통방문은 읽히지만 시장은 맞혀지지 않는다: 금통위 당일 국채선물·국고채 방향 예측과 decision model Laya·Jev, 2012–2026

채권 퀀트 데스크
실험 0019_laya_mpc_tone · 0023_mpc_event_cash_futures · 0025_mpc_walkforward

Abstract

한국은행 금융통화위원회(금통위) 당일, 기준금리 결정 발표와 통화정책방향 결정문(통방문) 공개 뒤 국채선물과 국고채가 어느 쪽으로 움직일지를 *decision model*이 맞힐 수 있는지 세 실험으로 물었다. 첫 실험은 통방문 텍스트만으로 공개 직후 3년·10년 국채선물 방향을 오픈소스 decision model Laya(회사 PC에서 도는 3억 2천만 파라미터 트랜스포머)로 예측했고, 판을 여섯 번 바꿨다. 두 번째 실험은 과제를 두 사건(결정·통방문) × 선물·현물 8계약 × 1분–장 마감 시계 2,249문항으로 넓히고, 같은 질문을 API decision model Jev에도 똑같이 던졌다. 결과는 한 방향을 가리킨다. **통방문의 톤은 읽힌다** — LLM 이 매긴 요소 점수는 다음 회의의 인상 대 인하를 AUC 0.97–1.00 으로 가르고, Jev 는 그 점수를 rank correlation 0.62–0.91 로 재현한다. **그러나 방향은 맞혀지지 않는다** — 두 실험의 어떤 모델·판도 시험 구간에서 우연 범위를 벗어나지 못했고(첫 실험 1분 accuracy 0.38–0.62, 두 번째 실험 22회의 701문항 0.46–0.50), Laya fine-tuning과 Jev 의 차이는 −2.9%p(회의 단위 bootstrap 95% −23–+19%p)였다. 원인은 «읽은 톤 → 가격» 사이에 있다: «직전보다 매파면 하락» 규칙의 부호가 2020–2022 와 2023–2025 사이에 뒤집히고(0.77 → 0.17–0.25), 시장의 사전 기대(프리뷰 기사)는 통방문 톤을 rank correlation 0.90 으로 미리 알며, 결정 서프라이즈의 반응은 결정 직후 맞는 방향(0.58–0.68)에서 그날 마감에는 거꾸로(0.22–0.26) 되돌려진다. 세 번째 실험은 학습 구간을 2025 년까지 늘리고, 2024–2026 의 22회의마다 직전 회의까지 전부로 다시 학습하는 walk-forward 로 같은 질문을 물었다: 질문 단위 accuracy 는 0.575 로 올랐지만 회의 단위로는 22회 중 11회만 절반 넘게 맞혔고(우연), 같은 자료·seed 로 두 번 학습한 모델이 한 회의를 0.00 과 0.93 으로 갈랐다. decision model은 «톤 판독기» 로는 쓸 만하지만, 금통위 당일 방향 신호를 그 자체로 만들지는 못한다.

1 Introduction

데스크 노트의 매매 아이디어는 단순하다: 통방문이 공개되면 매파·비둘기를 판단해 국채선물을 사고판다. 통방문은 10:30 정시에 풀리고(2020 년 이후), 가격은 공개 뒤 수 초–수십 초에 걸쳐 움직인다. 사람이 읽기에는 늦고, 생성형 LLM 은 느리고 출력이 글이라 코드가 바로 쓰기 어렵다. 그래서 *decision model* — 글을 생성하지 않고, 입력(state)을 읽어 타입이 정해진 질문마다 선택지 확률을 내는 모델 — 이 이 일을 대신할 수 있는지가 물음이다.

두 실험을 묶는다. 실험 A(0019)는 «통방문 텍스트 → 공개 직후 3선·10선 방향» 하나를 깊게 팠다: 공개 시각을 2012 년까지 복원하고, 레이블 시계를 5분에서 초 단위 틱 1분으로 옮기고, 방법을 다섯 가지(fine-tuning·요소 가중합·distillation·LLM 요소 + 로지스틱·기대 톤 서프라이즈)로 바꿨다. 실험 B(0023)는 과제를 넓혔다: 결정 발표(E1)와 통방문 공개(E2) 두 사건, 선물 2개와 국고채 지표물 6개, 1분부터 장 마감까지 여러 시계, 그리고 온프렘 Laya 와 API 모델 Jev 의 *같은 조건* 비교.

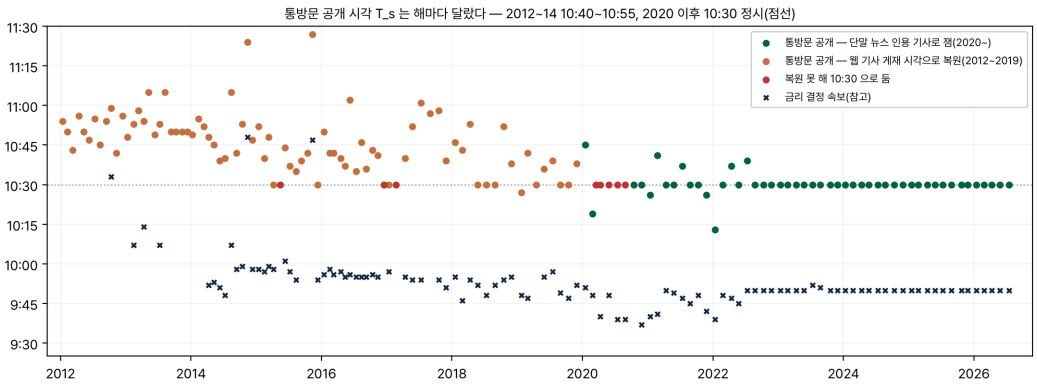


Figure 1: 회의별 통방문 공개 시각 T_s(점)와 결정 속보 시각 T_d(×). 2020 년부터 10:30 정시(점선). (실험 A)

Contributions.

1. **세 실험, 같은 결론.** 좁게(텍스트 → 선물 1분) 보든 넓게(두 사건 × 8계약 × 시계) 보든, Laya · Jev · LLM 요소 · 규칙 어느 것도 시험 구간 방향을 우연 이상으로 맞히지 못한다(§4, §5).
2. **읽기와 예측의 분리.** 톤 읽기는 강하게 성립한다(LLM 다음 결정 AUC 0.97 – 1.00, Jev rank correlation 0.62 – 0.91, distilled Laya 0.52 – 0.72). 실패는 읽기가 아니라 «톤 → 가격» 사상에 있다(§7).
3. **서프라이즈의 행방.** 결정 서프라이즈는 결정 직후 반응을 설명하지만 당일 마감까지 되돌려지고, 통방문 톤은 프리뷰가 거의 미리 안다 — 남는 서프라이즈가 작다(§8).
4. **API 대 온프렘.** Jev 는 0.24초 · 두 판 합계 $0.06 으로 빠르고 싸며 읽기가 낫지만, 방향 예측에서 Laya fine-tuning과 차이가 없다. 확률은 비결정적(같은 입력 최대 0.16 – 0.33 차)이고 보정하면 방향 정보가 사라진다(§5).
5. **학습을 늘려도, 회의마다 다시 배워도.** 학습 구간을 2025 년까지 늘리거나(2026 시험 0.416) 회의마다 다시 학습해도(walk-forward 0.575) 회의 단위로는 우연이다. fine-tuning 은 회의마다 한 방향을 내고, 그 방향이 학습의 작은 흔들림에 달려 있다(§6).

2 Task and Data

Notation. E1 · T_d 는 기준금리 결정 발표와 그 시각, E2 · T_s 는 통방문 공개와 그 시각이다. 3선 · 10선은 3년 · 10년 국채선물 (1틱 = 0.01), 국고N년은 국고채 N년 지표물 금리(bp = 0.01%p)다. *강세*는 선물 가격 상승 = 현물 금리 하락이다. 시계 1m · 5m · 30m 은 사건 뒤 1 · 5 · 30분, to_s 는 결정 뒤 통방문 공개 직전까지, close 는 그날 장 마감까지다.

Release-time reconstruction. 통방문 PDF 에는 공개 시각이 없다. 2020 년 이후는 뉴스 단말의 속보 · 인용 기사 첫 게재 시각으로, 2012 – 2019 년은 웹 기사 번호를 이분 탐색해 결정 기사 · 인용 기사 게재 시각으로 복원했다(웹 결정 기사는 속보보다 +3 – 4분 늦다, 2020 – 2021 13회 대조). 결정 시각은 속보 분 – +3분 안의 3선 거래량 급증 봉으로 정련했다. Figure 1 처럼 T_s 는 2012 – 2014 년 10:40 – 10:55 에서 점점 당겨져 2020 년부터 10:30 정시다. 2023 년 이후 10선 체결틱의 5초 거래량은 10:30:05 – 10:30:40 에 몰려 정시 공개를 확인한다.

Prices and labels. 선물은 연결 1분봉(2012 – 2026, 스탬프 = 봉의 끝 = 직전 체결가)과 초 단위 체결틱(2023 – 2026, 30회), 현물은 지표물 1분 격자(2022 – 2026, 체결량 0 인 행은 이월값이라 가격으로 쓰지 않는다 — 장내 체결과 대조해 스탬프 규약을 확인, 400분 중 392분 일치)다. 레이블은 사건 직전 가격 대비 시계 끝 가격의 부호이고, 선물 1틱 미만 · 현물 0.5bp 미만 움직임과 체결 없는 현물 시계는 질문을 만들지 않는다.

Samples and splits. 모든 분할은 시간순이다(Table 1). 실험 B 는 회의 단위로 하나만 나눈다 — 계열마다 나누면 같은 회의(같은 입력)의 질문 사이로 정답이 샌다(Figure 2).

Table 1: 두 실험의 과제·표본. 실험 A 의 주 레이블은 판 v5(공개 뒤 1분).

	실험 A (0019)	실험 B (0023)
사건	E2 통방문 공개	E1 결정 · E2 통방문
계약	3선 · 10선	3선 · 10선 · 국고 2·3·5·10·20·30년
시계	1분(주) · 30초 · 5분	선물 1m·5m·30m·to_s·close / 현물 5m·30m·to_s·close
입력	통방문 본문 + 직전 대비 변경 문장 + 결정	결정·설문·기대 톤·통방문 (+ 시장 움직임 문장 판)
분할	학습 2012–2022 · 시험 2023–2026	학습 ≤2023 · 시험 2024–2026
시험 표본	47문항(30회의, 초 단위 틱)	701문항(22회의)
모델	Laya 다섯 방법 · LLM 요소 · 규칙	규칙 4 · Laya zero-shot/fine-tuning · Jev zero-shot/보정

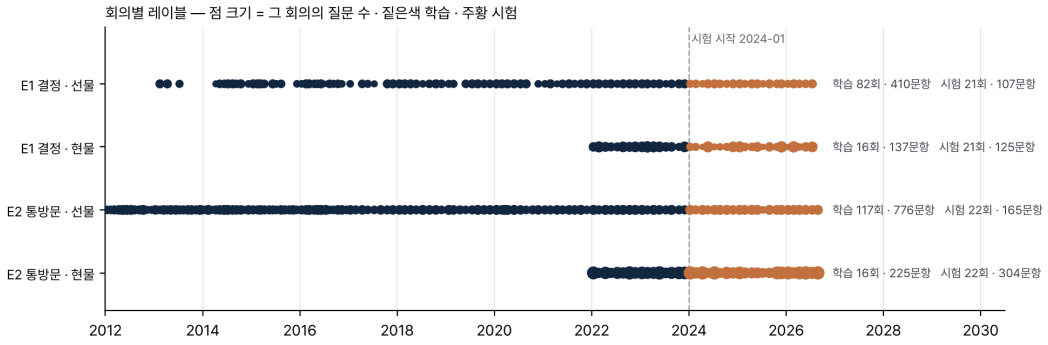


Figure 2: 실험 B 의 회의별 레이블. 짙은 점 학습, 주황 점 시험, 점 크기는 그 회의의 질문 수. (실험 B)

3 Models

Laya. `convaiinnovations/laya-multilingual`. 질문 하나를 «[CLS] 질문 [SEP] [MASK] 선택지$_1$ [MASK] 선택지$_2$ [SEP] 상태 [SEP]» 한 줄(최대 4,096 토큰)로 조립하고, 선택지 앞 [MASK] 자리의 은닉 상태를 점수로 바꿔 softmax 한다. 질문마다 전체 열을 다시 읽는다. fine-tuning은 공식 레시피(소프트 교차엔트로피 + 정책경사, Kaggle T4×2, seed 3개)를 따르고 시험 표본은 epoch 선택·온도 보정에 쓰지 않았다. 실험 A 는 2지선다 fine-tuning(C), 요소 8개 5단계 점수의 가중합 끝단 학습(B), LLM 요소 점수로 가르친 distillation(D)를, 실험 B 는 zero-shot(L0)과 선물·현물·두 사건 한 모델 fine-tuning(L1, 4epoch)을 돌렸다.

Jev. TypeSafe 의 API decision model(`jev-1.13.0`). 같은 state·질문 형식을 받고 질문마다 선택지 확률을 돌려준다. 가중치에 접근할 수 없어 학습은 못 하고, 학습 구간 예측으로 Platt 보정($P = \sigma(a \operatorname{logit} p + b)$)만 했다(J0c). 산술·개수·날짜를 못 한다는 공식 제약이 있어 숫자는 낱말로 줬다.

LLM factors and rules. LLM 이 통방문마다 요소 8개(성장·물가·환율·금융안정·가이던스·불확실성·전반 톤·직전 대비 변화)를 0–4 로 매긴 점수를 «선생 레이블» 로 쓴다. 규칙 기준선은 학습 구간 다수 방향(B0), 결정 규칙(인하 → 강세), 결정 서프라이즈(채권 전문가 설문 보다 매파 결정 → 약세, B2), 기대 톤 서프라이즈($s = \frac{1}{2}[(\text{전반 톤} - \text{기대}) + (\text{가이던스} - \text{기대})]$, B3)다. 기대 톤은 회의 전 프리뷰 기사를 LLM 이 같은 척도로 읽어 만들었다.

4 Results A — Statement Text to Futures Direction Right After Release

실험 A 의 시험(2023–2026 초 단위 틱 1분, 47문항, 상승 61.7%)에서 zero-shot 0.38, fine-tuning(C) seed 0.47/0.62/0.57, 요소 가중합(B) 0.43–0.49, distillation(D) 0.47, LLM 요소 + 로지스틱(E) 0.55 로 모두 0.5 대비 이항검정 $p \geq 0.14$ 다(Figure 3). 잘 나온 fine-tuning seed 는 거의 전부 «상승» 을 찍어 시험 기저율과 같고, 보정 온도는 세 seed 모두 상한 10.0 이었다 — 학습 구간 안에서도 확신을 낼 근거를 찾지 못했다는 뜻이다. 판을 바꿔도(결정 속보 뒤 5분 → 통방문 공개 뒤 5분 → 1분, 띄어쓰기 보정, 질문 문구 교정) 결론은 같았다.

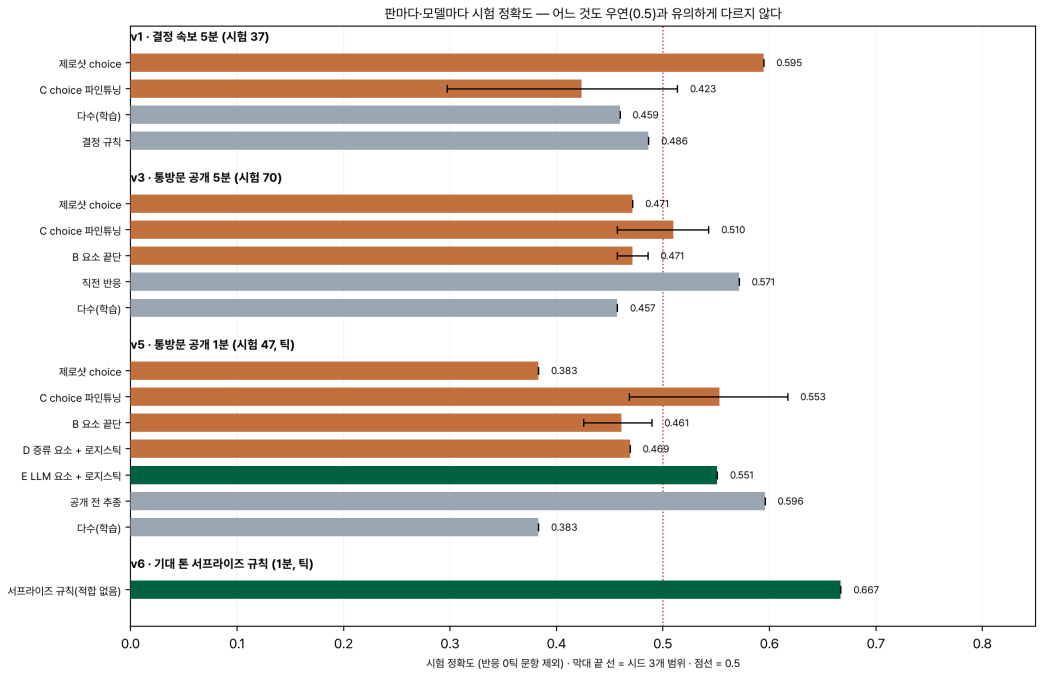


Figure 3: 실험 A 의 판·모델별 시험 accuracy. 막대 끝 선 = seed 3개 범위, 점선 = 0.5. (실험 A)

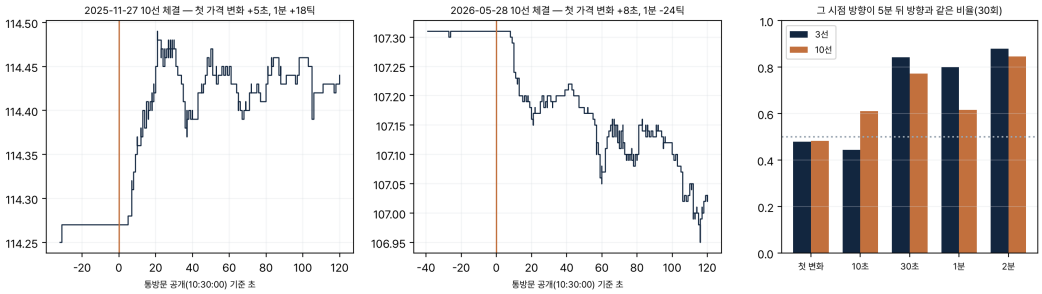


Figure 4: 공개 직후 10선 체결가(왼쪽 둘)와 시점별 방향이 5분 뒤 방향과 같은 비율(오른쪽, 30회). (실험 A)

공개 직후 반응의 모양도 쉽지 않다(Figure 4): 10선의 첫 가격 변화는 공개 뒤 중앙값 6.5초, 3선은 19초이고, **첫 가격 변화의 방향과 5분 뒤 방향은 48% 만 같다**. 즉 «통방문의 방향» 을 하나로 정의하기부터 어렵다.

5 Results B — Two Events, Futures and Cash, Laya vs. Jev

Table 2 와 Figure 5 에서 모든 모델이 0.5 근처다. B0 가 0.5 아래인 것은 학습 구간의 다수 방향(대부분 약세)이 시험 구간 (2024 년 이후 인하기)에 뒤집혔기 때문이다. 같은 질문 짝의 accuracy 차이를 회의 단위 군집 bootstrap(2,000회)으로 보면 어느 비교도 구간이 0 을 벗어나지 않는다(Figure 6): 주 비교 L1 − J0 = −0.029 (−0.232 – +0.191), 시장 판 +0.021 (−0.142 – +0.166), 시장 문장의 효과는 Laya +0.003, Jev −0.047. 정확 McNemar 검정은 같은 회의의 질문을 독립으로 세어 p 가 작게 나오므로 판정에 쓰지 않았다. 공개 뒤 첫 1분 선물 방향이 그 뒤 현물 방향을 예고하는지도 보았으나(국고 3·10·30년, T+1분 – T+h) 0.39 – 0.63, 모두 $p > 0.35$ 였다.

What the probabilities say. Laya 는 강세 확률이 0.5 근처에 몰리고(zero-shot은 97% 를 약세로 답함), Jev 는 0·1 쪽으로 강하게 갈리지만 신뢰도 곡선이 평평하다(Figure 7). 학습

Table 2: 실험 B 시험 accuracy(22회의 · 701문항). _m = 시장 판(결정 전·통방문 전 시장 움직임 문장을 입력에 더함).

가족	문항	B0	B2	B3	L0	L1	J0	L0_m	L1_m	J0_m
전체	701	.439	.466	.475	.485	.474	.502	.459	.476	.455
E1 선물	107	.430	.467	.430	.505	.467	.533	.318	.495	.364
E1 현물	125	.472	.528	.472	.576	.448	.560	.592	.544	.400
E2 선물	165	.497	.491	.539	.424	.461	.539	.358	.418	.479
E2 현물	304	.398	.428	.457	.474	.493	.447	.510	.474	.497

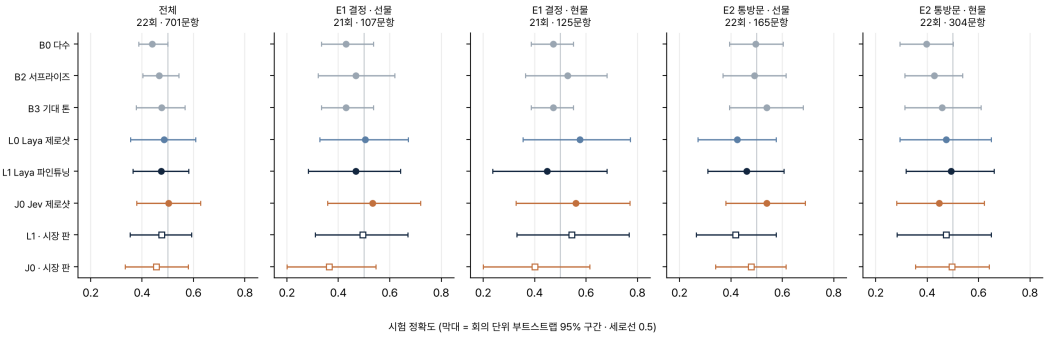


Figure 5: 실험 B 가족별 시험 accuracy와 회의 단위 bootstrap 95% 구간. 빈 네모 = 시장 판. (실험 B)

구간으로 적합한 Platt 기울기는 L0 +0.098, J0 +0.011 로 0 에 가깝다 — 원래 확률이 방향과 무관하다는 뜻이고, 보정하면 Brier 0.331 → 0.251, ECE 0.26 → 0.04 로 «확률로서» 나아지지만 출력이 한 값으로 모인다. Jev 는 같은 입력을 두 번 보내면 강세 확률이 최대 0.16(공개 판)·0.33(시장 판) 달랐다. L1 의 시험 log loss은 첫 epoch부터 $\ln 2 = 0.693$ 에 붙어 있었고 보정 온도는 6개 중 5개가 상한 10.0 이다 — 실험 A 와 같은 증상이다.

6 Results C — Longer Training and Walk-Forward Retraining

실험 B 는 학습 ≤ 2023 · 시험 2024– 로 한 번 나눴다. 통방문 공개 양상이 2023 년 무렵부터 일정해졌다는 데스크의 관찰에 따라, 세 번째 실험(0025)은 질문·state·모델·학습 설정을 그대로 두고 *분할만* 바꿨다: (S25) 학습 ≤ 2025 · 시험 2026 의 6회의, (WF) 2024–2026 의 22회의마다 그 회의 *직전 회의까지 전부*로 Laya 를 다시 fine-tuning 해 그 회의 하나만 예측(학습 항목 1,239 → 1,779, 회의마다 seed 1개, Kaggle T4×2 에서 29번 학습). Jev·Laya zero-shot 은 학습이 없어 실험 B 의 예측을 그대로 쓰고, 다수 기준선과 Platt 보정만 분할마다 다시 맞췄다.

표 3 에서 학습을 2025 년까지 늘린 S25 는 2026 년에 0.416 으로 오히려 0.5 아래였다(이항 $p = 0.02$, seed 별 0.356 · 0.361 · 0.441). WF 는 0.575(2026 만 0.698)로 실험 B 의 0.474 보다 높았지만, 같은 질문 짝의 회의 단위 군집 bootstrap 은 WF − S23 +0.101(−0.029–+0.235), WF − J0 +0.073(−0.126–+0.280)로 0 을 벗어나지 않는다. 연도별 WF accuracy 는 0.496 → 0.553 → 0.698 로 오르지만(Figure 8), 회의 단위로 세면 4/8 · 4/8 · 3/6 회의다.

Accuracy 는 회의 수만큼의 동전이다. 같은 회의의 질문은 한쪽으로 쏠리고(회의별 정답 강세 비율 중앙 0.58), fine-tuning 된 Laya 는 그보다 더 한 방향만 낸다 — WF 22회의 중 12회의에서 강세 답 비율이 0.1 이하나 0.9 이상이다(Figure 9). 그래서 회의 하나가 통째로 맞거나 틀리고, 절반 넘게 맞힌 회의는 WF 11/22 · S23 9/22 · Jev 10/22 로 우연(11)과 같다. 질문 단위 0.575 는 질문이 많은 몇 회의가 맞은 결과다.

같은 학습이 같은 답을 내지 않는다. S25 seed 0 과 WF(2026-01-15) seed 0 은 학습 자료(회의일 < 2026-01-01)·설정·seed 가 같고, 다른 것은 epoch 마다 기록용으로 평가하는 시험 표본뿐이다. 두 모델은 2026-01-15 회의 30문항을 0.00 과 0.93 으로 맞혔다(Figure 10). 학습 로그에서 epoch 3–4 에 학습 손실이 0.65 에서 1.6 으로 튀고 시험 log loss 가 0.13–6.3 으로

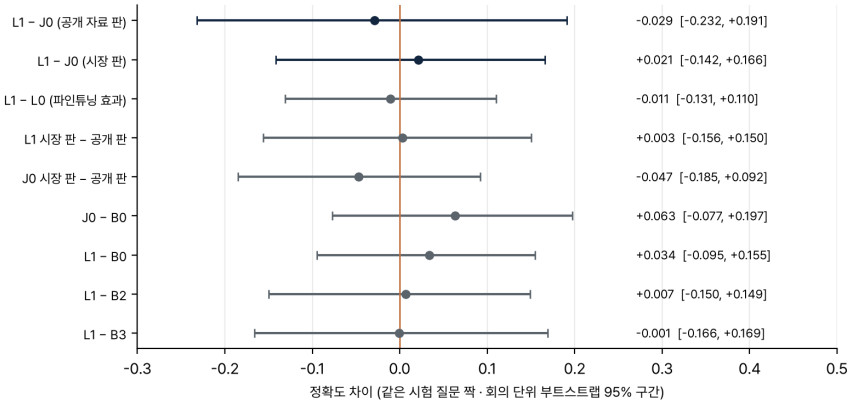


Figure 6: 같은 시험 질문 짝의 accuracy 차이와 회의 단위 bootstrap 95% 구간. (실험 B)

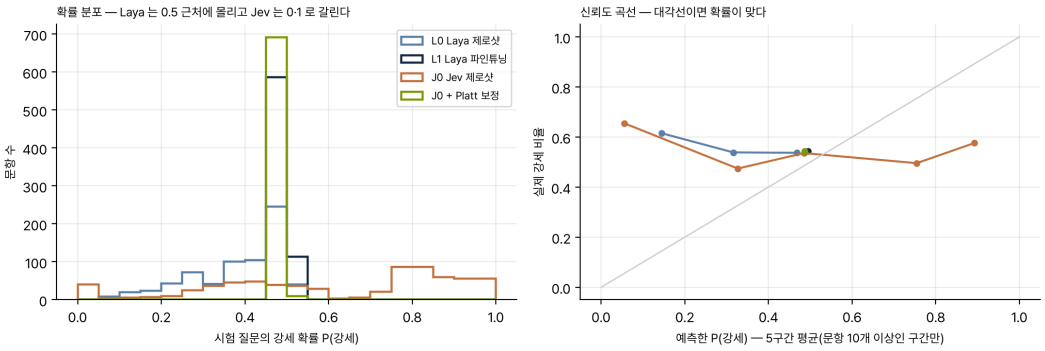


Figure 7: 시험 질문의 강세 확률 분포(왼쪽)와 reliability diagram(오른쪽). (실험 B)

갈라진다. 즉 이 과제의 fine-tuning 은 회의마다 한 방향을 정하는 불안정한 과정이고, WF 2026 의 0.698 은 그 흔들림의 한 표본이다.

7 Tone Reading Works

방향 예측과 대조적으로, 통방문 *읽기*는 세 방향에서 성립한다(Figure 11, 12). (1) LLM 요소 점수가 회의 t 의 통방문만 보고 회의 $t+1$ 의 인상 대 인하(34회)를 AUC 전반 톤 0.998 · 물가 0.986 · 가이던스 0.965 로 가른다(zero-shot Laya 0.34 – 0.61). LLM 이 이후 역사를 알고 있을 수 있어 사후 지식 누설은 배제하지 못하지만, 금리 사이클의 지속성이 대부분을 설명한다. (2) LLM 점수로 distillation한 Laya 가 시험 통방문(2023 –)에서 그 점수를 rank correlation 성장 0.73 · 가이던스 0.72 · 전반 톤 0.52 로 재현한다(나머지 요소는 과적합). (3) Jev 는 학습 없이 요소 8개 모두 0.62 – 0.91(성장 0.91 · 물가 0.87 · 금융안정 0.84 · 환율 0.81 · 불확실성 0.76 · 가이던스 0.70 · 전반 톤 0.68 · 변화 0.62)로 LLM 점수를 따라 한다. 같은 질문의 Laya zero-shot은 −0.12 – 0.44 다.

8 Where the Surprise Goes

읽기가 되는데 예측이 안 되는 이유를 세 관찰이 가리킨다.

The tone-to-direction sign flips. «직전 통방문보다 매파로 바뀌면 하락» 규칙(적합 없음)은 2020 – 2022 에 0.77(30문항, $p = 0.005$)이었지만 2023 – 2025 에는 해마다 0.17 – 0.25 로 뒤집혔다(Figure 13). 분봉 1분과 틱 1분의 부호가 46문항 모두 같아 측정 차이가 아니다. 학습 구간에서 배운 «매파 → 하락» 이 시험 구간에서 거꾸로 작동하면, 어떤 학습 모델도 시험에서 우연 이하로 떨어질 수 있다.

Table 3: 세 분할의 질문 단위 accuracy. L1 = Laya fine-tuning(S23·S25 seed 3개 평균, WF 회의마다 seed 1개). (실험 C)

시험	분할	B0	L0	J0	J0c	L1
2026 · 6회의 · 202문항	S23	.500	.515	.525	.480	.520
	S25	.470	.515	.525	.550	.416
	WF	.465	.515	.525	.475	.698
2024–2026 · 22회의 · 701문항	S23	.439	.485	.502	.469	.474
	WF	.438	.485	.502	.439	.575

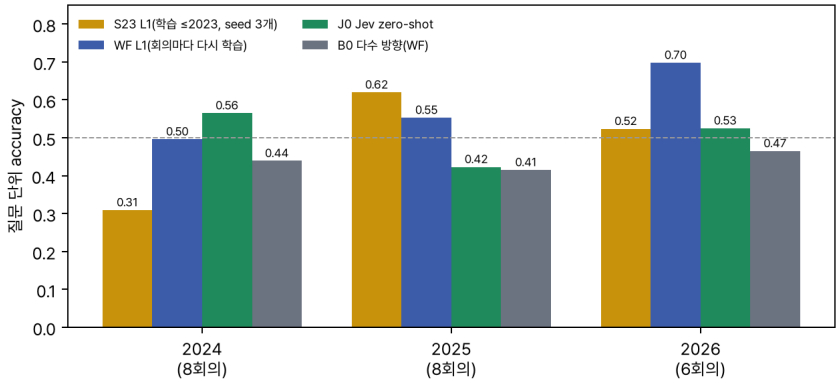


Figure 8: 연도별 질문 단위 accuracy — S23 L1 · WF L1 · Jev zero-shot · 다수 기준선. 괄호는 회의 수. (실험 C)

The market already knows the tone. 회의 전 프리뷰 기사로 만든 기대 톤이 실제 통방문 톤을 rank correlation 전반 톤 0.903 · 가이던스 0.906 으로 맞히고, 기대 결정 적중은 94.4%(54회 중 51)다. 그래서 톤 서프라이즈가 54회 중 28회 0 이고(Figure 14), 남은 서프라이즈로 낸 규칙은 두 구간 모두 0.5 위(2023– 틱 1분 0.667, $p = 0.15$ · 2020–2022 0.529)로 부호가 안 뒤집힌 유일한 규칙이지만 유의하지 않다. 실험 B 의 시험(2024–)에서 같은 규칙(B3)은 0.475 였다.

The decision surprise reverses within the day. 채권 전문가 설문 대비 결정 서프라이즈는 결정 속보 1분 반응과 3선 rank correlation -0.61($p = 0.021$, 14회)로 측정이 맞다. 서프라이즈가 있던 14회의 전부에 «매파 서프라이즈 → 약세» 를 적용하면 E1(결정 직후)은 0.58–0.68 로 맞는 쪽이지만, 같은 규칙이 E2 *마감*은 선물 0.22 · 현물 0.26 으로 거꾸로 맞힌다(Figure 15). 결정 서프라이즈는 09:5x 속보에서 값이 매겨지고, 통방문 공개 뒤 장 마감까지 되돌려진다. 2024 년 이후 시험 22회의에는 서프라이즈가 6회뿐이라 이 신호가 시험 accuracy 에 드러날 자리도 작다.

9 Latency and Cost

가격은 공개 순간 뛰지 않는다. 밀리초 체결·호가 두 회(2026-07-16·08-27)에서 첫 가격 변화는 7–21초, 3틱까지 22–28초였고, 방향을 완벽히 안다면 20초 늦게 들어가도 몫이 거의 그대로였다(Figure 16). 이 기준에서 Jev(사례 하나 0.24초)는 물론 Laya(CPU, 질문 하나 약 4초 · 질문 약 20개를 묶은 통방문 사례 84초)도 질문 수만 줄이면 늦지 않다. 비용은 Jev 두 판 합계 입력 150만 토큰 약 $0.06 이었다. **즉 병목은 속도·비용이 아니라 방향 정보다.**

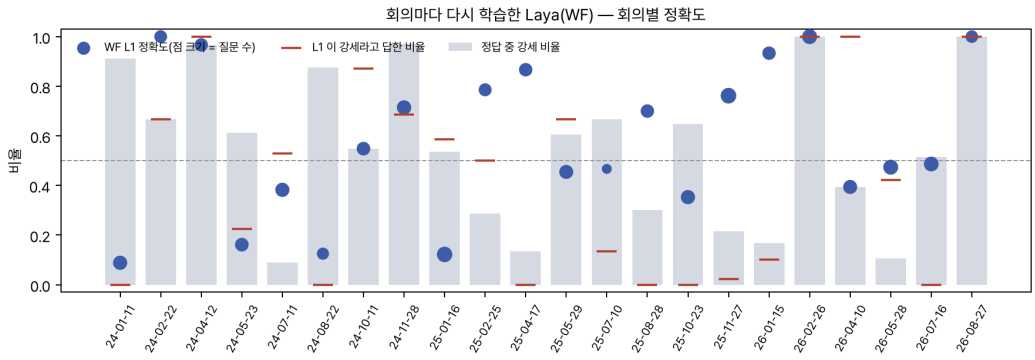


Figure 9: WF 회의별 — 정답 중 강세 비율(회색 막대), L1 이 강세라고 답한 비율(빨간 선), L1 accuracy(점, 크기 = 질문 수). (실험 C)

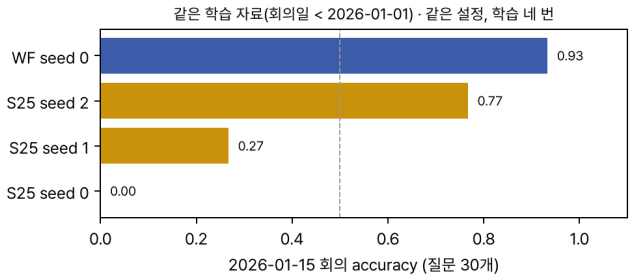


Figure 10: 같은 학습 자료·같은 설정으로 학습한 네 모델의 2026-01-15 회의 accuracy. (실험 C)

10 Limitations

- **표본이 작다.** 실험 A 시험 47문항(30회의), 실험 B 시험 22회의. 같은 회의의 질문은 서로 얽혀 실효 표본이 회의 수에 가깝고, 최소 검출 차이는 15–20%p 다. 이 리포트는 «큰 차이가 없다» 까지만 말한다.
- **국면이 둘뿐이다.** 2020–2022 인상기와 2023–2026 동결·인하기 사이에서 부호가 뒤집혔다. 다른 국면에서 학습된 사상이 통할지는 이 자료로 알 수 없다.
- **복원한 시각.** 2012–2019 의 결정·공개 시각은 웹 기사 시각으로 복원한 값이다(대조 +3–4분).
- **시장 문장은 거칠다.** 실험 B 의 시장 판은 방향과 한 단계 크기만 주었다.
- **walk-forward 는 seed 1개.** 회의마다 seed 3개(66번 학습)는 Kaggle 주간 GPU 한도를 넘었다. 같은 설정의 학습이 크게 흔들리므로 판정은 회의 단위로 했다.
- **LLM 선생의 사후 지식.** LLM 요소 점수와 기대 톤은 회의일 이후 지식을 쓰지 말라는 지시만으로 만들었다.

11 Conclusion

decision model은 통방문을 읽는다 — 특히 Jev 는 학습 없이 LLM 의 요소 판단을 따라 하고, 빠르고 싸다. 그러나 세 실험 어디에서도 그 읽기가 금통위 당일의 선물·현물 방향으로 이어지지 않았다 — 학습 구간을 늘려도, 회의마다 다시 학습해도 회의 단위로는 우연이었다. «톤 → 가격» 의 부호는 국면마다 바뀌고, 시장은 톤을 미리 알며, 결정 서프라이즈는 당일 되돌려진다. 남은 후보는 읽기를 *방향*이 아니라 *조건*으로 쓰는 것이다: 프리뷰가 예상하지 못한 문장 단위 서프라이즈, 국면을 함께 넣은 조건부 사상, 그리고 결정 서프라이즈의 당일 되돌림 자체를 매매 대상으로 삼는 것. decision model은 그때 톤 판독기로 다시 쓸 만하다.

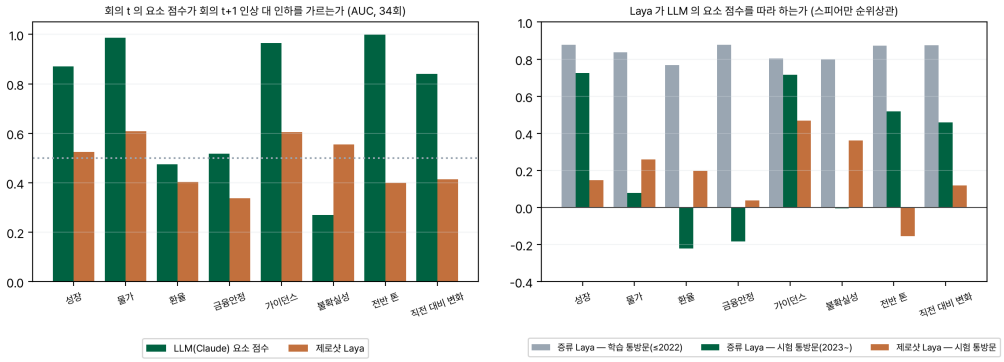


Figure 11: 왼쪽 — 회의 t 의 요소 점수로 회의 $t+1$ 인상 대 인하를 가르는 AUC. 오른쪽 — distilled Laya 의 LLM 점수 재현(rank correlation). (실험 A)

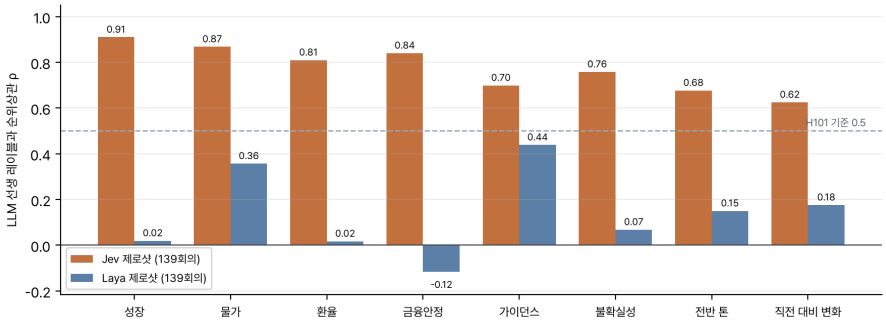


Figure 12: 요소 8개 읽기의 LLM 점수와의 rank correlation(139회의) — Jev zero-shot 대 Laya zero-shot. (실험 B)

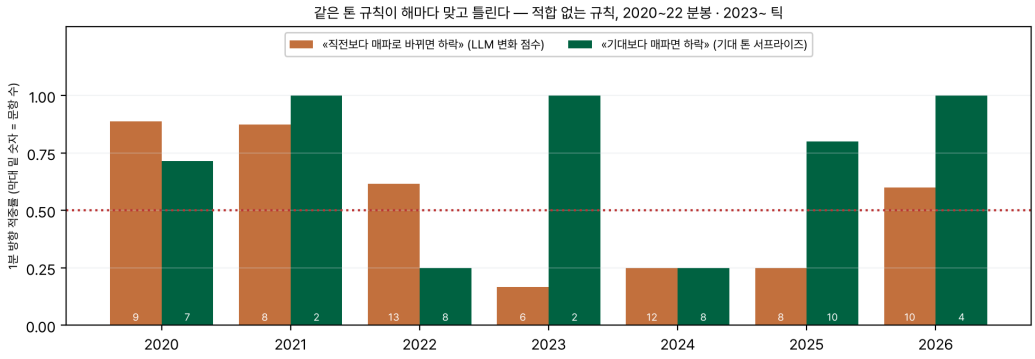


Figure 13: 적합 없는 두 규칙의 해마다 1분 적중률 — «직전보다 매파면 하락»(주황)과 «기대보다 매파면 하락»(초록). (실험 A)

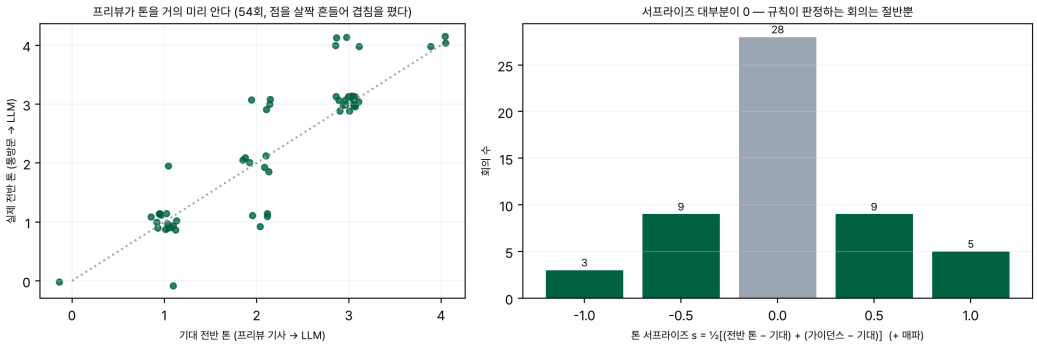


Figure 14: 기대 전반 톤 대 실제 전반 톤(왼쪽, 54회)과 톤 서프라이즈의 분포(오른쪽). (실험 A)

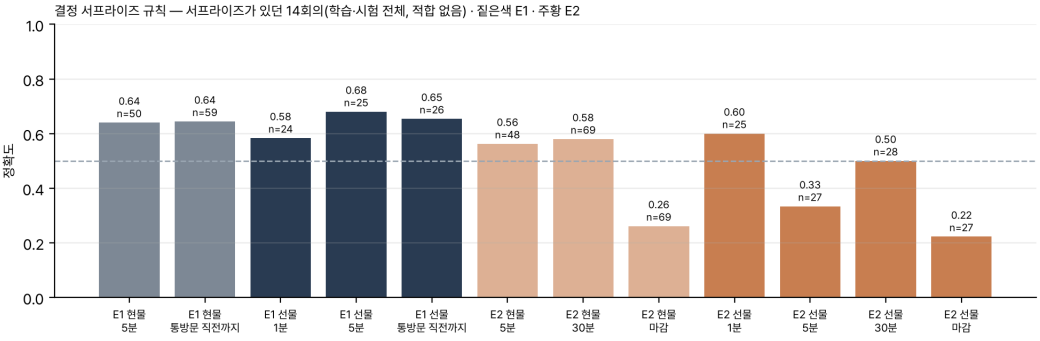


Figure 15: 결정 서프라이즈 규칙의 사건·계열·시계별 accuracy(서프라이즈 14회의, 적합 없음). 짙은색 E1, 주황 E2. (실험 B)

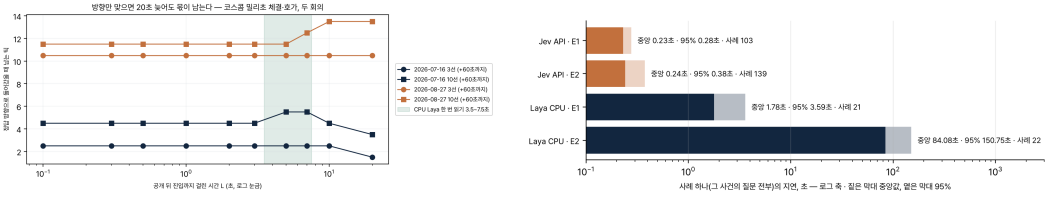


Figure 16: 왼쪽 — 방향을 안다고 치고 공개 뒤 L 초에 들어갈 때 60초 뒤 남는 틱(밀리초 자료 2회, 실험 A). 오른쪽 — 사례 하나의 지연, Jev API 대 Laya CPU(실험 B).

Appendix

A Dataset Samples

본문의 표와 그림은 아래 실물에서 나왔다. 숫자와 원문은 결과 파일에서 그대로 옮겼다(줄인 곳은 «중략» 으로 적었다).

A.1 Label Ledger — Rows of One Meeting

레이블 원장(`labels.csv`, 2,858행 중 레이블 2,249개)의 2026-07-16 회의 일부. 한 행 = (회의, 사건, 계약, 시계). `delta` 는 사건 직전 가격 대비 변화(선물 틱 · 현물 bp), `label` 은 부호를 «강세(가격 상승 = 금리 하락)/약세» 로 맞춘 정답이다. 선물 1틱 미만 · 현물 0.5bp 미만이면 질문을 만들지 않는다(`why`).

사건	앵커	계열	계약	시계	delta	단위	정답
E1	09:52	fut	3선	1m	+7	tick	강세
E1	09:52	fut	3선	5m	+4	tick	강세
E1	09:52	fut	3선	to_s	+2	tick	강세
E1	09:52	fut	10선	1m	+11	tick	강세
E1	09:52	fut	10선	5m	+6	tick	강세
E1	09:52	fut	10선	to_s	+1	tick	강세
E1	09:52	cash	국고3년	5m	-1.4	bp	강세
E1	09:52	cash	국고3년	to_s	-0.7	bp	강세
E1	09:52	cash	국고10년	5m	-0.8	bp	강세
E1	09:52	cash	국고10년	to_s	-0.2	bp	— flat
E2	10:30	fut	3선	1m	-3	tick	약세
E2	10:30	fut	3선	5m	-4	tick	약세
E2	10:30	fut	3선	close	+0	tick	— flat
E2	10:30	fut	10선	1m	-6	tick	약세
E2	10:30	fut	10선	5m	-9	tick	약세
E2	10:30	fut	10선	close	+14	tick	강세

A.2 Structure of One Case

모델의 입력 단위는 «사례»(회의 하나의 사건 하나)다. 사례 242개(학습 199 · 시험 43), E2 사례의 state 길이 중앙값 2,041자. 아래는 `2026-07-16_E2` 의 뼈대다 (state 본문은 다음 절에 실었다).


```
{
 "id": "2026-07-16_E2",
 "split": "test",
 "state": {
  "기준금리 결정": "인상 - 새 기준금리 2.75% (직전보다 25bp 인상)",
  "시장 예상(채권 전문가 설문)": "응답자 약 66%가 인상 예상",
  "예상 대비 실제 결정": "예상보다 긴축적(매파 서프라이즈)",
  "시장의 사전 기대(프리뷰 기사)": "예상 결정 인상 · 기대 전반 톤 긴축적 · 기대 포워드 가이던스 긴축적",
  "직전 통방문 대비 새로 등장한 문장": "<문장 6개>",
  "직전 통방문에서 사라진 문장": "<문장 6개>",
  "통방문 본문": "<문장 7개>"
 },
 "questions": {
  "3선_1m": "<choice: 강세/약세>",
  "3선_5m": "<choice: 강세/약세>",
  "3선_30m": "<choice: 강세/약세>",
  "...": "외 18개"
 },
 "gold": {
  "3선_1m": {
   "label": "약세",
   "probabilities": {
    "강세": 0.0,
    "약세": 1.0
   }
  },
  "3선_5m": {
```

```
   "label": "약세",
   "probabilities": {
    "강세": 0.0,
    "약세": 1.0
   }
  }
 },
 "meta": {
  "day": "2026-07-16",
  "event": "E2",
  "decision": "인상",
  "surprise": 1,
  "q": {
   "3선_1m": {
    "kind": "fut",
    "contract": "3선",
    "horizon": "1m",
    "big": true
   },
   "3선_5m": {
    "kind": "fut",
    "contract": "3선",
    "horizon": "5m",
    "big": true
   }
  }
 }
}
```


회의일은 state 에 넣지 않는다(사전학습 기억으로 그 시기 시장을 떠올리지 않게). `meta.q.big` 은 «크게 움직였나» 한 비트로, 평가 층화에만 쓴다.

B Model I/O Samples — 2026-07-16 MPC Meeting

같은 회의의 두 사건을 두 모델에 실제로 준 그대로, 실제로 나온 그대로 싣는다. Jev 는 사례마다 한 번 호출하고 그 사건의 질문을 모두 묶는다. Laya 도 같은 state·같은 질문을 받는다(글자 하나 다르지 않다).

B.1 Rate Decision (E1)

Request meta

Jev: `api.typesafe.ai/v1/systemone` · 요청 모델 `jev-latest` → 응답 `jev-1.13.0` · 질문 16개 한 번에 · 입력 2,477 토큰 · 왕복 0.27초
Laya: `convaiinnovations/laya-multilingual` · 같은 입력 · max_len 4096 · L0 zero-shot / L1 fine-tuning(seed 3개 평균)

Input — state

기준금리 결정: 인상 — 새 기준금리 2.75% (직전보다 25bp 인상)
시장 예상(채권 전문가 설문): 응답자 약 66%가 인상 예상
예상 대비 실제 결정: 예상보다 긴축적(매파 서프라이즈)

Input — question (총 16개 중 하나를 그대로)


```
{
 "3선_1m": {
  "type": "choice",
  "instructions": "한국은행 금융통화위원회가 방금 기준금리 결정을 발표했다. 발표 직후 1분 동안 3년
  ↪  국채선물 가격은 발표 직전보다 어떻게 움직였나?",
  "criteria": {
   "강세": "가격 상승 - 금리 하락, 채권 강세",
   "약세": "가격 하락 - 금리 상승, 채권 약세"
  }
 }
}
```


나머지 질문은 같은 꼴에 계약·시계만 다르다: `3선_5m`, `3선_to_s`, `10선_1m`, `10선_5m`, `10선_to_s`, `국고2년_5m`, `국고2년_to_s`, `국고3년_5m`, `국고3년_to_s`, `국고5년_5m`, `국고5년_to_s`, `국고10년_5m`, `국고20년_5m`, `국고20년_to_s`, `국고30년_5m`

Jev response (원응답 JSON 중 4개 질문)


```
{"model": "jev-1.13.0",
 "answers": {
  "3선_1m": {"type": "choice", "choice": "약세", "confidence": 0.99, "probabilities":
  ↪ {"강세": 0.01, "약세": 0.99}},
  "3선_5m": {"type": "choice", "choice": "약세", "confidence": 0.99, "probabilities":
  ↪ {"강세": 0.0, "약세": 1.0}},
  "10선_1m": {"type": "choice", "choice": "약세", "confidence": 0.99, "probabilities":
  ↪ {"강세": 0.01, "약세": 0.99}},
  "10선_5m": {"type": "choice", "choice": "약세", "confidence": 0.98, "probabilities":
  ↪ {"강세": 0.01, "약세": 0.99}}
  ...(나머지 12개 중략) },
 "usage": {"input_tokens": 2477, "output_tokens": 584}}
```


Laya response — L0 zero-shot 확률 (// 뒤는 L1 fine-tuning seed 3개 평균)

```
"3선_1m": {"강세": 0.501, "약세": 0.499}   // L1 강세 0.495
"3선_5m": {"강세": 0.501, "약세": 0.499}   // L1 강세 0.496
"10선_1m": {"강세": 0.502, "약세": 0.498}   // L1 강세 0.496
"10선_5m": {"강세": 0.501, "약세": 0.499}   // L1 강세 0.496
```

Ground truth check — 질문 16개 (괄호 숫자 = 강세 확률, ✓= 맞음)

질문	정답	Jev(J0)	Laya L0	Laya L1
3선_1m	강세	약세 0.01	강세 0.50 ✓	약세 0.50
3선_5m	강세	약세 0.00	강세 0.50 ✓	약세 0.50
3선_to_s	강세	약세 0.01	약세 0.50	약세 0.49
10선_1m	강세	약세 0.01	강세 0.50 ✓	약세 0.50
10선_5m	강세	약세 0.01	강세 0.50 ✓	약세 0.50
10선_to_s	강세	약세 0.01	약세 0.50	약세 0.49
국고2년_5m	강세	약세 0.04	강세 0.50 ✓	약세 0.49
국고2년_to_s	강세	약세 0.05	약세 0.50	약세 0.49
국고3년_5m	강세	약세 0.05	강세 0.50 ✓	약세 0.49
국고3년_to_s	강세	약세 0.06	약세 0.50	약세 0.49
국고5년_5m	강세	약세 0.04	강세 0.50 ✓	약세 0.49
국고5년_to_s	강세	약세 0.05	약세 0.50	약세 0.49
국고10년_5m	강세	약세 0.05	강세 0.50 ✓	약세 0.49
국고20년_5m	강세	약세 0.04	강세 0.50 ✓	약세 0.49
국고20년_to_s	강세	약세 0.04	강세 0.50 ✓	약세 0.49
국고30년_5m	강세	약세 0.04	강세 0.50 ✓	약세 0.49
맞힌 수		0/16	11/16	0/16

B.2 Statement Release (E2)

Request meta

Jev: `api.typesafe.ai/v1/systemone` · 요청 모델 `jev-latest` → 응답 `jev-1.13.0` · 질문 21개 한 번에 · 입력 5,380 토큰 · 왕복 0.26초
Laya: `convaiinnovations/laya-multilingual` · 같은 입력 · max_len 4096 · L0 zero-shot / L1 fine-tuning(seed 3개 평균)

Input — state

기준금리 결정: 인상 — 새 기준금리 2.75% (직전보다 25bp 인상)
시장 예상(채권 전문가 설문): 응답자 약 66%가 인상 예상
예상 대비 실제 결정: 예상보다 긴축적(매파 서프라이즈)
시장의 사전 기대(프리뷰 기사): 예상 결정 인상 · 기대 전반 톤 긴축적 · 기대 포워드 가이던스 긴축적
직전 통방문 대비 새로 등장한 문장 (6개) ... 외 4개(중략)

- 성장세가 수출과 투자를 중심으로 강화되고 있는 가운데 물가상승률은 상당기간 목표수준을 상회할 것으로 보이고 금융안정 측면의 리스크도 지속되고 있는 만큼 기준금리를 0.25%p 인상하는 것이 적절하다고 판단하였다.

- 세계경제는 중동상황의 불확실성이 이어지고 있지만 견조한 AI 투자로 완만한 성장세를 지속할 것으로 보이며, 물가는 그간의 에너지 가격 상승이 시차를 두고 영향을 미치면서 당분간 높은 오름세를 나타낼 것으로 예상된다.

직전 통방문에서 사라진 문장 (6개) ... 외 4개(중략)

- 중동전쟁 영향으로 물가상승 압력이 높아진 반면 성장세는 수출 호조 등에 힘입어 예상보다 확대되고 있고 금융안정 측면에서도 리스크가 지속되고 있지만, 중동사태 전개 및 파급 영향과 관련한 불확실성이 여전히 높은 만큼 현재의 기준금리 수준을 유지하면서 사태의 추이와 성장· 물가에 미치는 영 ...(중략, 전체 190자)
- 세계경제는 AI 관련 투자 확대에도 중동전쟁에 따른 에너지 원자재 가격 상승 및 공급 차질의 영향으로 성장세가 둔화되겠으나 물가상승 압력은 상당폭 확대될 것으로 보인다.

통방문 본문 (7개) ... 외 5개(중략)

- 금융통화위원회는 다음 통화정책방향 결정시까지 한국은행 기준금리를 현재의 2.50% 수준에서 2.75%로 상향 조정하여 통화정책을 운용하기로 하였다. 성장세가 수출과 투자를 중심으로 강화되고 있는 가운데 물가상승률은 상당기간 목표수준을 상회할 것으로 보이고 금융안정 측면의 리스크도 지속되 ...(중략, 전체 201자)
- 세계경제는 중동상황의 불확실성이 이어지고 있지만 견조한 AI 투자로 완만한 성장세를 지속할 것으로 보이며, 물가는 그간의 에너지 가격 상승이 시차를 두고 영향을 미치면서 당분간 높은 오름세를 나타낼 것으로 예상된다. 국제금융시장에서는 미 연준 금리인상 기대 및 중동상황의 변화 등에 영향 ...(중략, 전체 312자)

Input — question (총 21개 중 하나를 그대로)

```
{
 "3선_1m": {
  "type": "choice",
  "instructions": "한국은행이 방금 통화정책방향 결정문을 공개했다. 공개 직후 1분 동안 3년 국채선물
  ↪  가격은 공개 직전보다 어떻게 움직였나?",
  "criteria": {
   "강세": "가격 상승 - 금리 하락, 채권 강세",
   "약세": "가격 하락 - 금리 상승, 채권 약세"
  }
 }
}
```

나머지 질문은 같은 꼴에 계약·시계만 다르다: `3선_5m`, `3선_30m`, `10선_1m`, `10선_5m`, `10선_30m`, `10선_close`, `국고2년_5m`, `국고2년_30m`, `국고2년_close`, `국고3년_5m`, `국고3년_30m`, `국고5년_5m`, `국고5년_30m`, `국고10년_5m`, `국고10년_30m`, `국고10년_close`, `국고20년_5m`, `국고20년_30m`, `국고30년_5m`, `국고30년_30m`

Jev response (원응답 JSON 중 4개 질문)

```
{"model": "jev-1.13.0",
 "answers": {
  "3선_1m": {"type": "choice", "choice": "약세", "confidence": 0.98, "probabilities":
  ↪  {"약세": 0.99, "강세": 0.01}},
  "3선_5m": {"type": "choice", "choice": "약세", "confidence": 0.99, "probabilities":
  ↪  {"약세": 0.99, "강세": 0.01}},
  "10선_1m": {"type": "choice", "choice": "약세", "confidence": 0.99, "probabilities":
  ↪  {"약세": 0.99, "강세": 0.01}},
  "10선_5m": {"type": "choice", "choice": "약세", "confidence": 0.99, "probabilities":
  ↪  {"약세": 0.99, "강세": 0.01}}
  ...(나머지 17개 중략) },
 "usage": {"input_tokens": 5380, "output_tokens": 779}}
```

Laya response — L0 zero-shot 확률 (// 뒤는 L1 fine-tuning seed 3개 평균)

```
"3선_1m": {"강세": 0.498, "약세": 0.502}   // L1 강세 0.496
"3선_5m": {"강세": 0.498, "약세": 0.502}   // L1 강세 0.497
"10선_1m": {"강세": 0.497, "약세": 0.503}   // L1 강세 0.497
"10선_5m": {"강세": 0.497, "약세": 0.503}   // L1 강세 0.497
```

Ground truth check — 질문 21개 (괄호 숫자 = 강세 확률, ✓ = 맞음)

질문	정답	Jev(J0)	Laya L0	Laya L1
3선_1m	약세	약세 0.01 ✓	약세 0.50 ✓	약세 0.50 ✓
3선_5m	약세	약세 0.01 ✓	약세 0.50 ✓	약세 0.50 ✓
3선_30m	약세	약세 0.01 ✓	약세 0.50 ✓	약세 0.50 ✓
10선_1m	약세	약세 0.01 ✓	약세 0.50 ✓	약세 0.50 ✓
10선_5m	약세	약세 0.01 ✓	약세 0.50 ✓	약세 0.50 ✓
10선_30m	약세	약세 0.01 ✓	약세 0.50 ✓	약세 0.50 ✓
10선_close	강세	약세 0.01	약세 0.50	약세 0.50
국고2년_5m	약세	약세 0.02 ✓	약세 0.48 ✓	약세 0.49 ✓
국고2년_30m	약세	약세 0.02 ✓	약세 0.49 ✓	약세 0.49 ✓
국고2년_close	강세	약세 0.04	약세 0.49	약세 0.49
국고3년_5m	약세	약세 0.03 ✓	약세 0.48 ✓	약세 0.49 ✓
국고3년_30m	약세	약세 0.03 ✓	약세 0.48 ✓	약세 0.49 ✓
국고5년_5m	약세	약세 0.02 ✓	약세 0.48 ✓	약세 0.49 ✓
국고5년_30m	약세	약세 0.02 ✓	약세 0.48 ✓	약세 0.49 ✓
국고10년_5m	약세	약세 0.02 ✓	약세 0.48 ✓	약세 0.49 ✓
국고10년_30m	약세	약세 0.02 ✓	약세 0.49 ✓	약세 0.49 ✓
국고10년_close	강세	약세 0.03	약세 0.49	약세 0.49
국고20년_5m	약세	약세 0.02 ✓	약세 0.49 ✓	약세 0.49 ✓
국고20년_30m	약세	약세 0.03 ✓	약세 0.49 ✓	약세 0.49 ✓
국고30년_5m	약세	약세 0.02 ✓	약세 0.48 ✓	약세 0.49 ✓
국고30년_30m	약세	약세 0.02 ✓	약세 0.49 ✓	약세 0.49 ✓
맞힌 수		18/21	18/21	18/21

B.3 Factor Reading — the Same Meeting's Statement

Input — factor question (score 5단계, 본문을 읽는 질문 7개 중 하나)

```
{
 "inflation": {
  "type": "score",
  "instructions": "이 결정문은 물가 상승 압력을 어떻게 평가하는가?",
  "criteria": [
   "물가 크게 둔화·목표 크게 하회",
   "물가 안정·목표 하회 우세",
   "목표 수준 근방에서 안정",
   "목표 상회·오름세 지속",
   "높은 오름세·상방 압력 매우 큼"
  ]
 }
}
```

Jev response — 요소별 기대 수준(0–4, 높을수록 매파)

```
{"growth": 3.98, "inflation": 3.11, "fx": 2.38, "finstab": 2.97, "guidance": 3.49,
↪  "uncertainty": 2.59, "overall": 3.27}
```

C Walk-Forward Samples (Experiment C)

회의마다 Laya 를 처음부터 다시 fine-tuning 했다. 아래 표의 한 행이 한 번의 학습이다(학습 로그 `train_logs.json` 과 예측 `predictions.csv` 에서 그대로).

시험 회의	학습 항목	보정 항목	온도(choice)	질문	정답 강세	답 강세	accuracy
2024-01-11	1,239	309	10.00	34	0.91	0.00	0.09
2024-02-22	1,266	316	10.00	27	0.67	0.67	1.00
2024-04-12	1,288	321	1.00	30	0.97	1.00	0.97
2024-05-23	1,312	327	8.04	31	0.61	0.23	0.16
2024-07-11	1,336	334	10.00	34	0.09	0.53	0.38
2024-08-22	1,364	340	10.00	24	0.88	0.00	0.12
2024-10-11	1,383	345	10.00	31	0.55	0.87	0.55
2024-11-28	1,408	351	10.00	35	0.97	0.69	0.71
2025-01-16	1,436	358	10.00	41	0.54	0.59	0.12
2025-02-25	1,468	367	10.00	28	0.29	0.50	0.79
2025-04-17	1,491	372	2.44	30	0.13	0.00	0.87
2025-05-29	1,515	378	10.00	33	0.61	0.67	0.45
2025-07-10	1,541	385	10.00	15	0.67	0.13	0.47
2025-08-28	1,553	388	10.00	30	0.30	0.00	0.70
2025-10-23	1,577	394	10.00	34	0.65	0.00	0.35
2025-11-27	1,604	401	10.00	42	0.21	0.02	0.76
2026-01-15	1,638	409	10.00	30	0.17	0.10	0.93
2026-02-26	1,662	415	1.00	38	1.00	1.00	1.00
2026-04-10	1,692	423	10.00	33	0.39	1.00	0.39
2026-05-28	1,719	429	10.00	38	0.11	0.42	0.47
2026-07-16	1,749	437	10.00	37	0.51	0.00	0.49
2026-08-27	1,779	444	9.50	26	1.00	1.00	1.00

온도가 상한 10.0 이면 보정이 «확률을 0.5 쪽으로 최대한 눌렀다» 는 뜻이다(학습 구간 마지막 20% 에서 방향 정보가 없었다).

Request meta — Kaggle 노트북의 학습 계획 한 줄(커널 A)

```
{"name": "WF_2026-01-15", "train_lt": "2026-01-15", "test": ["2026-01-15"], "seeds": [0]}
```

학습 = 회의일이 `train_lt` 보다 이른 사례 전부, 시험 = 그 회의의 E1·E2 사례. `torchrun --nproc_per_node=2 train_ddp.py MODEL OUT DATA 0 4`.

Laya training log — S25 seed 0 — 학습 회의일 < 2026-01-01, 시험 2026 6회의

```
n_train 1638 · n_calib 409 · temperature [10.0, 1.2, 1.2]
epoch 1: train_loss 0.667 · test_acc(기록용) 0.520 · test_logloss 0.696 · 480s
epoch 2: train_loss 0.682 · test_acc(기록용) 0.644 · test_logloss 0.687 · 974s
epoch 3: train_loss 0.696 · test_acc(기록용) 0.356 · test_logloss 0.724 · 1480s
epoch 4: train_loss 0.692 · test_acc(기록용) 0.356 · test_logloss 0.851 · 1977s
```

Laya training log — WF seed 0 — 같은 학습 자료, 시험 2026-01-15 회의만

```
n_train 1638 · n_calib 409 · temperature [10.0, 1.2, 1.2]
epoch 1: train_loss 0.708 · test_acc(기록용) 0.400 · test_logloss 0.762 · 465s
epoch 2: train_loss 0.652 · test_acc(기록용) 0.000 · test_logloss 0.992 · 932s
epoch 3: train_loss 1.601 · test_acc(기록용) 0.967 · test_logloss 0.132 · 1399s
epoch 4: train_loss 1.427 · test_acc(기록용) 0.933 · test_logloss 0.136 · 1858s
```

Ground truth check — 2026-01-15 회의 30문항

S25 seed 0: 0.00 · WF seed 0: 0.93. 학습 자료·설정·seed 가 같고 다른 것은 epoch 마다 기록용으로 평가한 시험 표본뿐이다 — 같은 학습이 같은 답을 내지 않는다(본문 §6).

D Full Configuration

항목	값
MAX_LEN, HEAD_MAX_LEN	4096, 256
MICRO_BATCH, GRAD_ACCUM	2, 1
GROUP_SIZE	4
LR_ENCODER, LR_HEAD	2.0e-5, 1.0e-4
SIGMA_START, SIGMA_END	0.4, 0.1
CALIB_FRAC, CALIB_MIN	0.2, 10
epoch · seed	4 · 0,1,2 (마지막 epoch 저장, 시험으로 epoch 를 고르지 않음)
보정	학습 구간 시간순 마지막 20% 로 온도 적합(choice · score · noul)
분할	회의 단위 — 학습 ≤ 2023-12-31 · 시험 2024-01-01 –
Jev	jev-latest(응답 jev-1.13.0) · 사례당 1회 호출 · 같은 입력 2회로 결정성 점검
기준선	B0 학습 구간 가족별 다수 · B1 결정 · B2 설문 대비 서프라이즈 · B3 기대 톤 서프라이즈

학습 스크립트 상수는 train_ddp.py 에서 그대로 읽었다.