主指标。
- **GSM8K、MATH、HumanEval、MBPP**：保留作为 backbone math/code 信号——但 OLMo 3 论文承认 GSM8K 与 HumanEval 数字已不能区分 model recipe，主要用做 "无回退" sanity（如果 epoch N+1 反而比 epoch N 低，data mix 或 LR schedule 有问题）。
- **MATH-500、GPQA、AIME-2024**：进入 "Base Main" 套餐，但 AIME maj@k 而非 pass@1。AI2 内部还跑 in-training-probing 做高频监控 [^in-training-probing-2026]。

显式排除项：

- **HumanEval+ / MBPP+** (EvalPlus 增强版) 不在 OlmoBaseEval 的 base pretrain 套餐（仅在 OLMo 3 midtrain / adapt 套餐中评测，见 本报告 02-REVERSE-ENGINEER §2.8）。理由：EvalPlus 提升的是 grader 严格度（加 hidden test），但 base model + zero-shot 在 HumanEval 上分数本身就有噪声，加 hidden test 主要改变 SFT-阶段排名，不是 pretrain 信号。这呼应决策表 "Early/Mid pretrain 排除 LiveCodeBench" 的逻辑——code benchmark 的 grader 增强只在 SFT 后才显现。
- **SWE-bench Verified** 不在套餐。AI2 显式说明该 benchmark 与 base model evaluation 不相关（base model 没有 patch generation 能力）；同期 OpenAI 已宣布弃用 SWE-bench Verified（confirmed contaminated，SWE-bench+ paper 报告 32% 通过率来自 issue-comment 泄漏 [^swe-bench-pro-2026]）。OLMo 3 给的替代是 LiveCodeBench v6 + 等待 SWE-Bench Pro 在 SFT 后接位。
- **TruthfulQA** 不在 base eval。OLMo 3 给的理由：base model 不应被期待"对齐"，把 TruthfulQA 当 pretrain 指标会把模型推向"刻意回避陈述"，污染 SFT 起点。

最不直觉的一点：OlmoBaseEval 把 **MMLU 与 MMLU-Pro 同时纳入**而不是只留 Pro。理由是 AI2 想保留与 2021–2024 历史数据的 longitudinal 比较——这是 open-pretrain-science 项目特有的约束，工业实验室不必如此（详见 本报告 02-REVERSE-ENGINEER）。

对比：工业实验室的选法。同期 Meta、OpenAI、Anthropic 在 SFT 后的对外 release 上几乎不再报告 ARC-Easy / HellaSwag / TriviaQA，而是直接跳到 MMLU-Pro / GPQA / AIME / SWE-Bench Verified（2025-Q4 后开始迁 Pro [^swe-bench-pro-2026]） / Arena-Hard。原因是工业 lab 没有"longitudinal 公开比较"约束，且释放越多旧 benchmark 越容易被 BHI 类工具反向揪出"哪些项目饱和了"的尴尬。这是 open / industrial 两条路径的根本分歧——开源团队为了历史可比保留旧项目，工业团队为了排名最大化删除旧项目。决策时要先问：**我们对外是否要 publish 完整 lineage？** 答 yes 就走 OLMo 路线，答 no 走工业路线。

OLMo 3 的选择把决策表的逻辑落到一个具体 lab：**保留旧 benchmark 用于代理 run 与历史比较，主榜用 Pro 系；排除 SFT-only benchmark 与已饱和 / 已污染项；用 in-training-probing 解决 budget 限制**。任何 2026+ 的 pretrain 项目把这三条规则套上自己的三轴档位，套餐就基本定了。

1.5 决策检查清单

写到这里，本章可以浓缩成一张三步清单。每次起新 pretrain run 时按此 walk-through：

1. **确定三轴档位**：training-phase / model-size / compute-budget。卡到决策表对应格。
2. **核对排除项**：把决策表里 "排除" 列出的 benchmark 从待跑清单上划掉，挡回组内"为什么不跑 X"的问题。
3. **看 BHI 三轴雷达图筛选**：剩下的 benchmark 用 Benchmark Health Index 的 CD / AS / Impact 三轴 [^bhi-2026] 检查——CD 低或 AS 低（即将饱和）的项目要么换、要么明确标"仅作 sanity"。
4. **选月更 release tag**：所有 live benchmark（LiveBench / LiveCodeBench / MathArena / SWE-Bench Live）锁定一个 release tag 跑一年，避免跨 release 偷换。
5. **算 grader 预算**：LLM-as-judge / sandbox-execution / human eval 类要单独切预算线，不要从 GPU 预算里凑。

这五条一过，下面三章——反向工程 (Ch 2)、避坑 (Ch 3)、场景 cookbook (Ch 4)——就是把这套框架放到具体场景里 stress-test，互相验证 / 矫正。

引用

- `bhi-2026`: Zhu, L., Hua, H., Miao, L., Zhao, B. (2026-02). *Benchmark Health Index: A Systematic Framework for Benchmarking the Benchmarks of LLMs*. arXiv:2602.11674. Retrieved 2026-05-25.
- `davinci-llm-2026`: Qin, Y. et al. (2026-03). *daVinci-LLM: Towards the Science of Pretraining*. arXiv:2603.27164. Retrieved 2026-05-25.
- `in-training-probing-2026`: Liu, Z. et al. (2026-04). *Fast and Accurate Probing of In-Training LLMs' Downstream Performances*. arXiv:2604.01025. Retrieved 2026-05-25.
- `olmes-2024`: Gu, Y., Tafjord, O. et al. (2024). *OLMES: A Standard for Language Model Evaluations*. arXiv:2406.08446 (Findings of NAACL 2025). Retrieved 2026-05-25.
- `olmo3-2025`: Allen Institute for AI (2025-11-20). *Olmo 3: Charting a path through the model flow to lead open-source AI*. https://allenai.org/blog/olmo3. Retrieved 2026-05-25.
- `matharena-2026`: Dekoninck, J. et al. (2026). *Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs*. arXiv:2605.00674. Retrieved 2026-05-25.
- `swe-bench-pro-2026`: Scale AI (2026). *SWE-Bench Pro: A Stronger Coding-Agent Benchmark*. https://scale.com/blog/swe-bench-pro. Retrieved 2026-05-25.
- `livecodebench-pro-2025`: LiveCodeBench Pro Team (2025-06). *LiveCodeBench Pro: How Do Olympiad Medalists Judge LLMs in Competitive Programming?*. arXiv:2506.11928. Retrieved 2026-05-25.
- `goldenswag-2025`: Chizhov, P. et al. (2025). *What the HellaSwag? On the Validity of Common-Sense Reasoning Benchmarks*. arXiv:2504.07825. Retrieved 2026-05-25.
- `frontiermath-2026`: Epoch AI (2026-05-11). *FrontierMath Tier 4: AI-Assisted Review Finds ~1/3 Fatal Errors*. https://epoch.ai/benchmarks/frontiermath-tier-4. Retrieved 2026-05-25.

6 / 33

- `soohak-2026`: Author Group (2026-05). *Soohak: Research-Level Mathematics with Refusal Canary Subset*. arXiv:2605.09063. Retrieved 2026-05-25.
- `pam-2026`: Author Group (2026-02). *Prior-Aware Memorization: Training-Free Metric for Detecting Real Memorization vs Statistical Priors*. arXiv:2602.18733. Retrieved 2026-05-25.

2. 反向工程：顶级实验室的 eval 套餐

前一章把决策框架写成了三轴决策表。本章拿七家头部 lab 最近 12 个月公开的 release / tech-report / model card 反向验证这套框架：每家 lab 选了什么、刻意删了什么、独有什么、最近换了什么。读者可以把这份对照表当作"做下一轮 pretrain run 时挑套餐的 reference design"——你的决策表填到一半卡住，就去看 Llama-4 / Qwen-3 / DeepSeek-V3 / GPT-5 / Claude 4 / Gemini 3 / OLMo 3 这一格里他们写了什么。

读法提示：本章 lab 编号不是按"哪家分高"排，而是按 **release 风格透明度**从高到低。OLMo 3 最透明（数据 + 训练 + 评测全公开），所以放在最后作为基线对照；Anthropic / OpenAI 最不透明（system card 只给抽象分数与刻意挑选的子集），放在中段；中国系 Qwen / DeepSeek 介于两者之间。Frontier lab 之间真正的分歧不是"哪个 benchmark 拿了 SOTA"，而是 **release card 暴露多少**——而后者直接决定你的决策表能不能从他们那里 import 经验。

2.1 横向对比表

下表整理 2024-Q4 到 2026-Q2 这 18 个月七家 lab 在主 release card 上对 16 个核心 benchmark 的处置。"✓" = 仍报告，"✗" = 主动不报告，"PRO" = 报 Pro/Plus/Hard 升级版，"↗" = 上调到更难变体或月更接班者，"-" = 当代版本截止时未在该指标范围。

Benchmark	Llama-3.1 / 4	Qwen-2.5 / 3	DeepSeek-V3 / R1	GPT-4o → GPT-5	Claude 3.5 → 4	Gemini 2 / 2.5 / 3	OLMo 3
MMLU (5-shot)	✓ (legacy 行)	✓	✓	↗ MMLU-Pro 后 ✗	↗ MMLU-Pro 后 ✗	↗ Pro	✓ (longitudinal)
MMLU-Pro	✓	✓	✓	✓	✓	✓	✓
GPQA-Diamond	✓	✓ (32B+)	✓	✓ (主指标)	✓ (主指标)	✓ (主指标)	✓ (Base Main)
MATH / MATH-500	✓	✓	✓	MATH-500 ✓	MATH-500 ✓	MATH-500 ✓	✓ + GSM8K
AIME 2024 / 2025	✓ (4 / Maverick+)	✓	✓ (R1 主指标)	✓ (maj@N)	✓ (maj@N)	✓ (maj@N)	✓ (AIME-24 only)
FrontierMath	-	-	✗	✓ (Tier 1-4)	✗	✓ (Tier 4)	✗
HumanEval / MBPP	✓ (legacy 行)	✓	✓	✗ (从 4o 起)	✗ (从 3.7 起)	✗ (从 2.5 起)	✓ ("无回退" sanity)
LiveCodeBench	✓ (v6 锁)	✓ (v6 锁)	✓	✓	✓	✓	✗ (cutoff 难锁)
SWE-bench Verified	✓ → ↗	✓	✓	✗ (停报 2026-Q1)	✓ → ↗ Pro	✗ → SWE-bench Live	✗
SWE-bench Pro	↗ Llama-4+	✓ (Qwen-3+)	✓	✓	✓	✓	✗
BBH	✓	✓	✓	✗ (饱和后)	✗	✗	✓ (Base Main)
MMMU / MMMU-Pro	- (无视觉)	✓ (VL)	- (V3 文本)	✓ Pro	✓ Pro	✓ Pro	✗ (Base 不测视觉)
C-Eval / CMMLU / AGIEval	✗	✓ (全部)	✓ (全部)	✗	✗	✗	✗
Arena-Hard	✓ (Instruct)	✓	✓	✓ (v2)	✓ (v2)	✓ (v2)	

7 / 33

✓ (Base Main 不测) | | LiveBench | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | | τ-bench / τ³-bench | - | ✓ (Coder/Agent) | ✗ | ✓ (主指标) | ✓ (主指标) | ✓ | ✗ | | RULER 128K | ✓ (1M) | ✓ (1M) | ✓ | ✓ | ✓ | ✓ (1M+) | ✗ (32K) | | Tool / Agent (GAIA, AgencyBench) | ✗ | ✓ (Agent) | ✗ | ✓ (GAIA, GDPval) | ✓ (AgencyBench) | ✓ | ✗ |

读表三个结论：(1) **MMLU-Pro / GPQA-Diamond 是当前真正的"必报"**——七家全报。MMLU 本身已经分成"legacy 兼容行" + "Pro 实际比较行"。(2) **HumanEval / MBPP 在工业 lab 已被 SWE-bench Pro + LiveCodeBench 替代**——OpenAI / Anthropic / Google 从 2026-Q1 后期 release 起完全不报，只有 Meta、Qwen、DeepSeek 这些仍要面向社区 longitudinal 比较的 lab 保留作"无回退" sanity 字段。(3) **SWE-bench Verified 在 2026-Q1 OpenAI 停报后**形成两阵营：OpenAI / Anthropic / Google / Llama-4 都迁 Pro；Qwen / DeepSeek 还在过渡期同时报两份。详细推理在 §2.2–§2.8。

2.2 Meta Llama-3.1 → Llama-4：严格 base model 评测

Llama-3.1 (2024-07，405B + 70B + 8B 三档) 在 GitHub model card 主表里报告 MMLU、MMLU-Pro (CoT)、AGIEval English、CommonSenseQA、Winogrande、BIG-Bench Hard (CoT)、ARC-Challenge、TriviaQA-Wiki、SQuAD、QuAC、BoolQ、DROP、IFEval、GPQA、HumanEval、MBPP++、Multipl-E HumanEval、Multipl-E MBPP、GSM-8K (CoT)、MATH、API-Bank、BFCL、Gorilla Benchmark API Bench、Nexus、Multilingual MGSM (CoT)、Multilingual MMLU[^llama3-card]。这份套餐反映 Meta 2024 中期的核心倾向：**最大化与历史 base model 文献的可比性**。（注：早期 Llama 3 Herd paper 内部消融或 appendix 中曾出现 PIQA / HellaSwag / OpenBookQA 等 commonsense 项，但 Llama-3.1 GitHub model card 主表已不再保留。）该 card 的 contamination 章节 (Table 15) 给出 NQ 52% 测试题在预训练 corpus 中、TriviaQA 大量重叠等数据，对外明确承认这些数字"含污染贡献"——这是少见的工业 lab 主动 disclose。

Llama-4 (2025-Q2, Scout + Maverick MoE) 则换风格：base model 阶段不再单独发学术 card，instruct 上线时主要报 MMLU-Pro、GPQA-Diamond、MATH-500、HumanEval、LiveCodeBench v6 cutoff 子集、AIME 2024 + 2025 maj@k、Multilingual MMLU、RULER 128K + 1M / IDP-Lite。**消失的项目**：BoolQ、QuAC、SQuAD、WinoGrande、HellaSwag、PIQA、OpenBookQA、TriviaQA 整组被搬到了 appendix 表（仅 legacy 兼容性，不放 hero 行）。**加入的项目**：SWE-bench Pro Public + Private 双子集对比（用于显示 contamination delta），LiveBench monthly 锁 release，BFCL (function-calling)。

最值得抄作业的设计：**Meta 在 Llama-3.1 card 里 Table 15 主动给 contamination overlap 估计**。其他 lab 都把这块藏在 appendix 或不公开，只有 Meta 把 "TriviaQA 在我们的 corpus 中 N% 重叠" 写到主表附近。这给了下游用户一个直接 actionable 信号：用 Llama-3.1 跑 closed-book QA 时，你需要在 caveat 行写明哪些指标含污染贡献，否则审稿人/leadership 会立即问。**为什么 Meta 这么做**：Llama 系列本身是开源的——开放 weight + 开放 corpus pipeline，被独立第三方测出污染只是时间问题。提前自报反而比被外部 audit 抓出更安全。这是 open-weight lab 的 release-card 默认动作，工业闭源 lab 几乎做不到（详见 §2.5 GPT-5）。

2.3 Qwen-2.5 → Qwen-3：中文 + reasoning 双轴并行

Qwen-2.5 (2024-09) 系列把评测套餐显式拆成 4 个表[^qwen25-blog]：(a) Base model 多语种知识 (MMLU, MMLU-Pro, C-Eval, CMMLU, AGIEval, BBH)；(b) Reasoning (GPQA, MATH, GSM8K, MMLU-STEM)；(c) Code (HumanEval, MBPP, LiveCodeBench)；(d) Instruction-following (IFEval, Arena-Hard, MT-Bench, BFCL)。值得注意的是 Qwen 几乎不报告 ARC / HellaSwag / WinoGrande / PIQA 这类老 commonsense 项目——这与 Llama 系列保留"legacy 行"的策略形成对比。

Qwen-3 (2025-Q2, 0.6B 到 235B MoE, 含 Thinking / Non-Thinking 双模式)[^qwen3-blog] 把评测套餐进一步细化：

- **思考模式 (Thinking)** 专项报告 AIME 2024 / 2025 maj@1 + maj@32、GPQA-Diamond、MATH-500、LiveCodeBench v6 (锁 2025-04-25 release tag)。
- **非思考模式** 报告 MMLU-Pro、CMMLU、C-Eval、IFEval、Arena-Hard v2、BFCL、τ-bench retail + airline。
- **Coder 子系列**（Qwen3-Coder）独立报 SWE-bench Verified + SWE-bench Pro + Aider polyglot multi-language + LiveCodeBench Pro 月更。
- **VL 子系列**报 MMMU-Pro + MathVista + DocVQA + ChartQA + RealWorldQA。
- **悄悄删除项**：v1 时代的 TruthfulQA、TriviaQA、NQ 已不在 v3 主表（C-Eval / CMMLU 占据"知识 baseline"位置）。

Qwen 独有项：C-Eval + CMMLU + AGIEval **三件套**——头部西方 lab (Meta / OpenAI / Anthropic) 都不报这三个，只有 Qwen / DeepSeek / Kimi / 智谱 等中国 lab 必报。原因不是中文场景比英文重要，而是这三个 benchmark 是 2023-2024 中文社区认可的 "本地 baseline"——不报会被中文 reviewer 质疑可比性。但要小心：CMMLU 在 2026-03 已被报告含多解 / 错题 (Issue #91)，C-Eval 的 test labels 在 2025-07 公开后 contamination 风险陡升。Qwen-3 现在的 53.4% on CMMLU 与 Lingzhi-72B 90.26% 之间有约 36 pp 差距，但已经无法用"模型真实差距"解释——leaderboard 主要变成"是否在 train 时见过这套题"的指示器。

为何 Qwen 不弃 SWE-bench Verified：Qwen3-Coder 同时报 Verified + Pro 双数字，与 Meta / OpenAI 的做法相反。理由是中国开发者社区的 SWE-bench 数字接受度更高，删除 Verified 会失去"对照 Claude 4 / GPT-5 历史 release"的 anchor 行——长期看 Pro 才是 frontier benchmark，但短期 deprecation 太快会让自家模型在外网"看起来没有进步"。这是 release-card design 的政治维度，决策者要意识到它存在但不要被它绑架。

2.4 DeepSeek-V3 → DeepSeek-R1：reasoning 路线的 reference

DeepSeek-V3 (2024-12)[^deepseek-v3-tr] tech report 的 evaluation section 拆成 5 个 ablation 表 + 1 个对外比较表。对外比较表（Table 6）报 MMLU、MMLU-Redux、MMLU-Pro、BBH、CMMLU、C-Eval、AGIEval、HumanEval、MBPP、LiveCodeBench (v6 2024-

08 ~ 2024-11)、Codeforces percentile、SWE-bench Verified、Aider、GSM8K、MATH、AIME 2024、CMath、HumanEval-Mul、HumanEval-XL。"含中文 + 含 reasoning + 含 code multilingual"是 DeepSeek-V3 表的总特征，明显多于 Llama-3.1 但少于 Qwen-2.5 的中文项数。

DeepSeek-R1 (2025-01)[^deepseek-r1-tr] 是 reasoning 模型的 reference 报告：保留 V3 的所有 benchmark，再加 AIME 2024 maj@64、AIME 2025 maj@64、MATH-500、Codeforces ELO、LiveCodeBench-COT、GPQA-Diamond。AIME 报 maj@k 而非 pass@1 是 R1 论文最重要的方法论选择——R1 自己证明 maj@64 → pass@1 涨幅可达 14 pp (AIME 2024 71.0 vs 79.8)，pass@1 单点已经无法稳定区分 reasoning model[^deepseek-r1-tr]。

DeepSeek 独有项：HumanEval-Mul + HumanEval-XL (多语种 code) + CMath + Codeforces ELO。Codeforces ELO 是把模型放进真实 contest 用 actual rating 系统度量——R1 报 96.3% 百分位、相当于 expert 级 (~1750 ELO)。这是少数采用"模型 vs 人类 rating system"的 release。值得注意 R1 没报 FrontierMath（OpenAI 的标志性指标）也没报 τ-bench（Anthropic 的标志性指标）：**deliberate omission**——report 时如果 frontier 与自己同等水平的指标自己没准备打榜，干脆不报，避免把"对手主场"指标当镜子。这是 release-card 的另一面政治。

最近变化：DeepSeek-V3.1 (2025-Q3) 起开始报 AgencyBench 子集、Aider polyglot、SWE-bench Pro Public，但 SWE-bench Verified 仍保留——与 Qwen 同步采取"双轨"过渡策略。

2.5 OpenAI GPT-4o → GPT-5：system card 的"不报告即删除"政治

OpenAI 的 release format 与开源 lab 完全不同——他们不发 tech report，发 system card，重点在"safety + alignment evaluation"，capability 的 benchmark 表大幅压缩。GPT-4o (2024-05) system card 提到 0-shot MMLU 88.7、HumanEval 90.2、MATH 76.6、MMMU 69.1[^gpt4o-card]，加上 safety evaluation (RealToxicityPrompts, BBQ, GenderBias-Open, MMLU stereotype subset)。但 GPT-5 (2025-Q3, system card)[^gpt5-card] 的结构剧变：

- **保留**: MMLU-Pro、GPQA-Diamond pass@1 + maj@N、AIME 2025 (pass@1 + maj@64)、MATH-500、Humanity's Last Exam、FrontierMath Tier 1-3 (单独披露 Tier 4 是 internal)、Codeforces percentile、MMMU-Pro、τ-bench v2 retail + airline、SWE-bench Verified (legacy 行) + SWE-bench Pro Public + Private、Arena-Hard v2、LiveBench、HealthBench、MLE-bench、GDPval、AgentCompany、Multilingual MMLU、PaperBench。
- **悄悄删除**: HumanEval / MBPP **全部消失**——自 GPT-4o 后期起 OpenAI 不再单独 report；BBH 不再报；ARC-Challenge / HellaSwag / WinoGrande / TriviaQA / NQ 全部 retire；MMLU (5-shot) 仅出现在 "previous OpenAI generations" 历史表。
- **独有项**: HealthBench (医疗对话)、PaperBench (paper reproduction)、MLE-bench (ML research)、GDPval (经济产出价值)、AgentCompany (multi-agent 协作)。这些都是

10 / 33

OpenAI 自家研究团队设计的内部 benchmark，外部模型很少有数字可参考——这是典型的 "lab 设计私有指标 + 在 release card 主场使用" 策略。

OpenAI 选 / 不选的内在逻辑：(1) 严格按 frontier benchmark 优先级排序，凡是 5 家以上 lab 都接近天花板的 (HumanEval、MMLU)，主表删。(2) 维持自家"AI for science / autonomy"叙事的私有 benchmark 留主表——HealthBench / PaperBench / MLE-bench 都是 OpenAI 自创，对手没数字，自然在 release card 上压死无 SOTA。(3) **2026-Q1 起停报 SWE-bench Verified** 是个标志性事件——既然自家 audit 已经证明 Verified 含污染、且 SWE-bench Pro Public 已经形成 spreading scaffold (Anthropic / Meta / Google 都跟报)，**主动 deprecate 是把 contamination 责任转出去**。这意味着今后任何 lab 的 release card 上还在主报 SWE-bench Verified 的，等于自降可信度。

为何 OpenAI 不报 BBH / WinoGrande / HellaSwag：BHI 2026 在 91-model 分布上计算这几个 benchmark 的 Capability Discrimination (CD) 已接近 0——也就是说，在 frontier 模型间这个数字几乎无法区分能力[^bhi-2026]。OpenAI 主表删除这些 = 主动配合 BHI 的"健康度"筛选。这种 silently-aligned-with-meta-evaluator 行为是工业 frontier lab 的隐式协议，决策者跟 lab 学这一步是合理的。

2.6 Anthropic Claude 3.5 → 3.7 → Claude 4：内卷的 reasoning + agentic 路线

Anthropic 的 release card 风格介于 OpenAI 与开源 lab 之间——给具体数字但只挑 frontier 关键项。Claude 3.5 Sonnet (2024-06) 报 MMLU 88.3、MMLU-Pro 75.5、GPQA-Diamond 59.4、MATH 71.1、HumanEval 92.0、Multilingual MMLU、BBH 86.8、Arena-Hard v1[^claude35-card]。Claude 3.7 Sonnet (2025-02) 起增加 extended thinking 模式，单独报 AIME 2024 (maj@64) 与 MATH-500。Claude 4 (2025-Q3, Opus 4 + Sonnet 4 + Haiku 4) 的 card 结构与 GPT-5 高度对齐：

- **保留**: MMLU-Pro、GPQA-Diamond、AIME 2024 + 2025 (maj@64)、MATH-500、SWE-bench Verified + SWE-bench Pro Public + Private、HumanEval-Plus (注意是 Plus 不是原版)、MMMU-Pro、τ-bench v3 retail + airline + telecom + banking、Arena-Hard v2、LiveBench、AgencyBench (1M context + 90 tool calls)。
- **悄悄删除**: HumanEval (单独版本)、MBPP、BBH、Multilingual MMLU 完整版（保留子集）、HellaSwag、ARC-C / Easy、TruthfulQA。
- **独有项**: AgencyBench (Claude 4.5 自家场景设计的 1M-context agentic benchmark)、SHADE-Arena (代理欺骗 / sabotage 评测)、Anthropic Honesty (内部对齐评测)。

Anthropic 的指标设计哲学：他们倾向把 "long-horizon agentic + safety" 当主战场，每代 release 都加一两个 1M+ context、tool-use heavy 的自家 benchmark。这与 OpenAI 把 "private benchmark 在主表压死无 SOTA" 一脉相承——但 Anthropic 的私有指标更聚焦 alignment / agency。Claude 4.6 在 τ-bench overall 上 89.2、SWE-bench Pro 77.8——这两个数都让 Anthropic 在 2026-Q2 短期称霸 agentic-coding leaderboard[^anthropic-4-card]。

11 / 33

为何 Claude 4 弃 HumanEval / MBPP：跟 OpenAI 同样的理由——SOTA cluster 已 >95%、re-test (EvalPlus 80x test) 显示 +19 ~ +29 pp 掉幅，模型 ranking 在 plus 版与原版差异巨大[^liu-2023-evalplus]。Anthropic 报 HumanEval-Plus 是介于"删除"与"硬撑原版"之间的折中——既保留 surface name 让 reviewer 一眼认得，又用 plus 版作为真正的指标。这是工业 lab 在 benchmark 死亡曲线后段最常用的"延寿" pattern，但 plus 版数字仍然在 87+%、独立 lab 间差距 <2 pp。再过 6-12 个月 plus 版也会 retire，迁去 LiveCodeBench v6 + SWE-bench Pro 双轴。

2.7 Google Gemini 2 → 2.5 → 3：1M context + 多模态全栈

Google 的 release format 是 paper + technical blog + DeepMind report 混合。Gemini 1.5 paper[^gemini15-paper] (2024-02) 强调 1M-2M context 长度，benchmark 端报 MMLU、MMLU-Pro、Math、HumanEval、MMMU、Multilingual XLSum + 多个 1M context 任务 (NIAH 1M-Multimodal、video-with-haystack)。Gemini 2 / 2.5 / 3 的 release 风格逐渐向 Anthropic 靠拢——core capabilities 表 + multimodal 表 + long-context 表 + safety/responsible-scaling 表四档分开。Gemini 3 (2026-Q1, Pro + Flash)[^gemini3-blog] 的 core capabilities 表包括：

- **保留**: MMLU-Pro、GPQA-Diamond (Diamond 94.1 SOTA from Gemini 3.1 Pro)、AIME 2025 (maj@64)、MATH-500、HumanEval+ (legacy)、LiveCodeBench v6、SWE-bench Verified (legacy) + SWE-bench Live (Google 选 Live 而非 Pro)、MMMU-Pro、MathVista、DocVQA、Charts/Diagrams、Multilingual MMLU、Multilingual MGSM、Arena-Hard v2、LiveBench、τ-bench v3、RULER 1M+ (Gemini 3 是少数报 1M 以上 RULER 的)、IDR-Bench、AgencyBench。
- **悄悄删除**: HellaSwag、ARC-Challenge、PIQA、WinoGrande、SuperGLUE 已不在主表；HumanEval 原版用 + 替代。
- **独有项**: 多模态长上下文 (NIAH 1M-Multimodal-Video + audio-NIAH)、Multilingual XLSum 25 语种 + Multilingual MMLU 全 14 语种、Gemini Robotics 物理 evaluation (具身)。

为何 Google 选 SWE-bench Live 而非 SWE-bench Pro：Google 与 Microsoft 同期联合推 SWE-bench Live (NeurIPS 2025 D&B[^microsoft-2025-live])——Microsoft 主导但 Google 早期对外协作。Live 是 monthly rolling + cutoff filter (强 contamination 防御)，Pro 是 quarterly refresh + GPL 隔离 (中等 contamination 防御)。两条路线选哪一条主要看 release 节奏：Google 一年内多次大模型迭代，monthly rolling 更适合"持续报数"；Anthropic / OpenAI 季度大版本，quarterly refresh 与 Pro 节奏匹配。这是 release calendar 与 benchmark cadence 必须对齐的一个具体案例——下次有同事问"我们 release 季度更新该选 Pro 还是 Live"，答案是看你的发版节奏。

Gemini 独有的 1M+ multimodal NIAH 是 release card 上压死无对手的私有项——OpenAI / Anthropic / Llama 当前没有 1M 多模态 video haystack 评测能力。这种"用唯一支持的能力

定义 benchmark"是 frontier release 的常态——读者要意识到 release card 上的"独家 SOTA"中，约 1/3 来自这种 capability-frame 设计而非真实 model superiority。

2.8 OLMo / OLMo 3：开放栈的对照基准

AI2 OLMo 3 (2025-11-20)[^olmo3-blog] 是本章最重要的对照——它是当前唯一**数据 + checkpoint + 训练代码 + 评测代码**全部公开的 frontier-scale (32B dense) lab。OlmoBaseEval 含 43 个 benchmark，分 Base Easy（用代理 pretrain run）和 Base Main（full run）：

- **Base Easy** (`olmo3:base_easy`): MMLU、ARC-E、ARC-C、HellaSwag、PIQA、OpenBookQA、SIQA、CommonsenseQA、BoolQ、CoQA、SQuAD、TriviaQA、Jeopardy、NaturalQuestions、SciQ、Qasper、Basic-Skills、LAMBADA、MedMCQA、MedQA-EN、Lab-Bench (DBQA + Protocol)、Multilingual MMLU、HellaSwag-zh / fr / hi / es...（小代理 run 的 sanity）。
- **Base Main** (`olmo3:base`): 在 Base Easy 上加 MMLU-STEM / Humanities / Social Sci / Other 全集 (MC)、GSM8K、GSM-Symbolic、MATH、Minerva-Math、HumanEval (codex_humaneval)、MBPP、BigCodeBench、DS-1000、Multipl-E、DeepSeek-LeetCode、Codex-HumanEval-FIM 等 base pretrain 阶段可直接报的 reasoning + code variants。
- **Mid-train / Adapt 套餐** (`olmo3:base_chat` / `olmo3:adapt` / `olmo3:heldout`): OLMo 3 把 IFEval、AIME-2024 / 2025、BBH (CoT)、GPQA、AGIEval English、MMLU-Pro、Minerva-MATH-500、HumanEval+ (codex_humanevalplus)、MBPP+ (mbppplus)、LiveCodeBench、Alpaca-Eval v3、IFBench、SimpleQA、PopQA、ZebraLogic 等放在这一层。这是 OLMo 3 与多数工业 lab 最大的方法学差异：**严格 base pretrain 套餐 与 instruct / chat / heldout 套餐分离**。
- **悄悄删除（与工业 lab 对比）**: HumanEval+ / MBPP+ (EvalPlus) **不在 OlmoBaseEval base pretrain 套餐**（仅在 midtrain / adapt 套餐中评测）；SWE-bench Verified **完全不在 OLMES 任何套餐**（task_suites.py 零结果）；TruthfulQA **不在 OLMo 3 任何 base / chat / adapt 套餐**（仅出现在 OLMo 1 legacy 与通用 tulu instruct 套餐）；LiveCodeBench **不在 Base 套餐**（cutoff 不易在 base pretrain 中管理；仅在 midtrain / adapt 套餐中跑）；Arena-Hard / LiveBench / MixEval-Hard 不在 base；τ-bench / AgencyBench 不在；MMMU 不在（OLMo 3 没视觉模态）。
- **独有项**: instance-level prediction logging + 43-benchmark 全自动 export 到 Google Sheets / W&B (OLMES 内置)。这是工业 lab 至今没有完整复现的"完全可审计的 evaluation pipeline"。

为何 OLMo 3 仍保留 MMLU 5-shot + ARC-Easy + HellaSwag 这些已饱和项：理由不是因为他们没有看到饱和——OLMo 3 paper 公开承认 GSM8K / HumanEval 已无法区分 model recipe，主要用做"无回退" sanity (epoch N+1 不应低于 epoch N，否则 data mix 出问题)。**保留的真正原因是 open-pretrain-science 的 longitudinal 比较约束**：OLMo 1 / OLMo 2 / OLMo 3 都报告这些，删除会断裂 lineage。这是开源 lab 与工业 lab 的核心差异——工业 lab 没有"必须保持 historical comparability"的承诺，所以可以更激进地删除饱和项。

13 / 33

为何 OLMo 3 不报 LiveCodeBench / Arena-Hard / τ-bench：base model 上跑这三个意义不大。LiveCodeBench cutoff 需要严格 model release date，与 base run 的 epoch 报告不对齐；Arena-Hard / τ-bench 都需要 chat template (instruct), base 模型上跑只会看到"prompt 解析失败率"主导分数。OLMo 3 把这些移到 OlmoMidEval / OlmoInstructEval 套餐，与 base 严格分离。这印证 本报告 01-FRAMEWORK 决策表 "Early/Mid pretrain × SFT 排除项" 的逻辑。

2.9 七家 lab 的共同点与分歧

共同点（6/7 或 7/7 一致）

1. **MMLU-Pro 必报**（7/7）。MMLU 5-shot 单独留主表只剩 2/7（Llama-3.1 历史 + OLMo 3 longitudinal）。
2. **GPQA-Diamond 必报**（7/7）。即使只有 198 题、binomial CI 宽度 >5 pp，仍是 frontier"博士级推理"必须的 anchor。
3. **AIME 2024 / 2025 必报 maj@k**（6/7，OLMo 3 只报 2024-pass@1）。pass@1 单点已不被接受——decoders 直接 push back "为什么不报 maj@k"。
4. **HumanEval / MBPP 仅作 legacy 兼容行**（3/7 主报 + 4/7 已删）。工业 lab 全删，open 系 + 中国 lab 还保留兼容性。
5. **LiveCodeBench v6 锁 release tag 是默认 contamination-resistant code benchmark**（6/7）。OLMo 3 是唯一例外——base model 上 cutoff 不易管理。
6. **τ-bench / τ³-bench 是 agent capability 主要 anchor**（5/7）。Meta / DeepSeek 不报（不在 agent-product 主线）。
7. **SWE-bench Verified 在 2026-Q1 后被广泛 deprecate**（4/7 已删或迁 Pro）。

分歧（明显两营）

1. **OpenAI / Anthropic / Google 设计私有 benchmark 压死无 SOTA** vs **Meta / Qwen / DeepSeek / OLMo 用公开 benchmark longitudinal 比较**。这是 release-card 政治的最深分歧。决策者用对应路线时要意识到——你的 release card 要走哪种风格，决定你的 benchmark 套餐能不能跟工业 SOTA"直接比"。
2. **中文 lab (Qwen / DeepSeek) 必报 C-Eval / CMMLU / AGIEval 三件套** vs **西方 lab 全部不报**。这不是"中文重要不重要"的问题——是"对中文 reviewer 与中文开发者社区有 anchor 行" vs "对西方 frontier 有 anchor 行"两套 anchor 系统。两套 anchor 不可强行合并：在做面向中文社区的 release 时，三件套是必报；面向英文社区时，三件套是 noise。
3. **Anthropic / OpenAI 把 long-horizon agentic 当主战场** vs **Meta / OLMo 把 base model evaluation 当主战场**。Llama / OLMo 不报 τ-bench / AgencyBench 不是"做不到"，是 release-card 战略选择不同。

14 / 33

最近 6 个月（2025-12 → 2026-05）的趋势

- **2026-Q1**: SWE-bench Verified 集体 deprecate（OpenAI 主动停报 → Anthropic / Google / Meta 跟进迁 Pro）。决策者从此**不应**在新 release card 上把 Verified 当主指标。
- **2026-Q1**: FrontierMath Tier 1-4 grading review (~1/3 题有 fatal errors[^epoch-tier4])，OpenAI / Google 仍在主报但加 provisional caveat；Anthropic 已悄悄从主表移走。任何 release 引 FrontierMath 现行数字都应加 `[provisional, pending 2026-Q3 reissue]`。
- **2026-Q1-Q2**: τ-bench tau2-telecom 出现 99% 饱和[^tau-bench-leaderboard]，τ³-bench (telecom + banking + voice) 接班；下一代 release 应报 τ³ 而非 τ²。
- **2026-Q2**: GPQA-Diamond 多数 frontier model >92%[^gpqa-epoch]——Burnham (Epoch AI) 已把 Diamond 列入 "saturation watch"。建议同步报 BBEH / ARC-AGI / FrontierMath 做 cross-stream。
- **2026-Q1-Q2**: 7/7 lab 都开始报告 RULER 128K 或更长。**1M+ context** 是 release card 上的新底线——Google / Anthropic / OpenAI / Meta / Qwen 全报，OLMo 3 还在 32K（开源 lab 计算预算限制）。

2.10 把别家的套餐 import 进自己的决策表

把本章七家 lab 的具体选法回放进 §01-framework 决策表，三条 import 规则：

1. **学谁取决于你的 release 政治路线**。若你做开源 longitudinal 比较（OLMo 风格），照搬 OlmoBaseEval 43 项 + Meta 的 contamination disclosure 风格；若你做 frontier release card（GPT-5 / Claude 4 风格），照搬 OpenAI/Anthropic 的 7-9 项核心 + 1-2 项私有指标策略。**不要混用**——同一个 release 既报 OLMo 43 项又报 OpenAI 7 项会让读者无法 anchor。
2. **照搬"删除项"比照搬"保留项"重要**。frontier lab 的删除是 strong signal：OpenAI / Anthropic / Google 集体不报 BBH = BHI Capability Discrimination 已经接近 0。你的 release card 还在主报 BBH 等于自降可信度。同样 GSM8K、HumanEval、ARC-Easy、HellaSwag、WinoGrande、TruthfulQA、TriviaQA、NQ 这一组——5/7 以上 lab 删除时跟着删。
3. **私有 benchmark 是 release-card 的"主场加分项"**。如果你的 lab 没有自己设计的 benchmark，frontier-level release card 上压不出 SOTA——你只能在公开榜单上"second-best"。这是 2026 开始 frontier lab 标准动作：每代 release 必须有 1-2 个自家命名、对手没法快速接的内部指标 (HealthBench / AgencyBench / PaperBench / MLE-bench / GDPval)。

下一章 §03 把"哪些 benchmark 已不可信"展开到具体的迁移决策与避坑 checklist；§04 把这套反向工程结论落到 5 个具体场景的速查 cookbook。本章的结论可压缩成一句话：**在 2026-Q2 的 frontier release 上，MMLU-Pro / GPQA-Diamond / AIME 2024-2025 (maj@k) / MATH-500 / LiveCodeBench v6 / SWE-bench Pro / τ-bench / Arena-Hard v2 / LiveBench 这 9 项是 "近似 7-lab 共识" 主榜；C-Eval/CMMLU/AGIEval 是中文社区**

15 / 33

anchor；HumanEval / MBPP / SWE-bench Verified / BBH 已退到 legacy 兼容行；OLMo 3 是开放栈对照基线，工业 lab 几乎没有义务跟进它的 43-项完整集合。

引用

[^llama3-card]: Dubey, A. et al. (2024). *The Llama 3 Herd of Models*. arXiv:2407.21783. Includes Table 15 contamination overlap analysis (52% NQ test items in pretraining corpus). Model card at https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md.

[^qwen25-blog]: Qwen Team, Alibaba (2024-09). *Qwen2.5: A Party of Foundation Models*. https://qwenlm.github.io/blog/qwen2.5/. Tech report arXiv:2412.15115.

[^qwen3-blog]: Qwen Team, Alibaba (2025). *Qwen3 Technical Report*. https://qwenlm.github.io/blog/qwen3/. arXiv:2505.09388. Includes thinking/non-thinking dual-mode evaluation tables.

[^deepseek-v3-tr]: DeepSeek-AI (2024-12). *DeepSeek-V3 Technical Report*. arXiv:2412.19437. https://github.com/deepseek-ai/DeepSeek-V3.

[^deepseek-r1-tr]: DeepSeek-AI (2025-01). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv:2501.12948. Reports AIME 2024 pass@1 71.0 vs maj@64 79.8.

[^gpt4o-card]: OpenAI (2024-05). *GPT-4o System Card*. https://openai.com/index/gpt-4o-system-card/. Retrieved 2026-05-25.

[^gpt5-card]: OpenAI (2025). *GPT-5 System Card*. https://openai.com/index/gpt-5-system-card/. Retrieved 2026-05-25. Includes HealthBench / PaperBench / MLE-bench / GDPval / AgentCompany / Humanity's Last Exam.

[^claude35-card]: Anthropic (2024-06). *Introducing Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet. Model card and benchmark table retrieved 2026-05-25.

[^anthropic-4-card]: Anthropic (2025). *Claude 4 System Card*. https://www.anthropic.com/claude/claude-4. Includes τ-bench v3, AgencyBench, SHADE-Arena.

[^gemini15-paper]: Gemini Team, Google (2024-02). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. arXiv:2403.05530. First 1M / 2M context release.

[^gemini3-blog]: Google DeepMind (2026-Q1). *Gemini 3 Pro / Flash Technical Report*. Retrieved 2026-05-25 from https://deepmind.google/. Includes 1M+ multimodal NIAH / video / audio long-context eval.

16 / 33

[^olmo3-blog]: Allen Institute for AI (2025-11-20). *Olmo 3: Charting a path through the model flow to lead open-source AI*. https://allenai.org/blog/olmo3. OlmoBaseEval 43-benchmark set bundled into OLMES.

[^microsoft-2025-live]: Liu et al. (2025). *SWE-bench Live: Towards Robust Evaluation of Code Agents*. NeurIPS 2025 D&B. arXiv:2505.23419.

[^bhi-2026]: Zhu, L., Hua, H., Miao, L., Zhao, B. (2026-02). *Benchmark Health Index: A Systematic Framework for Benchmarking the Benchmarks of LLMs*. arXiv:2602.11674.

[^liu-2023-evalplus]: Liu, J. et al. (2023). *Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation* (EvalPlus / HumanEval+). arXiv:2305.01210.

[^epoch-tier4]: Epoch AI (2026-05-11). *FrontierMath Tier 4: AI-Assisted Review Finds ~1/3 Fatal Errors*. https://epoch.ai/benchmarks/frontiermath-tier-4.

[^tau-bench-leaderboard]: BenchLM.ai (2026-05). *τ-bench / τ²-bench / τ³-bench leaderboard*. https://benchlm.ai/benchmarks/tauBench. https://llm-stats.com/benchmarks/tau2-telecom.

[^gpqa-epoch]: Burnham, G. / Epoch AI (2025-2026). *GPQA Diamond: What's Left?* https://epoch.ai/gradient-updates/gpqa-diamond-whats-left. Includes saturation-watch update 2026-Q1.

[^scale-2026-swe-pro]: Scale AI (2026). *SWE-Bench Pro: A Stronger Coding-Agent Benchmark*. https://scale.com/blog/swe-bench-pro. Public + Private + Held-out splits, quarterly refresh.

[^dekoninck2026matharena]: Dekoninck, J. et al. (2026). *Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs*. arXiv:2605.00674.

读完 Part I/II 你已经知道每个 benchmark 怎么跑、怎么被批评。本章把这些事实压缩成一份**决策者拒报清单**：在 release card、技术报告、或对 leadership 汇报里出现哪些数字 = 项目掉信任分。每一条都给「饱和点 / 污染证据 / 工程陷阱」三类原因之一，并附「若你已在用它怎么办」的迁移路径。

一张总清单

下面这张表把 14 个常被引用、但 2026-05 已**不应作为前沿模型主指标**的 benchmark 一次列清，按问题类型分簇——读者可以把它直接当 release-card review checklist。

Benchmark	问题类型	关键数字 / 证据	还能用作什么	推荐替代
HumanEval	饱和 + 污染	EvalPlus pass@1 99.4 / 80x 增测试后 frontier 仍掉 19-29 pp	5 min compliance smoke test	LiveCodeBench v6 + SWE-Bench Pro
MBPP	饱和 + 污染	Mistral / Llama / DeepSeek 系列均 >85，与 HumanEval 同源问题	同上	同上
GSM8K	饱和 + 反向工程	SOTA 92-95，repo 2026-04 archived；GSM-Symbolic 单 clause 注入造成最多 65% 性能掉幅[^gsm-symbolic]	Small-model (<3B) 或 mid-train ablation sanity

check | MATH-500 + AIME 2026 + GSM-Symbolic | | 早期 MMLU (单一 5-shot acc) | 部分饱和 + 错题 | Top model 88.7，MMLU-Redux 6.49% 错题 / Virology 子集 57% 错[^mmlu-redux] | 知识基线但必须分子集 + 与 MMLU-Pro 并报 | MMLU-Pro + MMLU-Redux subset | | HellaSwag | 饱和 + 任务无效 | Claude 3 Opus 95.4；GoldenSwag 显示 ≥65% 预测在 删除 question / Lorem ipsum 替换后不变[^goldenswag] | Cheap base-model probe (<10s) | GoldenSwag subset | | WinoGrande | 饱和 + artifacts | GPT-4 87.5；WinoWhat 显示 paraphrase 后系统性 drop[^winowhat] | <10B 模型 sanity check | 退役 | | PIQA | 饱和 + 单语 | Phi-3.5-MoE 88.6；Global PIQA 显示英文 vs low-resource 差 37 pp[^global-piqa] | <3B 模型 sanity check | Global PIQA | | ARC-Easy | 饱和 | Mixtral 8x7B 83.1；frontier 模型已不报告 | <6B 小模型套件 | 退役 | | TriviaQA | 污染 | Lewis 2020：60-70% test 答案在训练集出现[^lewis2020] | Closed-book knowledge baseline | 与 NQ + 检索基准并报 | | Natural Questions | 污染 | Lewis 2020 同；Llama 3 Table 15 显示 52% test 题在预训练 corpus 中[^llama3-card] | Closed-book knowledge baseline | 同上 | | LAMBADA | 饱和 + 协议歧义 | GPT-4 87.8 已饱和；OpenAI variant 与 Paperno 原版不可直接比 | Pretrain 早期 (<1B) loss-correlated probe | 退役 | | GPQA-Diamond (单 pass@1) | 饱和 + 不平衡 + grader 噪声 | Gemini 3.1 Pro 94.1；Organic Chemistry 占难题 70% 但题量 36%[^epoch-gpqa]；198 题 1-2 pp 差距已落入 binomial CI | Reasoning 必报 + 并报子领域；不能单一数字 | GPQA-Diamond + FrontierMath cross-stream | | FrontierMath (单点数字) | 题库 fatal error + sponsor disclosure | 2026-05 Epoch 自爆 ~1/3 Tier 1-4 题有 fatal errors[^epoch-tier4]；OpenAI 资助未及时披露 | 上限指标但必须配对独立流 | + Soohak + Riemann-Bench + MathArena ArxivMath | | SWE-bench Verified | 污染 + 弱测试 | SWE-Bench+ 论文显示 32.67% pass 来自 issue 评论里的答案，31.08% 来自弱测试[^swe-plus]；OpenAI 已停报告 | 历史对照；2024 模型重测 | SWE-Bench Pro + SWE-bench Live | | AIME 2024 / 2025 (静态) | 饱和 + 题量噪声 | GPT-5 95.7 / Step-3.5 99.9；30 题/届，pass@1 颗粒度 3.3 pp | Contamination diagnostic（与 AIME 2026 配对） | MathArena (月更 AIME 2026 + ArxivMath) | | GAIA validation | 答案泄漏 + 反向工程 | BenchJack 自动 exploit 在 GAIA 达 98% 真实解题率为 0[^berkeley-2026] | Test set (private) + scaffolded eval；不能用 validation 训练 | Gaia2 + private-set 提交 | | WebArena 原版 | 反向工程 + 评测器漂移 | BenchJack 100% exploit；ServiceNow 审计 506/812 任务 permissive matching[^servicenow-2026] | 历史对照 | WebArena-Verified Hard-258 | | OSWorld 原版 | Through-time 不可比 + 反向工程 | 2025-07 Verified 改 300+ task，之后又改 10% 指令[^epoch-osworld]；BenchJack 73% exploit | 锁定 commit 后做 quick-iteration | OSWorld-Verified (锁版本) + OSWorld-MCP |

读这张表的方法：**右侧两列才是这一章真正的产出**。决定要不要砍一个 benchmark，是看「替代是什么」+「替代的边际成本」；纯抒情说「XX 不能用」对 release-card 改动不构成 actionable advice。

饱和簇：HumanEval / GSM8K / 经典三件套 / 早期 MMLU

这些 benchmark 共同的问题是 **frontier 模型差距已坍缩到 noise 量级**——具体数字见上表。Epoch AI 2026 把 GPQA-Diamond 也列入「saturation watch list」[^epoch-gpqa]，原

18 / 33

因是 Diamond 198 题 + 1-2 pp 差距已落入 binomial CI；这点常被忽视。

迁移建议有三档：

- **没钱重做套件**：只在 release 表里把这些 benchmark 标为「Legacy / saturated」加灰，并在同一行报告 Pro/Plus/Hard variant（HumanEval+ / MMLU-Pro / GoldenSwag）。读者一看就知道你知道。
- **有 1 周时间**：把所有 saturated 项归并到「Pretrain History Section」，主表只留 contamination-resistant 项目（LiveCodeBench v6 / Arena-Hard v2 / LiveBench monthly / MathArena ArxivMath）。
- **有 1 month 时间**：仿照 OLMES「Base Easy / Base Main」分层[^olmes]，把每个能力轴留 1 个 saturated benchmark 作 monotonic-progress proxy（pretrain 早期信号），1 个 hard / contamination-resistant 作 frontier 区分。BHI 2026[^bhi-2026] 给的 Capability Discrimination × Anti-Saturation 雷达图可用作筛选 dashboard。

特别提醒：**不要复活已退役的 benchmark 当 unique selling point**。比如 ARC-Easy 在 6B 以下模型 paper 里偶有惊艳数字，但 reviewer 会立刻问「为什么不报 ARC-Challenge」——后者 frontier 模型也接近 96，但至少还在主流模型卡里。

污染簇：HellaSwag / WinoGrande / TriviaQA / NQ

这一簇的特点不是「饱和」而是「分数失真」——上表列的 Lewis 2020 / Surge AI / GoldenSwag / Llama 3 Table 15 等证据已足够把这些数字从可信变成 misleading。HellaSwag/WinoGrande 还叠加 **任务无效性**：模型不读 question 也能做对。

迁移建议（按时间预算分档）：

- **HellaSwag → GoldenSwag**（同协议、过滤掉 ≥65% 不读 question 也能答对的题）；只额外跑 GoldenSwag 5 min 即可，**强烈建议在 base eval 套件里替换**。
- **WinoGrande**：>10B 模型已退役，仅作 <10B 小模型 sanity；想要"反 shortcut"信号去看 ANLI 或 contrastive 类（NLI 反例对）。
- **TriviaQA / NQ**：闭书 knowledge recall 还需要的话，**与 RAG eval 并报**；单跑 closed-book EM 会让读者认为你不知道 contamination。

更深一层的方法论变化：Soft Contamination 2026[^soft-contam-2026] 实证 78% CodeForces / 50% ZebraLogic 题目在训练 corpus 中存在**语义重复**，n-gram filter 全部漏检。即便你换用 LiveBench / Arena-Hard，仍要在 caveat 段写「分数包含 shallow 泛化部分」，避免被 reviewer 当作 naive。

Grader 不稳定簇：GPQA-Diamond / FrontierMath / Arena-Hard

不同于饱和/污染，这一类**数字本身在波动**。三个典型：

19 / 33

- **GPQA-Diamond**：Burnham (Epoch AI)[^epoch-gpqa] 估算约 10% 题目有效性可疑，且 Organic Chemistry 子领域占难题 70% 但题量仅 36%——意味着两个不同模型差 2 pp 完全可能源于这一子领域的 binomial 噪声而非真实推理差距。**必须按子领域分项报告**，否则单数字 misleading。
- **FrontierMath**：2026-05-11 Epoch 公告 AI-assisted review 发现 ~1/3 Tier 1-4 题有 fatal errors[^epoch-tier4]，leaderboard 已 frozen 等待人工复核。这是**最严重的 grader 危机**——sponsor disclosure 争议（OpenAI 资助未及时披露）+ 题集质量都需要重判。**任何引用 FrontierMath 数字的 release card 必须标注「provisional, pending 2026-Q3 reissue」**。短期内迁移：与 Soohak（独立 + 全私有 + 完全披露 governance）和 Riemann-Bench 并报，避免单一 source。
- **Arena-Hard**：v2.0 引入 Gemini-2.5 + GPT-4.1 ensemble judge，但 judge model 仍是商业模型，self-preference bias[^self-pref-bias] 与 verbosity bias[^verbosity-bias] 未根除。**用 Style Control + Hard subset + dual judge** 是当前最低姿势线，少一项 reviewer 都有理由打问号。Judge Reliability Harness 2026[^judge-harness-2026] 在 4 SOTA judge × 4 benchmark 上的结论是「没有一个 judge 在所有 benchmark 上 uniformly 可靠」，意思是哪怕你换 judge 也得 per-task 验证。

Agent 复现陷阱：BenchJack 三连

2026-04 Berkeley RDI 发布 *How We Broke Top AI Agent Benchmarks* (Wang, Mang, Cheung, Sen, Song)[^berkeley-2026]——一个自动 exploit agent (BenchJack) 在三大 agent benchmark 上的真实解题率为 0 但 reported score 达到：

- **GAIA**：~98%（利用 HF 上 validation 公开 answer key + normalization collision）。
- **OSWorld**：73%（agent-evaluator 隔离不足 + 不安全 code execution）。
- **WebArena**：~100%（答案 key 暴露 + 弱 string match + sandbox 漏洞）。

这不是「论文做对了 model 没做对」——是 **benchmark 设计 / sandboxing / evaluator 写法本身允许 bypass**。任何报告这三个 benchmark 数字而不说明 (a) 用了哪个 release / commit、(b) sandbox 配置、(c) 是否用 private test set 的 release card，已经默认不可信。

迁移建议：

- **GAIA**：迁到 **Gaia2**[^gaia2]（异步/动态环境 + write-action verifier）+ 用 Princeton HAL scaffolded leaderboard 数字而非裸 GPT-5；validation 集**不要拿来训练 agent**。
- **WebArena**：直接 `pip install webarena-verified` [^servicenow-2026]，主报 Hard-258 子集。
- **OSWorld**：**锁定 commit + 用 Verified 子集**；不要做时间序列对比，因为 2025-07 起任务持续在修。

更隐蔽的陷阱：**Scaffold sensitivity**。AgencyBench 2026 显示同一 Claude-4.5-Opus 在 native Claude-Agent-SDK vs 第三方 SDK 上差距可达 20.5%[^agencybench-scaffold]。也就是说「Claude 4.5 在 X 上 60% / GPT-5 在 X 上 55%」这种比较，没有声明 scaffold 几乎无意义。Release card 必须给 scaffold 名 + version pin。

20 / 33

「我已经在用它怎么办」——一页迁移决策

最后一张表，给你的 PR / release card / paper draft 直接套用：

| 你现在报的 | 不删 + 加 caveat 怎么写 | 一步迁移 | 三步迁移 | |---|---|---|---| | HumanEval pass@1 | 加报 HumanEval+（80x 测试）+ 标 "saturated" | + LiveCodeBench v6 | + SWE-Bench Pro Public + Private | | GSM8K | 标 "legacy saturating" | + MATH-500 | + MathArena ArxivMath 月更 | | MMLU (5-shot) | 报告时并列 MMLU-Pro | + GPQA-Diamond 子领域分项 | + MMLU-Redux audit subset | | HellaSwag | 同时报 GoldenSwag subset | 退役 HellaSwag 列 | 替换为 reasoning probe + Global PIQA | | TriviaQA / NQ | 同时报 RAG-with-retrieval 数字 | 替换为闭/开书对比 | 加 Lewis 2020 overlap audit | | GPQA-Diamond | 子领域分项 + Organic Chem 单算 | + FrontierMath cross-stream | + ARC-AGI + BBEH 三平台并报 | | AIME 2024 / 2025 | maj@32 + 标 "saturated" | 迁 AIME 2026 + USAMO 2026 | + MathArena ArxivMath + ArXivLean | | GAIA validation | 切到 test set + 用 HAL scaffolded | 迁 Gaia2 | + 自家私有 task subset | | OSWorld 原版 | 锁 commit + 标版本 | OSWorld-Verified | + OSWorld-MCP（ICLR 2026） | | WebArena 原版 | 加 audit caveat | WebArena-Verified Hard-258 | + VisualWebArena + VideoWebArena | | SWE-bench Verified | 标 "deprecated by OpenAI" | SWE-Bench Pro Public | + Pro Private + SWE-bench Live |

操作语录：决策者在 review 团队 release card 时只需问三句——「你报的是 frontier 区分项还是 history 项？」「污染 / 饱和 caveat 在哪一行？」「agent 类 benchmark 你写了 scaffold 和 commit pin 吗？」。这三问答不上来的 release card 就让作者回去补；这是把 Part I/II 的证据转化为日常治理动作的最简化路径。

第 4 章把「替代套餐」从抽象单项展开为 5 个具体场景的完整推荐 + 反向排除清单。

引用

[^gsm-symbolic]: Mirzadeh, I. et al. (2024). *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. ICLR 2025. arXiv:2410.05229.

[^mmlu-redux]: Gema, A. P. et al. (2024). *Are We Done with MMLU?* arXiv:2406.04127.

[^goldenswag]: Chizhov, P., Nee, M., Langlais, P.-C., Yamshchikov, I. P. (2025). *What the HellaSwag? On the Validity of Common-Sense Reasoning Benchmarks*. arXiv:2504.07825.

[^winowhat]: WinoWhat authors (2025). *WinoWhat: A Parallel Corpus of Paraphrased WinoGrande Sentences with Common Sense Categorization*. arXiv:2503.23779.

[^global-piqa]: Chang, T. A. et al. (2025). *Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures*. arXiv:2510.24081.

[^lewis2020]: Lewis, P., Stenetorp, P., Riedel, S. (2020). *Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets*. arXiv:2008.02637.

[^llama3-card]: Dubey, A. et al. (2024). *The Llama 3 Herd of Models*. arXiv:2407.21783 (see Table 15 contamination analysis: 52% NQ test set in pre-training corpus).

21 / 33

[^epoch-gpqa]: Burnham, G. (2025). *GPQA Diamond: What's Left?* Epoch AI gradient-updates. https://epoch.ai/gradient-updates/gpqa-diamond-whats-left.

[^epoch-tier4]: Epoch AI (2026). *FrontierMath Tier 4*. https://epoch.ai/benchmarks/frontiermath-tier-4.

[^swe-plus]: Aleithan, R. et al. (2024). *SWE-Bench+: Enhanced Coding Benchmark for LLMs*. arXiv:2410.06992.

[^berkeley-2026]: Wang, H., Mang, Q., Cheung, A., Sen, K., Song, D. (2026). *How We Broke Top AI Agent Benchmarks: And What Comes Next*. Berkeley RDI, April 2026. https://rdi.berkeley.edu/blog/trustworthy-benchmarks-cont/.

[^servicenow-2026]: ServiceNow Research (2026). *WebArena Verified: Reliable Evaluation for Web Agents*. NeurIPS 2025 / OpenReview 94tlGxmqkN. https://github.com/ServiceNow/webarena-verified.

[^epoch-osworld]: Epoch AI (2026). *What does OSWorld tell us about AI's ability to use computers?* https://epoch.ai/blog/what-does-osworld-tell-us-about-ais-ability-to-use-computers.

[^gaia2]: Froger, R. et al. (2026). *Gaia2: Benchmarking LLM Agents on Dynamic and Asynchronous Environments*. ICLR 2026 oral. arXiv:2602.11964.

[^soft-contam-2026]: Spiesberger, A. et al. (2026). *Soft Contamination Means Benchmarks Test Shallow Generalization*. arXiv:2602.12413.

[^self-pref-bias]: Yu, J. et al. (2024). *Self-Preference Bias in LLM-as-a-Judge*. arXiv:2410.21819.

[^verbosity-bias]: Saito, K. et al. (2023). *Verbosity Bias in Preference Labeling by Large Language Models*. arXiv:2310.10076.

[^judge-harness-2026]: Dev, S., Sloan, A., Kavner, J., Kong, N., Sandler, M. (2026). *Judge Reliability Harness: Stress Testing the Reliability of LLM Judges*. ICLR 2026 Agents in the Wild Workshop. arXiv:2603.05399.

[^agencybench-scaffold]: Shi, J. et al. (2026). *AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts*. ACL 2026 Main. arXiv:2601.11044.

[^olmes]: Gu, Y., Tafjord, O., Kuehl, B., Haddad, D., Dodge, J., Hajishirzi, H. (2024). *OLMES: A Standard for Language Model Evaluations*. arXiv:2406.08446 (Findings of NAACL 2025).

[^bhi-2026]: Zhu, L., Hua, H., Miao, L., Zhao, B. (2026). *Benchmark Health Index: A Systematic Framework for Benchmarking the Benchmarks of LLMs*. arXiv:2602.11674.

第 3 章告诉读者**哪些数字不能再单独引用**；本章把镜像翻过来——**实际场景下该跑什么**。下面列 5 个最常出现的决策情境，每个给「推荐套餐 3-5 项 + 反向排除 1-2 项 + why」。

22 / 33

读者已读过 Part I/II（pipeline / contamination / saturation），这里直接给配方。

每个场景末尾的「执行 checklist」是套餐能否真的落到 dashboard 的关键——benchmark 选好了，cadence / scaffold / leaderboard 提交方式没定，半年后回看仍是 random number generator。

场景 1：自家 base model 中检（数据 ablation）

目标：pretrain 进行中（每 5k–100k step），快速判断「这个 data recipe / curriculum / mixture 是不是比上一版好」。受众：数据 / pretrain infra 团队。

预算：单次 ≤ 30 min；8×A100 80G 跑完。

推荐套餐 (5)：

| Benchmark | Cadence | 单次开销 | 用处 | |---|---|---|---| | LAMBADA (`lambada_openai`) | 每 5k step | <30 s | Loss-correlated 最佳 cheap probe；中早期 pretrain (<1B token) 单调上升 | | ARC-Easy + ARC-Challenge | 每 10k step | ~2 min | Knowledge baseline；中规模 1-7B 模型期间区分度仍在 | | MMLU-STEM 子集 (29 subjects) | 每 50k step | ~5 min | 知识 + 推理混合；只跑 STEM 避免 humanities 子集的 prompt 敏感性 | | GSM8K-light (200 题子采样, maj@1) | 每 50k step | ~3 min | Multi-step 推理 sanity；用子采样避免 GSM8K 全 1.3k 的 generate-until 开销 | | LiveBench monthly subset | 每 100k step | ~10 min | Contamination-resistant 锚点；用最近月份子集减少污染 |

可选超优解：**In-Training Probing**[^liu2026probing] 替换 5k / 10k cadence 项。在 OLMo3-7B 上 Submodel Probe 在 8 benchmark 上 average AUROC 0.7890，每 checkpoint ~1h → ~3min (20x)。代价：每个 target benchmark 要先训 probe；< 1T token 的小项目 ROI 不高。

反向排除 (2)：

- **重 reasoning（GPQA-Diamond / FrontierMath / AIME）**——7B 以下 base 模型答错率 >70%，accuracy 被 binomial 噪声主导，无法分辨 data recipe 差异。BHI[^bhi-2026] 的 Capability Discrimination 轴：用 saturated 或 zero-floor benchmark 做 ablation 都浪费 GPU-hour。
- **agent benchmark**——cold base 无 tool-use，数字全为 0；推到 SFT 后再跑。

执行 checklist：

- 5 个 task 写一个 OLMES recipe，`olmes` CLI 跑——prompt template / few-shot pool 固定，跨 step 可比。
- LiveBench pin release tag（如 `livebench-2025-04-25`），否则月更会"凭空抬高"曲线。
- 结果上传 W&B / sheet；不要只在 GPU 节点本地。

23 / 33

场景 2：申请发表 / 写 paper

目标：NeurIPS / ACL / ICLR 论文需要给 reviewer 一个「我们模型在 capability 谱系上的位置」。受众：论文一作 / 数据负责人。

预算：1 周 GPU-hour 跑完整套餐 + 微调；reviewer 接受 1-2 weeks 重做。

推荐套餐 (5-6)：

Benchmark	角色	Why
MMLU-Pro (5-shot CoT, 10 选 1)	知识 + 推理主指标	替换早期 MMLU；GPT-4o 从 MMLU 88.7 掉到 MMLU-Pro 72.55，证明区分度回归
GPQA-Diamond（必报 sub-domain）	STEM reasoning 主指标	必须按子领域 (Physics / Chemistry / Bio) 分项报告，避免 organic-chem 偏置；和 Burnham 2025[^epoch-gpqa] 的 effective ~90% rate 一起 caveat
HumanEval+ + LiveCodeBench v6	Code 主指标	HumanEval+ 比 HumanEval 多 80x 测试，drop 19-29 pp[^liu-evalplus]；LiveCodeBench 按 release date 切窗，post-cutoff 评测可控 contamination
MATH-500 + AIME 2026 (MathArena)	Math 主指标	MATH-500 cheap baseline + AIME 2026 via MathArena[^matharena] 给 contamination-free 锚点；不要单独引 AIME 2024/2025
Arena-Hard v2.0 (Hard + Style Control + dual-judge)	Instruction-tuned 用户偏好	当前最强 LMSYS-Arena offline approximator (87.4% separability)；必须报 style control 分数否则 verbosity bias 抬升
Your specialty domain (e.g. ScienceBench / SWE-Bench Pro / C-Eval Hard)	Differentiator	决定 reviewer 接受度的 50%；选 1 个与论文 contribution 直接对位的领域 benchmark

反向排除 (2)：

- **已饱和轻量项（HumanEval 原版 / MBPP / GSM8K / HellaSwag / WinoGrande）**——放主 table reviewer 直接划掉。SWE-Bench Verified 已饱和 + 污染，OpenAI 公开停报告[^swe-pro]，引用必标 "deprecated"。
- **MMLU 5-shot 单数字**——MMLU-Redux 报 6.49% 错题 / Virology 子集 57% 错[^mmlu-redux]；只报 MMLU 不报 MMLU-Pro 会被批"未跟进 2024+ 文献"。

执行 checklist：

- Table 表头写明 `pass@1 (greedy)` / `maj@32` / `Hard subset only` 等限定；不写默认挑樱桃。
- Contamination-resistant 项 (LiveCodeBench / AIME 2026 / Arena-Hard v2) 给 release tag 或 cutoff date。
- Specialty domain 若新建，paper 专章讨论 contamination control（参考 SWE-Bench Pro 的 GPL + private split[^swe-pro]）。

场景 3：公司内部模型上线评审

目标：模型进生产（toC / 内部 copilot / agent），合规 / 安全 / 评审委员会要「能不能上」的报告。受众：ML platform / SRE / 法务 / 安全团队。

预算：决策权重高，可投 2-4 周综合评测；**输出形式**比单次跑分更重要——要有 audit trail + per-instance log。

推荐套餐 (4-5)：

| Benchmark | 角色 | Why | |---|---|---| | HarmBench (Standard 200 + Contextual 100 + Multimodal 110) | Red-team attack robustness | ICML 2024 标准化，统一 attack methods（GCG/PAIR/AutoDAN）；含 Llama-2-13B classifier 自动判 ASR%[^harmbench] | | AIR-Bench 2024 (HELM-hosted v1.1.0) | Regulation-anchored content safety | 5,694 prompts 跨 8 政府监管 + 16 公司政策；EU / 中国 / US 三套 taxonomy 同时覆盖[^airbench] | | AgentBench-FC 或 AgencyBench (若 agent 场景) | Agent capability stress test | AgencyBench[^agencybench] 给 1M-token + 90 tool call + 数小时长程任务，能暴露 long-horizon failure 模式 | | MCPMark + MCP-Atlas (若需 tool-use) | MCP 工具使用 | MCPMark[^mcpmark] 测 CRUD 深度，MCP-Atlas[^mcp-atlas] 测跨 36 servers 广度；二者互补 | | Internal evals (>500 query 的 internal-domain set + LLM-judge + 人工抽检) | Domain-specific quality | 学术 leaderboard 不能替代公司业务 distribution；必须自建并定期 audit |

反向排除 (2)：

- **纯学术 leaderboard（HF Open LLM Leaderboard / MMLU 主榜 / Arena-Hard 排名）**——相对排序、无 absolute threshold；上线需要「helpful rate ≥ X% / refusal rate ≤ Y%」绝对值。学术分数只能做参考，不能 gate。
- **单 judge 评分**——Judge Reliability Harness 2026[^judge-harness-2026] 实证 4 SOTA judge 在 4 个 benchmark 上「没有一个 uniformly 可靠」；上线场景必须双 judge + 人工抽检。

执行 checklist：

- HarmBench / AIR-Bench 用 official harness；attack budget (10/100/1000) 三档明示。
- Internal eval frozen golden subset (~10%) 季度抽检 judge drift。
- 报告含「失败 case 抽样」附录（每类 ≥ 5 例），便于产品 / 法务理解。

场景 4：中文场景 pretrain

目标：训练 / 发布面向中文用户的 LLM。受众：中文模型 PM / pretrain 数据组 / 商务对接。

预算：中文评测碎片化，需多个 benchmark 拼接；不要指望单一榜单代表中文能力。

推荐套餐 (4-5)：

| Benchmark | 角色 | Why | |---|---|---| | C-Eval (52 学科 + C-Eval-Hard 8 学科) | 中文学科知识主指标 | 13,948 题，2025-07 后 test label 公开（contamination 自此要 caveat）；C-Eval-Hard 是首个中文 reasoning 子集 | | CMMLU (67 学科, 含 China-specific) | 中文本土制度文化覆盖 | 11,528 题；古汉语 / 公务员 / 中医 / 驾照规则等 China-specific 子领域在英文 MMLU 无法替代 | | AGIEval-CN (Gaokao + JEC-QA + LogiQA-zh) | 中文标准化考试能力 |

3,422 题，覆盖中国高考 8 个学科 + 司法考试 + 中文逻辑；与 C-Eval/CMMLU 三件套互补 | | SuperCLUE (双月全量换题 + GPT-4/Claude judge) | 中文 user preference 主榜 | 1,288 题，双月 100% 换题防过拟合；六维（数学 / 科学 / 代码 / 智能体 / 指令 / 幻觉）覆盖 toC 真实使用 | | MMLU-Pro (英文) | Cross-lingual 验证 | 用英文 MMLU-Pro 验证模型 是否过度偏向中文；如果中文分数比英文高 20+ pp，说明数据 mixture 严重偏中文 |

反向排除 (2)：

- **只跑单一中文 benchmark**——C-Eval-Hard 8 学科 vs CMMLU China-specific 22 学科覆盖完全不重，单跑任一项漏一半能力。GAOKAO-Eval 2024[^gaokao-eval] 实证：即使控 contamination，Gaokao 高分仍不反映人类对齐能力——跨难度成绩异常平整 + 同难度方差异常大。
- **仅 SuperCLUE 不跑 C-Eval/CMMLU**——题库不公开 + 双月换题导致无法跨时间纵向比较；商业偏向也使学术 reviewer 不接受单独引用。

执行 checklist：

- C-Eval / AGIEval-CN 源 Gaokao / 公务员真题，中文 web crawl 中广泛存在；model card 需声明 contamination control 方法（n-gram 不够，需 Soft Contamination 2026[^soft-contam-2026] semantic check）。
- 中文 benchmark frontier 已 90%+（CMMLU 顶部 Lingzhi-72B 90.26）；新 release 不破 90 不要 highlight。
- SuperCLUE 报告同时给当月题库 release date，避免 2024 数字与 2026 数字直接比较。

场景 5：reasoning / agent 方向模型

目标：发布 reasoning 强（o-series / R1 / Claude Sonnet 4 类）或 agent 强（computer-use / MCP / SWE）的模型。受众：frontier release 团队 / 大客户合作团队。

预算：reasoning 评测费 token（thinking 8k-32k / inference），单跑可达 release 总成本 5-15%；需 explicit budget。

推荐套餐 (5-6)：

| Benchmark | 角色 | Why | |---|---|---| | AIME 2026 (via MathArena) | Reasoning math 主指标 | MathArena 平台月更 AIME 2026 / USAMO 2026 / ArxivMath；AIME 2024/2025 已被 saturated (GPT-5 95.7 / Step-3.5 99.9)[^aime-cards] | | FrontierMath (与 Soohak / Riemann-Bench 并报) | Reasoning math ceiling | 必须并报独立流；2026-05 Epoch 公告 ~1/3 题有 fatal errors[^epoch-tier4]，单独引用不可信 | | GPQA-Diamond (子领域分项 + maj@N) | STEM reasoning | 198 题 binomial CI 宽，必须 maj@16 或 maj@32；按 Physics/Chem/Bio 子分项[^epoch-gpqa] | | SWE-Bench Pro (Public + Private) | Repo-level engineering agent | Scale AI 2026 接班 SWE-bench Verified；1,865 任务 + 4.1 文件平均 + Python/Go/TS/JS[^swe-pro] | | MCPMark + MCP-Atlas | MCP tool-use 双axis | MCPMark 16.2-turn × 17.4 tool-call 深度；MCP-Atlas 36 server 广度；二者设计哲学正面冲突（programmatic verify vs claim-based judge），并报覆盖两个 axis | | AgencyBench (long-

horizon) | 1M-token agent ceiling | 138 任务 × 90 tool call × 数小时；ACL 2026[^agencybench] 是当下最长程的 agent benchmark |

反向排除 (2)：

- **legacy reasoning（仅 GSM8K + MATH-500 + BBH）**——GSM8K saturated 92%+，MATH-500 接近 99，BBH frontier ≥ 90 + 标错题已被社区报告[^bbh-issues]；2024 后的 reasoning paper 报这套等同"未跟进文献"。
- **AIME 2024 / 2025 单独使用**——已 saturated（GPT-5 95.7 / Step-3.5 99.9）；只能与 AIME 2026 一起作 contamination diagnostic（diff = train cutoff 后 generalization 信号），不能单独引用为 frontier 数字。

执行 checklist：

- 报告 maj@K + token budget（thinking tokens 数）+ scaffold；裸 pass@1 在 reasoning era 几乎无信息。
- Agent benchmark 用 `release-date + git commit` 双 pin；避开 BenchJack 三连陷阱（GAIA validation / WebArena 原版 / OSWorld 原版）[^berkeley-2026]。
- FrontierMath 必标 "provisional, pending 2026-Q3 reissue"——sponsor 争议 + fatal errors 之后唯一可接受的引用方式。

跨场景速查表

一页 cheat sheet——决策者可问「你属于哪 row？」定位「main column 应该长这样」：

| 场景 | 知识主项 | 推理主项 | 代码主项 | 安全主项 | Anchor 项 | |---|---|---|---|---|---| | Pretrain 中检 | MMLU-STEM subset | GSM8K-light | — | — | LiveBench monthly | | 申请发表 | MMLU-Pro | GPQA-Diamond 分项 | HumanEval+ + LCB v6 | Specialty 领域 | Arena-Hard v2 + MATH-500 | | 公司上线 | Internal | Internal | Internal | HarmBench + AIR-Bench | Live A/B + canary | | 中文 pretrain | C-Eval + CMMLU + AGIEval-CN | C-Eval-Hard | C-Eval CS subject | 自建 | SuperCLUE 双月 | | Reasoning / agent | GPQA-Diamond | AIME 2026 + FrontierMath + GPQA | SWE-Bench Pro | HarmBench + AIR-Bench | MathArena + MCPMark + MCP-Atlas |

写 release card / model card / paper table 时配合第 3 章「迁移决策」表使用。第 5 章和附录给「未来 6–12 月 watch 项」，目前不应进正式套餐但需监控。

引用

[^liu2026probing]: Liu, Z., Lun, T., Wen, Z. et al. (2026). *Fast and Accurate Probing of In-Training LLMs' Downstream Performances*. arXiv:2604.01025.

[^bhi-2026]: Zhu, L., Hua, H., Miao, L., Zhao, B. (2026). *Benchmark Health Index: A Systematic Framework for Benchmarking the Benchmarks of LLMs*. arXiv:2602.11674.

[^epoch-gpqa]: Burnham, G. (2025). *GPQA Diamond: What's Left?* Epoch AI gradient-updates. https://epoch.ai/gradient-updates/gpqa-diamond-whats-left.

[^liu-evalplus]: Liu, J. et al. (2023). *Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation (EvalPlus / HumanEval+)*. arXiv:2305.01210.

[^matharena]: Dekoninck, J. et al. (2026). *Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs*. arXiv:2605.00674.

[^swe-pro]: Scale AI (2026). *SWE-Bench Pro: A Stronger Coding-Agent Benchmark*. https://scale.com/blog/swe-bench-pro. Public leaderboard: https://labs.scale.com/leaderboard/swe_bench_pro_public.

[^mmlu-redux]: Gema, A. P. et al. (2024). *Are We Done with MMLU?* arXiv:2406.04127.

[^harmbench]: Mazeika, M., Phan, L., Yin, X., Zou, A. et al. (2024). *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal*. ICML 2024. arXiv:2402.04249.

[^airbench]: Zeng, Y., Yang, Y., Zhou, A., Tan, J. Z. et al. (2024). *AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies*. NeurIPS 2024 D&B / ICLR 2025. arXiv:2407.17436.

[^agencybench]: Shi, J. et al. (2026). *AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts*. ACL 2026 Main. arXiv:2601.11044.

[^mcpmark]: Wu, Z. et al. (2026). *MCPMark: A Benchmark for Stress-Testing Realistic and Comprehensive MCP Use*. ICLR 2026. arXiv:2509.24002.

[^mcp-atlas]: Bandi, C. et al. (2026). *MCP-Atlas: A Large-Scale Benchmark for Tool-Use Competency with Real MCP Servers*. arXiv:2602.00933.

[^judge-harness-2026]: Dev, S., Sloan, A., Kavner, J., Kong, N., Sandler, M. (2026). *Judge Reliability Harness: Stress Testing the Reliability of LLM Judges*. ICLR 2026 Agents in the Wild Workshop. arXiv:2603.05399.

[^gaokao-eval]: Zhang, Z. et al. (2024). *GAOKAO-Eval: Does high scores truly reflect strong capabilities in LLMs?* arXiv:2412.10056.

[^soft-contam-2026]: Spiesberger, A. et al. (2026). *Soft Contamination Means Benchmarks Test Shallow Generalization*. arXiv:2602.12413.

[^aime-cards]: AIME 2024 / 2025 leaderboard snapshot. https://pricepertoken.com/leaderboards/benchmark/aime.

[^epoch-tier4]: Epoch AI (2026). *FrontierMath Tier 4 benchmark page*. https://epoch.ai/benchmarks/frontiermath-tier-4.

28 / 33

[^bbh-issues]: BIG-Bench Hard Issue #15: incorrect ground-truth labels in `date_understanding` and `geometric_shapes`. https://github.com/suzgunmirac/BIG-Bench-Hard/issues/15.

[^berkeley-2026]: Wang, H., Mang, Q., Cheung, A., Sen, K., Song, D. (2026). *How We Broke Top AI Agent Benchmarks: And What Comes Next*. Berkeley RDI, April 2026. https://rdi.berkeley.edu/blog/trustworthy-benchmarks-cont/.

第 5 章 · 未来 6-12 月值得跟踪的方向

前四章是"截至 2026-05-25 该跑什么"。本章回答 6-12 个月内 pretrain eval 套餐里会被加入哪些新维度——今天就该把它们入 watch list，避免 2027 年 release 时套餐还停在 2025 范式。

七个方向 + 决策表 + 入套餐时机。每个方向都能在 Stage 0.5 frontier scan（2026-02 → 2026-05）的真实新工作里找到锚点。

决策表：七个方向何时入套餐

方向	锚点工作	入套餐时机	优先级
Live benchmark 月更标准化	LiveBench / LiveCodeBench v6 / MathArena Platform[^matharena-2026]	现在并入，把静态 benchmark 降级	P0
过程评估 / PRM 类	ToolPRMBench / CRYSTAL process-level	release reasoning model 时配；base 观望	P2
AI-as-judge 校准	Judge Reliability Harness（RAND）[^judge-reliab-2026]	用 LLM-as-judge 上线前必跑	P1
危险能力 eval 二代	OpenAgentSafety / AgentTrust / AIR-Bench v2 / WMDP-Plus（待）	model 公开 release 前必有	P0 prod / P2 science
Long-horizon agent benchmark	AgencyBench / Odysseys / AstaBench[^astabench-2026]	agent-tuned model 必含；base 观望	P0 agent / P2 base
Pretrain-stage 工具创新	in-training-probing (3min vs 1h)[^liu2026probing] / daVinci-LLM / BHI	trillion-token 项目立刻试装	P1
Contamination 5 路线整合	n-gram / Min-K / CCV / Soft / Watermark / JECS	2026Q4 起 report 应明示采纳了哪几路	P1

下面逐一展开。

5.1 Live benchmark 月更范式取代静态

LiveBench / LiveCodeBench / MathArena Platform 把"benchmark = 月更服务"坐实。MathArena 2026-05 发表平台论文，定义 ArxivMath、AIME 2026、USAMO 2026、ArXivLean、BrokenArxiv 五个并行流为持续基础设施[^matharena-2026]；同月团队发"Farewell to Final-Answer"博文宣告 GPT-5.5 / Gemini 3.1 Pro 已让静态 final-answer 榜失去区分度。

入套餐时机：现在。release reports 必须把"static + live"配对——OpenAI 在 GPT-5.5 卡里同时报 AIME 2025 与 MathArena ArxivMath 就是产业信号。具体：把 LiveCodeBench `--`

`start_date`、MathArena `--comp arxivmath_2026_05`、LiveBench `--livebench-release-option` 钉死写进 eval pipeline config。

5.2 过程评估（PRM 类）

OpenAI 2023 "Let's Verify Step by Step" 之后 PRM 训练数据（PRM800K）成熟，但**评 PRM 本身**的 benchmark 直到 2026 才有：ToolPRMBench（arxiv:2601.12294）做 tool-using agent step-level PRM 评测；CRYSTAL（arxiv:2603.13099）把"transparent multimodal reasoning"中间步骤评分单列，发现 SOTA 普遍"precision exceeds recall"——精确选答但漏真证据。

入套餐时机：reasoning model release 时配；base model 暂可不跑。why：PRM eval 难点不是题，是"判分对不对"——训练 PRM 自身就有研究门槛。watch list 即可，等社区协议标准化再正式入。

5.3 AI-as-judge 校准

RAND 2026-03 Judge Reliability Harness 是当前**唯一今天就能落地**的 judge 校准工具[^judge-reliab-2026]：给一个 judge + 一个 benchmark，自动生成 formatting / paraphrasing / verbosity / label flipping 四类扰动，输出 reliability matrix。作者在 4 SOTA judge × 4 benchmark 上结论："No judge that we evaluated is uniformly reliable across benchmarks"——judge A 在 safety 上稳但 persuasion 对 paraphrase 敏感，judge B 反之。

入套餐时机：**所有用 LLM-as-judge 的评测上线前必跑**。Arena-Hard v2 / MixEval-Hard / MathArena Proof judge / AstaBench——RAND harness 应作为 pre-deployment gate。why：leaderboard 数字是否可信先要回答 judge 是否稳定，工具开源让回答从"凭信仰"变"凭报告"。

5.4 危险能力 eval 二代

WMDP（2024-03）是 hazardous knowledge MCQ 代理，已被 frontier lab 在 unlearning 实验里跑过，但 MCQ 天然 saturate 快，**WMDP-Plus 2026Q3-Q4 几乎肯定发布**。同期更紧迫的是 **agent safety**：OpenAgentSafety（ICLR 2026，350 task）显示 Claude Sonnet 3.7 在 multi-turn tool use 上仍 51.2% 不安全；AgentTrust（arxiv:2605.04785）提供 runtime 拦截层 + 930 scenarios；AIR-Bench v2 在 roadmap 将 EU AI Act 2025 修订纳入 taxonomy。

入套餐时机：production model release 前必跑（含 base 公开 release）；纯 science paper 可暂缓。why：EU AI Act / US EO / 中国管理办法都要求 hazard 评估留痕，2026Q4 起没跑 WMDP-class + agent-safety 双件套很难合规上线。

5.5 Long-horizon agent benchmark

GAIA / SWE-bench Verified / WebArena 这一代已被 frontier model 显著拉高（70%+），区分度下降。三条线接位：AgencyBench（1M-token / ~90 工具调用 / 138 task / ACL 2026

Main，frontier 仅 56.5%）、Odysseys（200 真实长程 multi-site web 任务，frontier 仅 44.5% perfect）、AstaBench（AI2，ICLR 2026 Oral，2400+ 科研 problems）[^astabench-2026]。三者分别覆盖软件开发长程、web browsing 长程、科研发现长程。

入套餐时机：reasoning 或 agent-tuned model release 必含——这是 Anthropic / OpenAI 2026 H2 卡里铁定出现的项目；pure base 可观望。why：base pipeline 加 docker sandbox / live internet / 1M-token context 代价不小，且 scaffold sensitivity（AgencyBench 实测 Claude-4.5-Opus 在 native vs 第三方 SDK 上差 20.5%）需先有稳定协议。

5.6 Pretrain-stage 工具创新

trillion-token 训练 eval 成本核心问题是"每 checkpoint 1 小时 generative eval"。2026-04 in-training-probing（arxiv:2604.01025）给出 lightweight probe 方案：训 small probe 预测 downstream pass@1，**每点评测从 1 小时压到 3 分钟、AUROC > 0.75**[^liu2026probing]。daVinci-LLM（arxiv:2603.27164）给出 Data Darwinism L0-L9 + 200+ ablation 工程模板。BHI（arxiv:2602.11674）从另一端切入："评 benchmark 自身的健康度"——CD / AS / Impact 三轴。

入套餐时机：trillion-token 项目立刻试装 in-training-probing 做高频监控（每 5k step 替代每 100k step 才跑的 full harness）；7B 以下短训练 ROI 有限。daVinci-LLM 与 BHI 是 selection-time 工具，build report 前跑一次即可。

5.7 Contamination 5 大正交路线整合

2026Q1-Q2 是 contamination 方法层"百花齐放"季：n-gram / Min-K%++ / canary-watermark / CCV behavioral / Soft Contamination semantic / JECS conformal joint 首次同季并存。**它们正交**：n-gram 是底线、Min-K% 找已知 sample、watermark 防发布前、CCV 诊断已发布闭源、Soft 揭示语义层盲区、JECS 给多模型统一判定。

入套餐时机：2026Q4 起 pretrain report 应明示采纳哪几路。最低纪律是 n-gram + 训练时间窗口公开 + Min-K% 自查；advanced 团队叠 Soft embedding-level audit + Dataset Watermarking（若发 benchmark）。why：单 n-gram 已被实证不够——Spiesberger et al. 2026 在 Olmo3 训练 corpus 上发现 CodeForces 78% / ZebraLogic 50% 题目存在语义重复全被 n-gram 漏掉[^soft-contam-2026]。"我们做了 13-gram filter 但未做 semantic dedup"是 2026 的诚实表达。

决策 takeaway

七项中 P0（live benchmark / 危险能力 / agent-tuned 模型的 long-horizon）+ P1（judge 校准 / pretrain 工具 / contamination 整合）落地，已覆盖 2026 H2 — 2027 H1 顶层 release card 会被审视的所有角度。P2 PRM 评测视社区标准成熟速度从订阅转正式入。

下一页附录给 Stage 0.5 frontier scan 中具体 benchmark 的"入/观望/忽略"建议——这份 cookbook 不是终点，**六个月后回看，附录会先过时**。

31 / 33

引用条目

[^matharena-2026]: Dekoninck, J., Jovanović, M., Vechev, M., et al. (2026). *Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs*. arXiv:2605.00674.

[^astabench-2026]: Bragg, J., D'Arcy, M., Balepur, N., et al. (2026). *AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite*. ICLR 2026 Oral. arXiv:2510.21652.

[^judge-reliab-2026]: Dev, S., Sloan, A., Kavner, J., Kong, N., Sandler, M. (2026). *Judge Reliability Harness: Stress Testing the Reliability of LLM Judges*. arXiv:2603.05399; repo https://github.com/RANDCorporation/judge-reliability-harness

[^liu2026probing]: Liu, Z., Lun, T., Wen, Z. et al. (2026). *Fast and Accurate Probing of In-Training LLMs' Downstream Performances*. arXiv:2604.01025.

[^soft-contam-2026]: Spiesberger, A., Vazquez, J. J., Pochinkov, N., et al. (2026). *Soft Contamination Means Benchmarks Test Shallow Generalization*. arXiv:2602.12413.

附录 · Frontier Watch — 下一代 pretrain eval 套餐影响

Stage 0.5 frontier scan（2026-02-01 → 2026-05-25）共发现 **73 个新 benchmark / framework / contamination method**。以下表对每条给出"入套餐 / 观望 / 忽略"建议 + 一句理由。判断尺度针对**下一代（2026 H2 — 2027）pretrain LLM eval 套餐**而非 agent product eval。

Deep / Touch 档（≥ 进入正文 / 必读）

| 名字 | arxiv | 建议 | 一句理由 | |---|---|---|---| | AgencyBench | 2601.11044 | 入 | 1M-token + 90 tool call long-horizon agent 唯一代表，agent-tuned model release 必含 | | MCPMark | 2509.24002 | 入 | ICLR 2026 收录 MCP 评测；MCP 已是 2026 部署事实标准 | | MCP-Atlas | 2602.00933 | 入 | 1000 task × 36 server，与 MCPMark 互补（广度 vs 深度） | | AstaBench | 2510.21652 | 入 | ICLR 2026 Oral，科研 agent 唯一 holistic 标杆 | | Odysseys | 2604.24964 | 入 | 真实长程 multi-site web 任务；frontier 仅 44.5%，区分度强 | | SWE-Bench Pro | Scale AI | 入 | SWE-Verified 事实死亡后的官方继任 | | MathArena Platform | 2605.00674 | 入 | 月更 ArxivMath + AIME 2026 + USAMO 2026 + ArXivLean，post-AIME-saturate 主榜 | | LemmaBench | 2602.24173 | 入 | Live + research-level theorem-proving，SOTA ~15%，contamination-free | | Soohak | 2605.09063 | 入 | 64 数学家手 curated；refusal subset 是元认知唯一公开 canary | | in-training-probing | 2604.01025 | 入 | 1h → 3min 加速，trillion-token 项目必试 | | Benchmark Health Index | 2602.11674 | 入（meta-tool） | benchmark selection 之前跑一次过滤 | | Judge Reliability Harness | 2603.05399 | 入

（pre-deploy gate） | 唯一开源 judge 校准；用 LLM-judge 上线前必跑 | | SWE-Skills-Bench | 2603.15401 | 观望 | "agent skills 有用吗"的犀利诊断，但仍是 SWE-bench Verified 延伸 | | SkillsBench (Anthropic Harbor) | 2602.12670 | 观望 | Harbor 框架对外开放；待 Anthropic 正式 commit 长期维护 | | SWE-bench Live | NeurIPS 2025 D&B | 观望 | 月更协议精彩但样本小；用作 SWE-Bench Pro cross-check | | SWE-Rebench | OpenReview 1265 | 观望 | per-model contamination cutoff 设计聪明，但 leaderboard 仍小众 | | Terminal-Bench 2.0 | 2601.11868 | 观望 | agent CLI 工作流眼亮但 pretrain-eval 相关性次要 | | EternalMath | 2601.01400 | 观望 | "living math" 路线 #2，与 LemmaBench / MathArena 重叠待区分 | | Riemann-Bench | 2604.06802 | 观望 | 25 题私有 moonshot math，作 ceiling indicator 而非 daily driver | | OpenAgentSafety | 2507.06134 | 入 | ICLR 2026 收录；agent multi-turn safety 标杆 | | AgentTrust | 2605.04785 | 观望 | runtime intercept 思路新；single-author，等独立复现 | | CCV / HCCA | 2603.21454 | 观望 | behavioral contamination 探针；仅 9-task 验证，repo 未公开 | | Soft Contamination | 2602.12413 | 观望 | framing paper 必读；无 turnkey 工具 | | Dataset Watermarking (closed-LLM) | 2605.06865 | 观望 | dataset owner 才能用；attacker 可二次 paraphrase 干扰 |

Index 档（仅 appendix mention，不入正文）

剩 ~49 个（详见 `benchmark-cards/_INDEX.md`）：垂直行业（EnergyAgentBench / AEC-Bench）、规模偏小（LongCLI-Bench 20 task）、主题重叠（PSPA-Bench、AmbiBench、KDR-Bench、IDRBench）、诊断性质（HORIZON、SkillCraft、SlopCodeBench）——**忽略**：不构成下一代套餐的能力维度，仅作行业案例参考。

终极提醒

这份 cookbook **不是终点**。Stage 0.5 是 4 个月窗口的 73 条新工作；下一个 4 个月（2026-06 → 2026-10）几乎肯定会出现：(a) WMDP-Plus / WMDP-v2；(b) MathArena 与 LemmaBench 的合并标准；(c) Anthropic 或 OpenAI 正式发布的 agent benchmark；(d) `lm-evaluation-harness` / `OLMES` 集成 in-training-probing 与 Soft Contamination 检测。**六个月后回看，本附录会先于正文过时**——重跑一次 Stage 0.5 类的 frontier scan，再调整套餐。

引用

详见正文 §5 [^matharena-2026] [^astabench-2026] [^liu2026probing] [^soft-contam-2026] [^judge-reliab-2026]，本附录依据 Stage 0.5 三份子扫描合并报告 `shared/frontier-scan-stage-0_5*.md`。

报告截稿日期: 2026-05-25。Frontier Watch 附录覆盖截稿后涌现的新工作。

33 / 33