1

FreshBench：

让 AI 评测，保持新鲜。

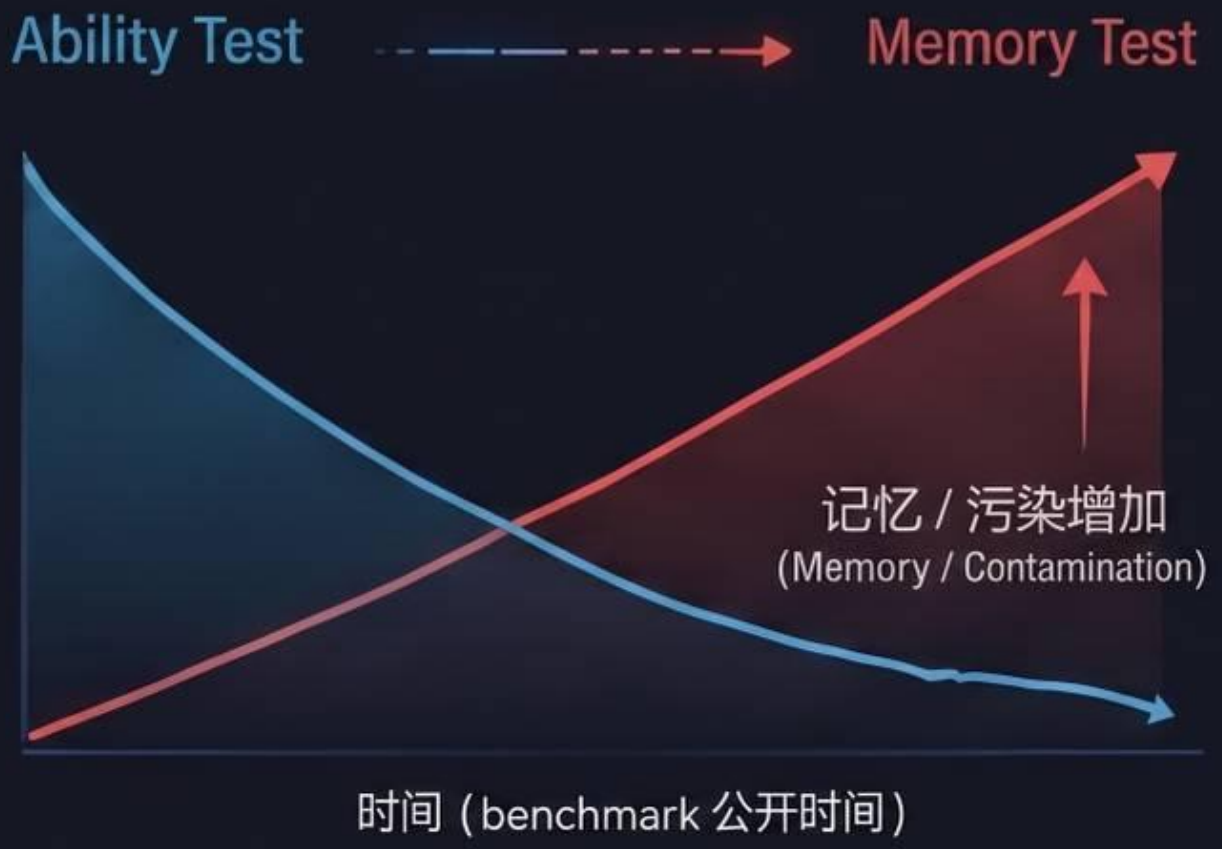
A fresh benchmark
for a smarter future.

2

一个越来越严重、但经常被忽略的问题：

很多 AI benchmark，
已经越来越难真实反映模型能力了。

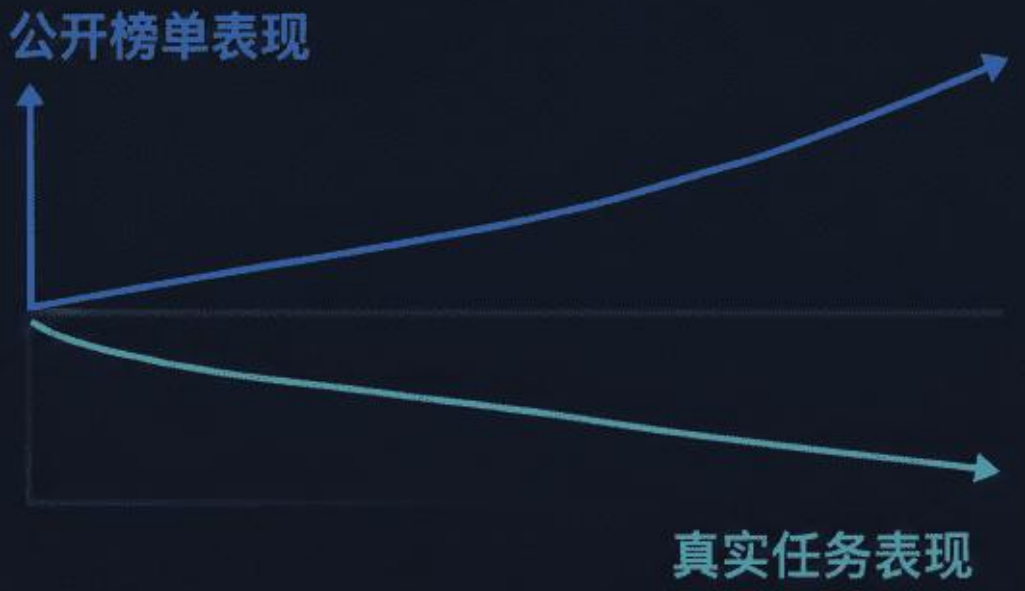
一旦公开太久，就会被训练数据污染，
被模型记住，最后从 ability test
变成 memory test。



3

现象：

模型在公开榜单上
分数越来越高，
但到了真实任务里持续出错



推理
(Reasoning)

工具调用
(Tool Use)

代码修复
(Code Fixing)

文档理解
(Use Understanding)

⚠ 持续出错　⚠ 持续出错　⚠ 持续出错　⚠ 持续出错

4

FreshBench 本质：

让 AI benchmark 持续“保鲜”的 Bittensor subnet 方案。

miner 提交 benchmark assets。

- 题目 (Problem)
- 可验证的答案 (Verifiable Answer)
- 能力标签 (Capability Tags)
- 难度估计 (Difficulty Est.)
- 新鲜度证据 (Freshness Proof)

5

validator：给 benchmark assets 做质量审查。

检查四件事：

1 

Schema Check

2 

Ground Truth Check

3 

Novelty Check

1 

Model Panel Calibration

只有通过这些检查的 benchmark asset，
才值得进入 active pool，成为真正有价值的评测资产。

6

为什么这件事适合用 Bittensor 来做?

开源协作
由全球社区共同完成。

持续供给
跨领域、跨语言、跨场景长期产出。

质量保障
公开竞争 + 可验证机制筛选优质结果。

方向一致
正是 Subnet 擅长的方向。

7

为了把这个想法落地，我们这次黑客松
先做了一个最小可运行的 demo。

在 v1 里，我们选择了

Document Structured Extraction

作为第一类任务，

不是因为 FreshBench 只做文档，
而是因为这个场景最容易清楚展示：

什么叫可验证
(Verifiable)

什么叫新鲜度
(Freshness)

什么叫高质量
benchmark
(High Quality)

8

这就是 FreshBench 想做的事情。

我们不奖励更多的问题，
we reward better benchmark assets.

不是让 AI
更会考试，

而是让 AI 的考试
重新变得可信。

谢谢大家！

Thank you!

FreshBench

Keep AI benchmarks fresh.