Foresight Backtest — 凍結權重模型的預測力遮蔽回測

判決報告 + 設計與限制 · 2026-06-13

Masked Backtest — 解讀（v0 + v1 判決，2026-06-12）

> 數字來源：`results/REPORT.md`（v0）、`results/REPORT_v1.md`（v1）。本檔是人讀的解讀層。
> 設計與限制：`README.md`。原始預測：`results/predictions*.jsonl`（v0 820 筆 + v1 1080 筆全留檔）。

⚖️ v1 判決（平衡 slate + 大模型時間機器 + 統計檢定）

一句話判決

**「凍結模型 + 純推理 → 預測未來」這條路，正式判死——而且死因比 v0 想的更深：

不是校準爛，是『沒有鑑別力』（AUC≈0.50，連排序都做不到）。

同一個實驗反手證明了 harness 本身有效：它當場抓到汙染的指紋。**

v1 設計（補掉 v0 全部缺陷）

- **平衡 slate**：120 題，yes/no 精確 50/50，2025（Manifold 60 題）/ 2026（Kalshi 60 題）各半 → naive 0.25 成為公平門檻
- **大腦時間機器**：加 gpt-oss:120b（cutoff 2024-06，邊界實測通過）
- **統計檢定**：bootstrap CI95 + AUC（rank-ordering 鑑別力）

判決證據（三條獨立證據鏈，全部指向同一結論）

證據 1：全 arm 輸給丟銅板，CI 不重疊（不是雜訊）

arm	Brier	CI95	vs naive 0.25
scout protocol	0.339	[0.285, 0.397]	**LOSES**
gpt-oss 120B protocol	0.383	[0.302, 0.466]	**LOSES**
其餘全部	0.35–0.49	全在 0.25 上方	**LOSES**

證據 2：AUC≈0.50 = 沒有鑑別力（這才是死亡證明）

arm	AUC 全體	AUC 2025	AUC 2026
scout（凍結）	0.514	0.502	0.524
gpt-oss 120B（凍結）	0.516	0.503	0.523
llama3.1（凍結）	0.469	0.500	0.454
modern qwen3.6（知道2025）	0.564	**0.679**	**0.463**
真錢市場 T-30	**0.777**	—	—

> 凍結模型 AUC 全部貼著 0.50 = 擲骰子。校準術救不了「沒有 resolution」的東西。
> 市場 0.78 = 真有料。**差別不是腦力，是即時資訊**。

證據 3：汙染的指紋被當場抓到（harness 自我驗證）

modern 模型的鑑別力**剛好在它的知識邊界蒸發**：2025 題 AUC 0.679（它讀過新聞）→ 2026 題 0.463（對它也是未來）。凍結模型則跨邊界平坦。這個 difference-in-differences 就是汙染的鐵證。

人類可讀的鐵證（2025 題、modern 對 / scout 瞎猜）：

- 「Kash Patel 當 FBI 局長？」→ modern **0.85✓**、scout 0.05（不知道）
- 「Tulsi Gabbard 確認情報總監？」→ modern **0.85✓**、scout 0.20
- modern 不是推理更強，是**記得**。這正是「假裝無知失敗」（Simulated Ignorance Fails）的活體標本。

一個比 v0 更深的發現：偏誤是「題目選樣」不是「結果選樣」

即使 slate 結果已平衡 50/50，凍結模型**仍系統性低估**（scout 說 4% 的事，實際發生 39–43%）。根因升級：**預測市場的題目天生活在「不確定區」**——「有人為它開盤」這個動作本身就篩掉了無聊題。模型套用「一般世界 prior」（戲劇性事件罕見），但市場題是「過濾後的集合」（戲劇性事件約半數成真）。這不是 slate 沒做平衡，是「prediction-market 題」這個物種的本質。

什麼死了、什麼沒死（精確版）

死了（強證據）：小/本地模型 + 無檢索 + 純推理，當「預測未來的 oracle」。連排序都不會，校準也救不回。

沒死，而且被照亮：

1. **市場 AUC 0.78** → 「看見未來」的正解是**資訊取得 + 聚合**，不是把推理關在盒子裡。
2. **汙染偵測器本身 = 驗證過的儀器**。乾淨量化 benchmark 汙染（當紅難題），這是 methods paper 級資產，與 SAVE 同族。

---

v0（原始輪） — 解讀

一句話

**Harness 端到端跑通（$0 API 費），第一輪就抓到一個比「LLM 準不準」更重要的方法學發現：

天真的回測量到的是「題目選樣偏誤」，不是「預測能力」。**

四個真發現

1. 量選樣偏誤 = 回測的頭號陷阱（本輪最大發現）

所有模型 arm 的 Brier（0.34–0.46）全部**輸給丟銅板**（0.25）。但看 calibration 表就懂為什麼：

Scout 說的機率	實際發生頻率
0.05	**0.41**
0.14	**0.55**
0.21	**0.57**
0.32	**0.61**

模型不是亂猜——是**系統性低估 YES**。根因：我們按交易量選市場，而高交易量市場 = 「戲劇性事件真的發生了」的題（slate 的 yes base rate 高達 0.56）。凍結在過去的模型對具體戲劇性事件理性地說「罕見」，在這個**非隨機 slate** 上就被屠殺。

含義：任何「LLM 回測打敗市場」的宣稱，若 slate 是熱門市場選出來的，先懷疑選樣。v1 必須分層抽樣（平衡 yes/no、結算年、類別）。

2. 「換最新模型」不是預測解方

現代模型（知道 2025）反而墊底：27B = 0.410、35B = 0.456，都輸給凍結在 2024-08 的 scout（0.350）。

原因：slate 109/110 題在 2026 結算——對 qwen3.6 而言也是未來。它沒有資訊優勢，而它「更理性的保守」在這個 yes-heavy slate 上傷更重。

（汙染對比本身**不確定**：2025 年結算的題只有 1 題進 slate——v0 的選樣缺陷，v1 修。）

3. 虛構自信率被量化了

Probe 階段的副產品——「宣稱知道結局」的傾向：

模型	宣稱 RESOLVED	其中宣稱發生但實際沒發生
llama3.1-8B (2023-12)	**66/117 (56%)**	11
llama4:scout (2024-08)	16/117 (14%)	1

世代/規模差在「知道自己不知道」上是 4 倍級距。這對「用 LLM 做判斷」的所有應用都是直接可用的數字。

4. Probe 真正攔到的是「同名事件混淆」

被剔除的「汙染」多數不是洩漏——是模型拿 cutoff 前的舊事件答新題（Pam Bondi 的佛州檢察長 vs 美國司法部長、FISA 702 的 2024 reauth vs 2026 reauth）。Probe 意外成為**歧義偵測器**。三層敏感度分析（strict/loose/none）下結論穩定，剔除規則不翻盤。

基準的位置

市場 T-30 Brier 0.199、T-7 0.140——資訊不對稱下市場仍是壓倒性 baseline（預期內）。protocol 腿（0.340）小贏直答（0.350），議事會腿樣本太小（n=19）無結論。

v1 該修的（按優先序）

1. **分層抽樣**：yes/no 平衡 + 結算年平衡（讓汙染對比有 2025 樣本）+ 類別配額
2. **更大的時間機器**：本地還有 gpt-oss:120b、nemotron-3-super:120b 可當凍結模型（cutoff 待驗證邊界）
3. **skill score**：用「informed base rate」當分母，而不是只比 raw Brier
4. **MiroFish-Offline 真模擬腿**（Neo4j + 本地舊模型），取代 sim-lite
5. 市場開盤價 baseline（資訊對等性更好的比較點）

成本帳

全程本地推理（M5 Max）：820+ 筆模型輸出、$0 API 費、約 1.5 小時運算。

Harness 可重複使用：換 slate / 換模型 / 換腿都是改 config 等級的事。

來源檔：~/Desktop/02-研究/foresight-backtest/FINDINGS.md

Masked Backtest_v1 — Results

generated 2026-06-12T18:49:53.220187 · 120 events · primary rule: strict exclusion

Arms vs naive baseline (verdict bar = p=0.5 on this slate)

Arm	N	Brier ↓	CI95	vs naive	Market T-30	Market T-7	p=0.5
GPT-OSS-120B direct (cutoff 2024-06)	35	**0.4887**	[0.361,0.603]	LOSES to naive	0.1488	0.0974	0.2500
GPT-OSS-120B protocol	91	**0.3827**	[0.302,0.466]	LOSES to naive	0.2217	0.1218	0.2500
Scout direct (cutoff 2024-08)	113	**0.3464**	[0.287,0.409]	LOSES to naive	0.2098	0.1058	0.2500
Scout protocol	113	**0.3392**	[0.285,0.397]	LOSES to naive	0.2098	0.1058	0.2500
Llama3.1 direct (cutoff 2023-12)	75	**0.3810**	[0.313,0.452]	LOSES to naive	0.2125	0.0862	0.2500
Qwen3.6-35B direct (modern)	120	**0.3876**	[0.317,0.459]	LOSES to naive	0.2101	0.1171	0.2500

Global: market T-30 0.2101 · T-7 0.1171 · 0.5-const 0.2500 · base rate yes 0.50

Sensitivity to exclusion rule

rule	arm	N	Brier
strict	gptoss/A	35	0.4887
strict	gptoss/B	91	0.3827
strict	scout/A	113	0.3464
strict	scout/B	113	0.3392
strict	llama31/A	75	0.3810
strict	modern35/A	120	0.3876
loose	gptoss/A	33	0.5183
loose	gptoss/B	88	0.3948
loose	scout/A	110	0.3550
loose	scout/B	110	0.3474
loose	llama31/A	60	0.4463
loose	modern35/A	120	0.3876
none	gptoss/A	36	0.4751
none	gptoss/B	92	0.3786
none	scout/A	120	0.3658
none	scout/B	120	0.3437
none	llama31/A	90	0.3533
none	modern35/A	120	0.3876

Probe / confabulation

model	probed	claimed RESOLVED	false YES claims	hard flags (excluded)
scout	120	21	6	7
llama31	120	82	16	18

Contamination contrast (modern vs frozen, direct prompt)

slice	N	Scout (frozen)	Qwen3.6 (modern)	Market T-30
close_2025: scout vs modern35	59	0.3673	0.2961	0.2028
close_2026: scout vs modern35	54	0.3236	0.4464	0.2189
close_2025: gptoss vs modern35	19	0.4815	0.2621	0.1136
close_2026: gptoss vs modern35	16	0.4972	0.6048	0.1944

Calibration — scout_B

bin	n	mean p	actual freq
0.0	31	0.04	0.39
0.1	7	0.13	0.57
0.2	20	0.21	0.60
0.3	23	0.32	0.43
0.4	3	0.42	1.00
0.5	9	0.55	0.56
0.6	14	0.60	0.36
0.7	3	0.70	0.33
0.8	2	0.82	0.50
0.9	1	0.98	0.00

Calibration — scout_A

bin	n	mean p	actual freq
0.0	44	0.04	0.43
0.1	4	0.15	1.00
0.2	10	0.20	0.40
0.3	35	0.31	0.49
0.4	2	0.40	0.50
0.5	1	0.55	1.00
0.6	9	0.60	0.44
0.7	4	0.70	0.25
0.8	3	0.80	0.67
0.9	1	0.99	0.00

Calibration — gptoss_B

bin	n	mean p	actual freq
0.0	32	0.04	0.59
0.1	15	0.14	0.47
0.2	5	0.21	0.20
0.3	6	0.31	0.33
0.4	3	0.43	1.00
0.5	5	0.52	0.40
0.6	4	0.61	0.25
0.7	6	0.71	0.83
0.8	3	0.82	0.67
0.9	12	0.96	0.58

Calibration — gptoss_A

bin	n	mean p	actual freq
0.0	18	0.04	0.56
0.1	4	0.16	1.00
0.2	3	0.24	0.67
0.3	5	0.35	0.60
0.4	1	0.42	0.00
0.5	1	0.58	1.00
0.6	2	0.67	0.00
0.9	1	0.96	0.00

來源檔：~/Desktop/02-研究/foresight-backtest/results/REPORT_v1.md

Foresight Backtest v0 — 遮蔽回測（true-ignorance time machine）

> 2026-06-12 啟動。目的：在「LLM 可能已見過答案」的汙染問題下，嚴格地回測 LLM 預測能力。
> 方法：用 cutoff 凍結在過去的本地模型，預測「它們物理上不可能見過、但已結算」的真實事件，

設計

元件	內容
事件源	Kalshi 已結算二元市場（CFTC 監管真錢市場），close 2025-02 ~ 2026-04，lifetime ≥14d，量排序
排除	bio/health、cyber、sports、crypto/資產價格、weather（關鍵字 + category 雙層過濾）
時間機器	llama4:scout（cutoff 2024-08，邊界已驗證：知 CrowdStrike 2024-07、不知 2024-11 大選）+ llama3.1:8b（cutoff 2023-12）
汙染對照組	qwen3.6:27b（modern，知道 2025）同題直答 → 量化 contamination 效應
Probe	每事件先問舊模型「這題你知道結果嗎」，答對結局者剔除（belt-and-suspenders）
三條腿	A 直答 / B superforecaster protocol（reference class → base rate → 證據兩面）/ C sim-lite 議事會（統計家+因果分析師+魔鬼代言人+主席，top-25 子集）
基準	市場 T-30d 賠率、T-7d 賠率、constant 0.5、base rate
指標	Brier（同事件集公平比較）+ calibration bins

已知限制（v0 誠實揭露）

- 模型資訊凍結在 cutoff，市場 T-30 賠率有 cutoff 後資訊 → 模型 vs 市場是「不對稱但誠實」的比較（量化推理能補多少資訊差）
- Kalshi 題目偏美國；題文寫於開市時，少數可能含 cutoff 後實體（合法：預測者拿到新問題的常態）
- sim-lite ≠ MiroFish 全模擬（v1 接 MiroFish-Offline + Neo4j + 本地舊模型）
- Polymarket 從此網路不可達（geo-block），故用 Kalshi

跑法

```
python3 scripts/fetch_markets.py   # 抓資料 → data/events_selected.json
./scripts/run_all.sh               # probe→A→B→C→score，idempotent 可重跑續傳
python3 scripts/score.py           # 單獨重算 → results/REPORT.md
```

進度：`tail -f logs/run.log`；結果：`results/REPORT.md`、`results/results.json`、原始 `results/predictions.jsonl`。

來源檔：~/Desktop/02-研究/foresight-backtest/README.md