동시호가 방향예측에서 상수 기준선을 넘지 못한 두 컨셉:
국채선물 15:45 종가 방향, 2016–2026

채권 퀀트 데스크
실험 `0001_contest_bar_regression` · `0002_contest_tick_direct`

Abstract

3년 국채선물 15:45 동시호가 종가의 방향을 15:34에 제출하는 과제에서, 독립적으로 설계한 두 예측 컨셉이 **모두 사전등록한 상수 기준선을 넘지 못했다**. 이 과제가 어려운 이유는 제출 컷오프가 채점 기준가를 가리기 때문이다 — 제출 뒤 1분의 움직임이 유효일의 16.41%에서 정답 부호를 뒤집는다. 컨셉 1은 1분봉 9년으로 이동폭을 회귀한 뒤 관측된 드리프트를 빼서 그 항을 제거하고, 컨셉 2는 체결틱 2년으로 기준가를 거의 관측한 상태(컷오프에서 0틱 일치 75.03%)에서 방향을 직접 분류한다. 연도 확장창 워크포워드와 공통 시험일 388일에 대한 짝지은 검정[1]으로 판정했다. 가장 두드러진 결과는 이것이다: 학습창의 다수 클래스로 미리 고정한 상수 예측("매일 하락")이 **61.08%**로 학습된 모든 갈래(최고 60.05%)를 이겼고, 컨셉 2가 컷오프 불확실성을 2.31배 줄였는데도 그것이 승률로 옮겨오지 않았다. 이 과제에서 방향 예측기는 비용 축이 없는 채점 게임이므로 아래 수치는 전부 **비용 차감 이전이 아니라 비용 개념이 없는** 승률이다.

1 서론

한국거래소 3년 국채선물은 15:45 동시호가로 종가가 결정된다. 이 논문이 다루는 과제는 매일 **15:34까지** 그 종가가 기준가보다 오를지 내릴지를 제출하는 것이다. 순위는 승률이고, 미제출은 오답이므로 커버리지 100%가 강제된다.

과제의 핵심 난점은 제출 시각과 기준가 확정 시각이 어긋난다는 데 있다. 채점 기준가 P_{ref}는 15:35 직전 마지막 체결가인데 제출은 그보다 앞선다. 즉 제출자는 정답의 기준점을 모르는 채로 방향을 낸다. 우리는 이 공백의 크기를 정식으로 재고(§3), 그것을 각자 다른 방식으로 걷어내는 두 컨셉을 설계해 판정했다.

기여.

1. **컷오프 상한의 정식 측정.** 정식 근월물 규칙으로 재측정하면 15:34 기준 방향과 실제 정답 부호가 갈리는 날이 유효일의 **16.41%**이고, 정답이 0틱이라 무효인 날이 **22.47%**다. 15:34만 보는 예측기는 완벽해도 상한이 83.6%다(§3).
2. **두 컨셉 모두 사전등록 상수 기준선에 미달.** 공통 시험일 388일에서 상수 61.08% > 컨셉 1 60.05% > 컨셉 2 54.64%이고, 짝지은 검정에서 세 갈래 전부 "상수만 맞힌 날"이 더 많다(§5.3).
3. **줄인 불확실성이 승률로 옮겨오지 않는다.** 컨셉 2는 같은 날 집합에서 컷오프 잔차를 57.73%→24.97%로 2.31배 줄였는데도 승률은 오히려 낮다(§5.2).
4. **방법론: 사전등록 문턱 셋이 전부 성립하지 않았다.** 우리가 실험 전에 박아 둔 판정 문턱 세 개가 각각 통과 불가능·항상 통과·정의상 불가능한 형태였다(§6.2). 사전등록은 값을 미리 정하는 것만이 아니라 *그 문턱이 통과 가능한지*를 미리 확인하는 것까지다.

Table 1: 15:34 컷오프의 구조적 상한. 대상일 2,381(결측 봉 제외) 기준.

항목	값
0틱 무효 (정답이 0틱)	22.47%
유효일	1,846
$\mathrm{sign}(P_{45}-P_{34}) \neq \mathrm{sign}(P_{45}-P_{\mathrm{ref}})$	**16.41%**
15:34 기준 0틱이라 방향을 못 내는 날	258
부호가 실제로 반대인 날	45
제출 1분 드리프트 $\lvert d\rvert \geq 1$틱	56.8%
정답 $\lvert P_{45}-P_{\mathrm{ref}}\rvert$가 정확히 1틱	47.1%

2 데이터와 가정

기호. 전 논문에서 고정한다. P_{34}는 컷오프 시점 가격(15:34 라벨 봉의 현재가), P_{ref}는 대회 기준가(15:35 라벨 봉의 현재가 = 15:35 직전 마지막 체결가), P_{45}는 대회 종가(15:45 동시호가 체결가)다. 틱 크기는 0.01포인트이고 모든 가격 차는 틱 정수로 다룬다. 제출 1분 드리프트 $d = (P_{\mathrm{ref}} - P_{34})/0.01$, 회귀 타깃 $y = (P_{45} - P_{34})/0.01$, 정답 부호는 $\mathrm{sign}(P_{45} - P_{\mathrm{ref}}) = \mathrm{sign}(y - d)$다.

원천. 1분봉은 인포맥스 단말 3년 국채선물 연도별 CSV (2016-12-14∼2026-09-07, 대상일 2,385)이고, 체결틱은 mficc 덤프 (2023-01-02∼2026-08-13, 5,654,218행, 873일)다. 체결틱 표에는 호가가 없어 5단계 호가 불균형·OFI 계열 피처가 원리적으로 불가능하고, 방향 판정은 Lee–Ready가 아니라 틱규칙(직전 체결가 비교)만 가능하다.

표본과 제외. 두 실험 모두 라벨을 1분봉 하나에서 만든다(정본 모듈 `contest_labels`). 정답이 0틱인 날은 방향이 없으므로 학습·판정에서 제외하되 건수를 항상 보고한다 — 컨셉 1은 2,385일 중 535일(22.43%), 컨셉 2는 899일 중 171일(19.02%)이다. 두 실험의 시험 표본이 388일에서 만나며, 차이 하루 (2025-11-10)는 그날 체결틱 표가 통째로 비어 있기 때문이다.

> **비표준 세션 가드.** 한국거래소는 수능일에 장을 1시간 늦게 열고 **16:45** 동시호가로 마감한다. 라벨 모듈이 15:34/15:35/15:45를 상수로 쓰고 있었으므로 그날은 "한 시간 전 아무 10분 구간의 등락"이 정답으로 나왔고, 봉이 존재하니 결측 표시도 붙지 않았다. 1분봉 전수를 훑어 근월물 마지막 봉이 16:45인 날 **10일**(2017-11-16, 2017-11-23, 2018-11-15, 2019-11-14, 2020-12-03, 2021-11-18, 2022-11-17, 2023-11-16, 2024-11-14, 2025-11-13)을 찾아 `NONSTANDARD_SESSION`으로 표본에서 제외하는 가드를 넣었다. **아래 표의 런은 이 가드 이전 판이다** — 다시 돌리면 시험 표본이 컨셉 1에서 4일, 컨셉 2에서 2일 줄고 학습 행도 달라진다. 빠지는 날이 전부 정답/전부 오답이었다는 양극단 가정에서 풀링 승률의 상한은 컨셉 1 $-0.15 \sim +0.19$%p, 컨셉 2 $-0.24 \sim +0.28$%p이고 어떤 결론도 뒤집지 않는다.

비용 가정. 이 과제는 방향 채점 게임이라 비용·한도 축이 없다. 실험 `config.yaml`의 `assumptions·limits`는 비어 있고 이 논문의 모든 승률은 비용 개념이 적용되지 않은 값이다. 라이브 전환 시에는 데스크 파라미터가 확정된 뒤 다시 계산해야 한다.

3 컷오프가 만드는 상한

정식 근월물 규칙(거래개시일·최종거래일 표 기반)으로 재측정한 결과가 표 1이다.

뒤집힘은 연도별 12.3%∼20.7%로 9년 내내 빠짐없이 나타난다 — 특정 레짐의 현상이 아니라 구조다. 따라서 P_{45}를 P_{34} 대비로만 예측하는 기계는 완벽해도 상한이 83.6%다. 두 컨셉은 각자 다른 방식으로 이 항을 걷어낸다.

Table 2: P_{obs} 대 봉 기준가 P_{ref} 일치율(873일). 경계마다 답하는 질문이 다르다.

경계	0틱 일치율	$\lvert\text{diff}\rvert \le 1$틱	max $\lvert\text{diff}\rvert$
무필터(그날 마지막 행 — 동시호가에 오염)	19.36%	53.2%	18틱
≤15:34:58 (모델이 제출 시점에 보는 값)	**75.03%**	99.8%	2틱
<15:35:00 (진짜 기준가 경계)	**100.00%**	100%	**0틱**

4 방법

4.1 컨셉 1 — 봉 회귀 + 관측 드리프트 보정

$P_{45} - P_{\text{ref}} = (P_{45} - P_{34}) - (P_{\text{ref}} - P_{34})$에서 앞항은 모델이 회귀하고 뒷항은 백테스트에서 봉 둘의 차로 *정확히 관측*된다. 그래서 틱 데이터 없이도 드리프트 보정의 값어치를 전부 잴 수 있다.

피처는 종가 축만 쓴다: 결정시각 5개(14:00, 15:00, 15:20, 15:30, 15:34) 각각에서 ret_5, ret_{10}, ret_{20}, ret_{60}, rng_{20}, pos_{20} 여섯 개다. d는 피처가 아니라 관측값이며 모델이 학습하지 않는다. 모델은 LightGBM 4.7.0 회귀 (`max_depth`=4, `num_leaves`=15, `min_data_in_leaf`=50, `learning_rate`=0.03, `n_estimators`=400, `feature_fraction`=0.8)이고, 연도 확장창 워크포워드로 2020~2025를 시험한다. 하이퍼파라미터는 실행 전에 `config.yaml`에 고정했고 시험 연도를 보고 다시 고르지 않았다.

> **롤 겹침일 오염.** 그날 근월 계약을 고정하지 않고 일자로만 묶으면 롤 겹침일(전체의 30.8%, 연도별 12.8%~72.0%)에 두 계약의 봉이 섞인다. 오염률이 연도마다 크게 다르므로 하필 연도별 비교를 정확히 망가뜨린다. 피처 빌더는 라벨 표의 계약 코드로 먼저 거르고, 코드가 둘 이상인데 라벨이 없으면 조용히 하나를 고르지 않고 예외를 올린다.

4.2 컨셉 2 — 관측 기준가 + 틱 직접 분류

컷오프를 **15:34:58**(제출 마감 직전 마지막 초)로 두고, 그 시점까지의 체결틱 집계 12열($P_{\text{obs}}^{34:58}$, 체결 건수·수량·대량비, 틱규칙 부호 평균, 체결 간격, 실현변동성, 미결제약정 변화, 1/5/20분 수익률)로 방향을 직접 분류한다. 모델은 LightGBM 4.7.0 분류(`max_depth`=3, `num_leaves`=7, `min_data_in_leaf`=30, `learning_rate`=0.03, `n_estimators`=300), 시드 5개, 시험 연도 2024·2025다.

> **누설 경로.** 체결틱 표 873일 *전부*에 컷오프 이후 행이 있고, 871일에 거래량이 큰 `15:45:00` 프린트가 있다. 그 프린트의 가격은 1분봉 근월 15:45봉 현재가(= P_{45})와 **0틱 일치 99.89%**(870/871) — 진짜 동시호가 체결이다. 피처가 이 행을 흡수하면 정답을 그대로 보는 것이므로, 피처 경로는 반드시 컷오프 가드를 통과하게 하고 컷오프 이후 4,501행 전부를 오염시켜도 피처가 한 셀도 변하지 않음을 검사로 확인했다.

기준가를 얼마나 아는가. 세 경계의 의미가 서로 다르다(표 2). `<15:35:00` 경계에서 873일 전부가 봉 기준가와 0틱 일치한다는 사실은 틱 표가 1분봉 근월 계열과 같은 계약임을 확정한다(원월 대조는 3.94%). 그러나 모델이 제출 시점에 실제로 보는 값은 ≤`15:34:58` 경계의 75.03%다 — 컨셉 2는 기준가를 *아는* 것이 아니라 *거의* 아는 것이고, 컷오프 문제를 없앤 것이 아니라 줄였다.

4.3 기준선과 검정

- **규칙** — 20분 모멘텀 부호. 무신호일에 방향을 내지 않으므로 커버리지가 100%가 아니다(컨셉 1 시험표본에서 80.6%).
- **상수(사전등록)** — *학습창*의 다수 클래스로 방향을 미리 고정한 뒤 시험한다. 컨셉 2의 학습창(2023)은 상승 48.48%이므로 방향은 "하락"이고, 이 값을 실행 전에 `config.yaml`에 박고 노트북이 학습창에서 재계산해 일치를 단언한다.

Table 3: 컨셉 1 워크포워드. 시험 2020–2025, 결정시각 15:34. 위 네 갈래가 사전등록 판정이고 아래 셋은 표본·커버리지를 맞춘 사후 비교다.

갈래	n	승률	이항 p(단측)	누적틱
(a) 무작위 50%	1,175	51.06%	0.242	3
(b) 규칙 20분 모멘텀	947	**58.61%**	6.6e-08	390
(c) sign($\hat{y}$) 보정 없음	1,175	56.77%	2.0e-06	315
(d) sign($\hat{y}-d$) 보정	1,175	56.51%	4.5e-06	277
사후 — 표본·커버리지를 맞춘 탐색적 비교				
(c) 동일표본 947일	947	57.02%	8.7e-06	264
(d) 동일표본 947일	947	56.49%	3.6e-05	214
혼합: 규칙 우선 + 무신호 228일만 모델	1,175	58.21%	9.9e-09	453

Table 4: 컨셉 2 워크포워드. 시험 2024–2025, $n = 388$, 다수 클래스 비율 61.08%. p는 둘 다 단측이며, 다수 클래스 대비 p는 시험표본에서 추정한 귀무를 쓰므로 상수 행은 자기참조다(그래서 §5.3의 짝지은 검정으로 판정한다).

갈래	n	승률	p vs 50%	p vs 다수	누적틱
무작위 50%	388	50.26%	0.480	1.000	54
상수 하락(사전등록)	388	**61.08%**	7.4e-06	0.522	172
규칙 20분 모멘텀	299	56.52%	0.014	0.953	49
틱 직접 분류(본 갈래)	388	**54.64%**	0.038	0.996	78

- **검정** — 이항검정(단측, 귀무 0.5)과 함께, 같은 날 짝지은 McNemar 검정[1]을 낸다. 후자는 `scipy`의 `binomtest(`b`,` $b+c$`, 0.5)` 양측 정확검정으로 구현했다. 시험 구간이 두 해뿐이므로 **어디에서도 "유의하다"고 쓰지 않는다.**

5 결과

5.1 컨셉 1

표 3이 사전등록 판정이다. 드리프트 보정 sign($\hat{y}-d$)가 보정 없는 sign($\hat{y}$)도, 고정 규칙도 넘지 못했다 — 가설은 풀링과 동일표본 두 비교 모두에서 반증됐다.

왜 보정이 손해인가. 두 후보 설명을 숫자로 갈랐다. 워크포워드 시험표본의 잔차 $\hat{y}-y$ 표준편차는 **2.99틱**인데 보정하려는 $|d|$는 평균 0.768틱, 중앙값 1.0틱이다. 즉 모델 오차가 보정량의 세 배 가까이 크고, 거기서 1틱짜리 양을 빼는 조작은 경계선 근처 부호를 뒤집는 잡음으로 작용한다. 두 번째 후보("모델이 d를 암묵적으로 학습해 이중 차감")는 기각된다 — corr($\hat{y}, d$) = 0.034이고 ret_* 대 d 상관도 전부 0.02∼0.07이다.

확률과 확신도. 학습창 잔차로 σ를 추정해 $P(L) = \Phi((\hat{y}-d)/\hat{\sigma})$를 만들었다($\hat{\sigma}_Y$는 연도 $Y-1$의 잔차 표준편차이고 그 모델은 $Y-1$ 이전만 학습한다 — 시험 연도를 보지 않는다). 결과는 음성이다: $P(L)$의 로그손실 0.7305, Brier 0.2604로 **상수 0.5 기준선(0.6931, 0.25)보다 나쁘다**. 보정 없는 $\Phi(\hat{y}/\hat{\sigma})$만 0.6917, 0.2490으로 근소하게 넘는다. 확신 상위 30% 승률은 59.09%로 전체 56.51%보다 높지만 대회 벤치마크가 보고한 폭(56.7%→62.2%)의 절반 이하다. 혼합의 모델 폴백 구간(228일)에서 확신 3분위별 승률은 하 59.21% / 중 48.68% / 상 61.84%이고 확신도와 정오의 스피어만 상관은 $\rho = 0.061$, $p = 0.363$ — **확신도는 정보를 갖지 않는다.** 부수 산출로 잰 0틱 분류 가능성도 음성이다: $|\hat{y}-d|$로 0틱일을 가려낸 AUC가 0.523이고, 최소 십분위의 0틱 비율이 17.0%로 기저율 20.2%보다 오히려 *낮다*(리프트 0.84).

5.2 컨셉 2

표 4가 판정이다. 시드 5개의 승률 표준편차는 정확히 0이었다 — 지정한 하이퍼파라미터에 확률적 요소(`bagging·feature_fraction`)가 없어 LightGBM이 결정적이기 때문이다.

Table 5: 교집합 388일 최종 비교. b는 그 갈래만 맞힌 날, c는 상수만 맞힌 날. McNemar p는 양측이다.

갈래	승률	상수 대비	b	c	McNemar p
상수 하락(사전등록)	**61.08%**	—	—	—	—
컨셉 1 $\text{sign}(\hat{y}-d)$	60.05%	−1.03%p	80	84	0.815
컨셉 1 $\text{sign}(\hat{y})$	56.70%	−4.38%p	56	73	0.159
컨셉 2 틱 직접 분류	54.64%	−6.44%p	58	83	0.043

Table 6: 연도별 상승 비율(컨셉 1 유효일 기준). 다수 클래스 방향이 2022년에 뒤집힌다.

연도	2020	2021	2022	2023	2024	2025
n	183	195	210	198	197	192
상승 비율	40.4%	44.1%	**51.4%**	48.5%	39.6%	38.5%

H4는 미지지이고 방향이 가설과 반대다. 컨셉 2는 컨셉 1보다 낮다. 그런데 설계 전제 자체는 검증됐다 — 같은 날 집합에서 컷오프 잔차가 0이 아닌 날의 비율이 컨셉 1의 57.73%에서 컨셉 2의 24.97%로 **2.31배** 줄었다. 문제를 줄이는 데는 성공했으나 그것이 승률로 옮겨오지 않았다.

5.3 두 컨셉을 같은 표에

교집합 388일에서 상수 기준선을 다시 계산하면 상승 38.92%, 즉 "매일 하락"이 61.08%다. 표 5가 최종 판정이며, 짝지은 검정에서 **세 갈래 전부 "상수만 맞힌 날"(c)이 "갈래만 맞힌 날"(b)보다 많다** — 이 결론은 추정된 귀무가설에 의존하지 않는다.

앙상블 재료가 없다. 두 컨셉이 갈리는 날은 147일(37.9%)이고 그 날들에서 컨셉 1이 57.14%, 컨셉 2가 42.86%다. 그런데 같은 날들에서 **상수가 63.27%**로 일치일(59.75%)보다도 잘 맞는다 — 갈리는 날은 정보가 있는 날이 아니라 상수가 유난히 잘 맞는 날이다. 사후에 구성한 최선의 혼합(둘이 같으면 그 방향, 갈리면 상수)조차 62.37%로 상수와 b=42 대 c=37, $p = 0.653$이다 — 재료라 부를 수 없다.

> **원래의 앙상블 문턱은 성립할 수 없었다.** 계획서는 "불일치일에 각자 50%를 넘으면 재료가 있다"를 판정 문턱으로 적었는데, 두 갈래가 ±1이고 불일치일에는 정의상 예측이 서로 반대이므로 *두 승률의 합은 정확히 1*이다(실측 57.14+42.86=100.0). 어떤 데이터에서도 통과할 수 없는 문턱이다. 여기서 답을 준 것은 원래 문턱이 아니라 상수 기준선 대비라는 대체 문턱이다.

6 논의

6.1 왜 둘 다 상수를 못 넘었는가

세 가지 관측이 하나의 해석으로 모인다. 첫째, 모델의 예측오차(2.99틱)가 걷어내려는 양(1틱 안팎)보다 훨씬 크다. 둘째, 확률로 바꿔도 상수 0.5보다 나쁘고 확신도가 정오와 무상관이다 — 모델이 "언제 자신 있는가"조차 모른다. 셋째, 컷오프 잔차를 2.31배 줄여도 승률이 오르지 않았다. 즉 **줄인 불확실성이 예측 가능한 부분이 아니었다.** 15:34→15:35 구간을 관측으로 대체하는 것은 정보를 더하는 일이 아니라 이미 무작위인 항을 다른 무작위 항으로 바꾸는 일에 가까웠다.

한편 상수가 강한 이유는 기저율이 50%에서 멀기 때문이다. 시험 구간의 상승 비율은 38.92%다. 그러나 이 값 자체가 연도마다 크게 흔들린다(표 6) — 2022년은 상승이 다수였다. 따라서 "매일 하락"을 고정 기준선으로 삼는 전략은 그런 해에 통째로 진다. **상수를 이기는 가장 그럴듯한 길은 상수를 무시하는 것이 아니라 상수의 방향을 국면에 따라 바꾸는 것**이고, 그러려면 기저율 자체의 예측 가능성을 먼저 재야 한다. 이 논문은 그것을 재지 않았다.

Table 7: 사전에 박은 판정 문턱 셋과 그 실패 방식.

문턱	실패 방식	왜
틱 표와 봉의 미결제약정 **완전일치** 95%	통과 불가	서로 다른 피드의 시점 스냅샷이라 정확히 같을 수 없다
컨셉 간 차이 > **시드 분산의 2배**	항상 통과	하이퍼파라미터에 확률적 요소가 없어 시드 분산이 0, 문턱도 0
불일치일에 **각자 50% 초과**	정의상 불가능	예측이 서로 반대라 두 승률의 합이 정확히 1

6.2 사전등록 문턱 셋이 전부 성립하지 않았다

이 연구의 방법론적 산출물은 여기에 있다. 우리는 실험 전에 판정 문턱 세 개를 박았고 **셋 다 성립하지 않는 형태였다**(표 7). 세 번 모두 구현자가 문턱에 걸려 멈췄고, 세 번 모두 문턱 쪽이 틀린 것이었다. 특히 첫 번째는 위험했다 — 임계 미달을 이유로 컨셉 2를 폐기했다면 실제로는 측정 코드의 정렬 버그 때문에 실험을 접는 셈이었다(그 버그를 고치자 일치율이 83.05%에서 100%로 수렴했다).

> **교훈.** 사전등록은 값을 미리 정하는 것만이 아니라 *그 문턱이 통과 가능한지*를 미리 확인하는 것까지다. 재사용 가능한 점검은 간단하다 — 문턱을 쓸 때 **통과하는 결과와 통과하지 못하는 결과를 각각 하나씩 실제로 적어 보라.** 둘 중 하나를 못 적으면 그 문턱은 판정을 하지 못한다.

7 한계

- **컨셉 2의 시험 구간은 두 해뿐이다.** 체결틱 덤프가 2023년부터라 그렇다. 이 논문은 어디에서도 "유의하다"고 쓰지 않으며, 표의 p는 방향의 참고값으로만 읽어야 한다.
- **상수 기준선 대비 p는 추정된 귀무를 쓴다.** 다수 클래스 비율을 시험표본에서 추정했으므로 상수 행의 p는 자기참조다. 이 약점을 공시하고 추정 귀무가 필요 없는 짝지은 검정을 별도로 냈다 — 결론은 그쪽에 의존한다.
- **커밋된 런은 비표준 세션 가드 이전 판이다.** 영향 상한은 §2에 적었고 결론을 뒤집지 않지만, 재현 런을 돌리면 표본 n과 학습 행이 달라진다.
- **비용·체결·캐파시티 축이 없다.** 방향 채점 게임이라 과제에 그 축이 없기 때문이지 무시한 것이 아니다. 라이브 제출 경로로 옮기려면 데스크 파라미터 확정 뒤 다시 계산해야 한다.
- **수능일 라이브 경로가 비어 있다.** 연구 표본에서 버리는 것까지만 했는데, 대회는 그날도 제출을 요구한다. 16:45 기준 라벨링은 열려 있다.
- **체결틱 원천이 아직 출처 대장에 등록되지 않았다.** 경로는 코드에서 단일 지점으로 파생하지만 해시·크기 pin이 없다.

8 결론

15:34 제출 컷오프를 각자 다른 방식으로 걷어내는 두 예측 컨셉을 설계해 판정했고 둘 다 폐기했다. 사전등록한 상수 기준선(61.08%)이 학습된 모든 갈래를 이겼으며, 이 결론은 같은 날 짝지은 검정에서 추정 귀무 없이 성립한다. 컨셉 2는 컷오프 불확실성을 2.31배 줄였는데도 승률이 오르지 않았다 — 줄인 불확실성이 예측 가능한 부분이 아니었다는 뜻이다.

다음에 재야 할 것은 셋이다. (1) **기저율 자체의 예측 가능성** — 상수를 이기는 길은 상수의 방향을 국면에 따라 바꾸는 것이고, 아직 아무도 그 예측 가능성을 재지 않았다. (2) **확신도를 모델보다 먼저** — 이 연구는 확신도가 무정보라는 벽에 세 번 부딪혔다(확신 필터·앙상블·사후 혼합). 무정보면 그 위에 얹을 수 있는 것이 없다. (3) **수능일 라이브 경로.**

References

[1] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947. ISSN 1860-0980.

DOI: 10.1007/bf02295996. URL http://dx.doi.org/10.1007/bf02295996.