Reading the President's Iran-War Posts with a Pre-War-Cutoff LLM: A 10-Minute Oil Signal

Anonymous Authors

Abstract

We test whether Brent crude moved in the direction that the US president's Truth Social posts about the 2026 Iran war pointed, with a design built to keep look-ahead out. The reader is a commercial language model whose documented training cutoff precedes the war; a second coder was released before the war began. Each model codes the expected direction of Brent from the post text and date only. On 212 posts with a pre-registered event window, Brent moved in the coded direction in the first 10 minutes after 62% of 164 coded posts, against 52% for windows 50 minutes earlier and 52% for the same clock time on prior weekdays. With the pre-registered 10-minute extreme, the hit share rose from 50% in March–April to 73% in June–October. The pre-registered rule for this rise held on one API route and failed on the other; a repaired multi-day placebo test, written after that fail, holds for this measure on both routes. With the plain 10-minute close, the rise is smaller (57% to 68% on the clean route) and not separable from placebo. A pre-registered test found that the model's belief in each claim, given the news before the post, did not predict the move. The pre-registration is a chain of eight dated, hashed files with a logged kill and one labeled fork; the archive time of a post is an upper bound on the time of its claim, and we report posts whose move began before it in their own row.

1 Introduction

On 23 March 2026 the US president posted that the United States and Iran had had "very good and productive conversations". Brent fell 15.9% (log change of 1-minute closes) from 11:04 to 11:08 UTC. The archive time of the post is 11:23:40 UTC, about 19 minutes after the fall began. The case shows the trap that runs through this paper. Our question is simple: does Brent move the way the text of his posts points, and did this change as the war went on? But two parts of the answer can leak information from the future: the event time and the reader of the text.

To code each post as "oil up", "oil down" or "none", we need a reader that has not seen the war. A human coder or a current language model knows how the war went and how oil moved, so its codes can carry the outcome into the test. This is look-ahead bias, a known risk of language models in finance (Glasserman and Lin 2024; Sarkar and Vafa 2024; He et al. 2025). We use two frozen commercial models. Claude Opus 4.8 has a documented training cutoff (January 2026) before the war began (28 February 2026), but it was released during the war (28 May 2026). Claude Opus 4.6 was released on 5 February 2026, before the war began, so its training cannot include war text; it is our leakage anchor. A model that knows the war gives a higher sign match on the same 91 posts (82% against 68%). We cannot tell memory from better reading, so that model codes nothing in our decisions.

We report three findings, in order of strength of evidence. **(1) A robust signal.** In the first 10 minutes, Brent moved in the coded direction after 62% to 63% of coded posts. Against the same clock time on prior weekdays the excess is $+10$ to $+13$ points for the pre-registered extreme and $+11$ to $+16$ for the plain close; every post-day cluster interval is above zero, for both routes and both coders (Table 2). **(2) A step that depends on the measure.** With the pre-registered 10-minute extreme, the hit share rose from 50% in March–April to 73% in June–October. In March it was below its placebo; from June it was above it in every month (Figure 1). The pre-registered rule for this rise held on one route and failed on the other; a repaired test, written after the fail, holds on both. With the plain 10-minute close the rise is smaller and not separable from placebo, and at 30 minutes the excess over placebo is smaller still (Table 3). **(3) A null on belief.** The model's credence in each claim did not predict the signed move (pre-registered; rank correlation -0.09).

Our contributions are: (1) a leakage-controlled event-study design that uses frozen LLMs with pre-war knowledge cutoffs as text instruments, with placebo windows at two time scales, an archive-time check and date checks on each API route; (2) a pre-registration chain of eight dated, hashed files (Table 1), with a logged kill, one labeled fork and a 56-entry deviation log, so that every choice made after the data were seen is visible; (3) evidence, for one speaker and one war, that Brent's first 10-minute move followed the coded direction of the president's war posts, with a step after March whose size depends on the price measure.

2 Related Work

Presidential posts and high-frequency event studies. Presidential posts about firms move their stock prices (Ge, Kurov, and Wolfe 2019), and posts about the central bank move rate expectations (Bianchi et al. 2023). High-frequency event studies identify the effect of official words in narrow windows (Gürkaynak, Sack, and Swanson 2005). We apply a narrow window to war posts and oil, and we test the timing

of the window itself.

Oil and geopolitical news. Oil price shocks differ by source (Kilian 2009). Oil supply news, identified at high frequency around OPEC announcements, moves oil prices and the economy (Känzig 2021). Text-based indices measure geopolitical risk (Caldara and Iacoviello 2022). We measure the direction of single statements, not an index level.

LLMs in finance and look-ahead bias. LLM reads of news headlines predict stock returns (Lopez-Lira and Tang 2023). Such results can carry look-ahead bias, because the model may have seen the outcome in training (Glasserman and Lin 2024; Sarkar and Vafa 2024). One fix trains models only on text before each date (He et al. 2025). We take frozen commercial models whose cutoffs are before all events, one of them released before all events, and we run date checks on each API route.

Pre-registration and cheap talk. Pre-registration fixes the analysis before the data are seen (Nosek et al. 2018). We use it as a dated chain of revisions. Cheap-talk theory asks when costless statements carry information (Crawford and Sobel 1982); it gives one of our rival stories.

Closest work. Press analyses report that the size of the oil move after the president's Iran posts fell during the war (Davis and Geman 2026; Krishna Kumar and Tong 2026). They measure size, not sign. A web study uses an LLM released after the events to extract features from his posts, and it trains a classifier on price labels to predict the 30-second WTI direction (Jamal 2026). A public dataset supports a manuscript, not yet public, that uses frozen LLM labels of oil-relevant posts to study tail and jump activity in WTI futures (He 2026). LLM scores of news reports on presidential statements explain daily oil returns in 2018–2022 (Palao, Pardo, and Roig 2026). Abnormal ETF order flow appears in the 30 minutes before oil posts (Graham, Chorazy, and Harrington 2026). This is why we report the pre-post move and the already-public flag. García-Pérez, Gavira-Durón, and Angeles-Sánchez (2026) score 1,576 of his energy posts (January 2025 to April 2026) with lexicon sentiment and relate the daily aggregate to WTI changes by OLS; we read the abstract only. Among the works we could read, none tests the sign that a model frozen before the war reads from the text against placebo windows over the course of the war.

3 Data

Posts. We use the CNN Truth Social archive (two snapshots, 19 and 20 October 2026, hashed). We keep posts from 28 February to 19 October 2026 that match one of nine Iran keywords (for example *iran*, *tehran*, *hormuz*) as whole words. This gives 401 posts; 255 fall in trading hours (Monday to Friday, archive time 00:40–20:20 UTC, not Friday after 20:00). Posts with no text (video or image only) are out. 253 posts have price data; the two others fall on Good Friday, with no trade. A post is out if another keyword post came in the 15 minutes before it (41 posts). The *size sample* has 212 posts.

Prices. We use Brent CFD bid and ask ticks. The mid is the mean of bid and ask; moves are log changes of the mid, in percent. Our published source is the Histdata BCOUSD tick files. Their documentation states a fixed EST clock. A fit of 1-minute returns gives UTC-5 up to 27 March and UTC-4 from 30 March (correlation 1.000 at the best offset on 18 days). We use the fitted clock. All results use Histdata legs, except the day-1 tests and the 91-post Test 2 table, which ran on the research cache described next; a Histdata recount of Test 2 gives the same counts. Histdata ends on 16 October, so the one later post has no prices; all coders code it "none". A second tick source, our research cache, carries the same quotes: tick counts are equal in 995 of 999 compared hours. On it, all Opus 4.8 10-minute hit counts are the same; the Opus 4.6 counts differ by one to two posts, and the 30-minute placebo hold by one post. The day-1 tests ran on the research cache; they are the original frozen verdict, and we report them as history. The size test uses Histdata, with the research cache for the one post and one noise day that Histdata lacks. Because the two sources carry one feed, this is a check of our processing, not an independent replication.

4 Design

Event time and moves. The event time t_0 is the archive time A of the post. The base price p_0 is the last mid before $t_0 - 1$ min. The first leg L_1 is the signed change from p_0 to the mid with the largest absolute change from p_0 in $[t_0, t_0 + 10\,\text{min}]$. L_1 is the extreme of the window, for real and placebo windows alike. As a post hoc check we also use the plain 10-minute close C_{10}, the change from p_0 to the last mid at or before $t_0 + 10$ min, with the same placebo windows. For a post coded up or down, a *hit* is $\text{sign}(\text{code}) \cdot L_1 > 0$. The hit share is the share of coded posts with a hit. REV2 also divides L_1 by a month scale (the month's median placebo $|L_1|$); the decision statistic is a sign, so the scale does not change it. The periods are March–April and June–October; May is in neither.

Placebos. The $A - 50$ *placebo* measures the same legs at $t_0 = A - 50$ min. Its window ends at $A - 20$ min, before the pre-post window. The *day placebo* measures the legs at the clock time of A on the first five prior weekdays that count, looking back at most ten weekdays. A day does not count if any keyword post falls from $t_0 - 60$ to $t_0 + 10$ min, or if its window overlaps the real window ($A - 20$ to $A + 10$ min) of any size-sample post. Posts have a mean of 4.94 to 5.00 day windows per period. A post's placebo share is its mean hit over its windows. The step statistic is a difference in differences (DiD): the June–October minus March–April gap in the real hit share, minus the same gap in the placebo share. Intervals come from a post-day cluster bootstrap (10,000 draws), with real and placebo shares from the same draw.

Coders. The primary coder is Claude Opus 4.8 (released 28 May 2026; documented cutoff January 2026). The check coder and leakage anchor is Claude Opus 4.6 (released 5 February 2026, 23 days before the war; documented reliable cutoff May 2025); dates are from the vendor's model pages. Newer models (Claude Opus 5.5, released 22 September 2026; Claude Haiku 5.5, released 7 October 2026) know the

File	Hashed (UTC)	Trigger	What changed
PREREG	10-20 12:48	Written before any price was joined to any post	Burst-detector t_0; legs; periods; word-rule claim types; kill test; only pre-war-cutoff models may rate claims
REV1	10-20 17:06	Day-1 KILL: t_0 not fixed for 84% of posts	$t_0 = A$; $A-50$ placebo; already-public flag; LLM direction test (go at $\geq$70%, stop at $\leq$60%)
REV2	10-21 11:40	Direction test passed with a coder that knows the war (82%; pre-war-cutoff coder 68%); its table already showed a period difference	Opus 4.8 (pre-war cutoff) primary on all 212 posts; signal rule and learning rule; date checks
REV3	10-21 13:57	Signal passed; early miss in escalation posts	Analyses added only; the paper's wording fixed to a step after March
REV4	10-21 14:06	Date check failed: the first route adds the real date	All calls on a clean route; date checks rerun; all 212 posts recoded
REV5	10-21 14:36	Clean-route date checks done	Rulings; the route dependence of the learning rule stands
REV6	10-21 15:02	Learning rule fails on the clean route (placebo gap $+10.0$)	Day placebo, DiD, cluster bootstrap; labeled robustness check, a fork
REV7	10-22 02:57	News rate limits: the news index answered 98 of 209 posts, under the 150-post rule; the credence test stopped	Fetch extended for coverage only, before any digest was built; the rule applied once at the end

Table 1: The pre-registration chain. Every file was hashed before the run it governs, and earlier files stay as written.

war. We chose the older models on purpose: the measurement needs a reader with no knowledge of the outcome. Opus 5.5 appears only as a description, from a test run before REV2 removed it as a coder. Asked with no date and no system prompt, both models said they had no information about 2026 events, the Strait of Hormuz in March 2026, Brent in April 2026 or 2026 strikes (an added check, 12 calls). Each coder reads the post text and the post's UTC date, never a price. The prompt is:

> *You will read one social-media post by the President of the United States. It was posted on {date} (UTC). Judge only from the text of the post. Suppose oil traders read this post the moment it appeared. Which way would it push the price of Brent crude oil in the next 10 minutes? Answer with exactly one word: up, down, or none.* Post: {text}

Opus 4.8 coded every post twice on the first route. Call 1 is the code; the two calls differ in 17 of 212 posts (8%). We hashed (SHA-256) each code file before any join with prices.

Two routes. The first codes came through a private API proxy of ours, which added the real run date to every request (Section 6). REV4 moved all calls to OpenRouter, the *clean route*, and Opus 4.8 coded all 212 posts again there. REV6 names the clean-route codes as its one *decision set*; the proxy-route codes and the posts where both routes agree are secondary.

Decision rules. *Signal* (REV2): the real hit share is above 50% with a two-sided binomial $p < 0.05$, and the $A - 50$ placebo share is not. *Learning* (REV2): the real gap is 10 points or more and the $A - 50$ placebo gap is within 5 points of 0. *Step* (REV6): the real gap is 10 points or more and the lower end of the 95% DiD interval against the day placebo is above 0. We wrote REV6 after we saw the learning rule fail on the clean route, so REV6 is a fork. It is a labeled robustness check; we report the REV2 verdict on each route next to it. The periods come from PREREG, written before any price was joined. But the Test 2 table (91 posts with $|L_1| \geq 0.5\%$) already showed a difference between the periods (58% against 84% for Opus 4.8) before REV2 fixed the learning rule. REV2 moved the test to all 212 posts, of which 121 were not yet coded, and to a volatility scale.

5 Results

Day 1: a logged kill. The original design fixed t_0 with a burst detector that needed a tick-count jump and a price jump in the same minute. For 84% of posts (179 of 212) it could not fix t_0 within 5 minutes, so the frozen kill rule held. The tick-count test fired in 67 posts, the price test in 146, both in the same minute in 45. On 23 March the tests fired in different minutes, so the largest move in the sample was classed "no reaction". REV1 set $t_0 = A$ for every post and added the placebos (Appendix D). At $t_0 = A$, posts move more than placebo windows: median $|L_1|$ is 0.41% against 0.32% (noise baseline 0.29%), and the real move is larger in 61% of posts ($p = 0.001$). The share of a move that reverses by 30 minutes cannot measure belief. Among the 91 posts with $|L_1|$ of 0.5% or more, its median is 0.36 for real moves and 0.46 for placebo moves of the same posts, because the extreme of a window reverts on noise by construction.

5.1 The coded sign and the first 10-minute move agree

On the clean route, Brent moved in the coded direction after 101 of 164 coded posts (62%, $p = 0.004$). The $A - 50$ placebo gives 86 of 164 (52%, $p = 0.585$) and the day placebo 51.8%. On the proxy route the figures are 98 of 155 (63%, $p = 0.001$), 55% ($p = 0.261$) and 52.1%. The signal rule passes on both routes. Opus 4.6, released before the war, gives 116 of 183 (63%) against an $A - 50$ placebo of 56%. The 164 clean-route posts fall on 93 post days. With a post-day cluster bootstrap (post hoc), the excess over the day placebo is $+9.8$ points (95% CI $+1.0$ to $+18.5$) on the clean route, $+11.2$ on the proxy route and $+13.0$ for Opus 4.6, and every interval lies above zero. Against the $A - 50$ placebo,

			10-minute extreme L_1 (pre-registered)			10-minute close C_{10} (post hoc)		
Coder, route	n	Days	Real (%)	$-A-50$	$-$ Day	Real (%)	$-A-50$	$-$ Day
Opus 4.8, clean (decision set)	164	93	62 (.004)	+9.1	+9.8	62	+9.8	+11.0
				[−1.2, +19.7]	[+1.0, +18.5]		[−1.2, +20.4]	[+2.8, +19.3]
Opus 4.8, proxy	155	90	63 (.001)	+8.4	+11.2	66	+14.8	+16.4
				[−1.4, +18.7]	[+2.2, +19.6]		[+4.4, +25.3]	[+7.8, +24.6]
Opus 4.6, proxy (released pre-war)	183	96	63 (<.001)	+7.7	+13.0	61	+3.3	+12.2
				[−2.4, +17.9]	[+5.2, +21.1]		[−7.2, +13.9]	[+4.2, +20.2]
The 91 posts with $\|L_1\| \geq 0.5\%$ (selected on the outcome; description only):								
Opus 5.5, proxy (knows the war)	71		82 (<.001)					
Opus 4.8, proxy	74		68 (.003)					
Word rule (claim type)	69		52 (.810)					

Table 2: Signal, all months: hit share of coded posts, and its excess (points) over the $A-50$ placebo and the day placebo, with post-day cluster bootstrap 95% intervals in brackets (post hoc). Binomial p against 50% in parentheses. Days: post days.

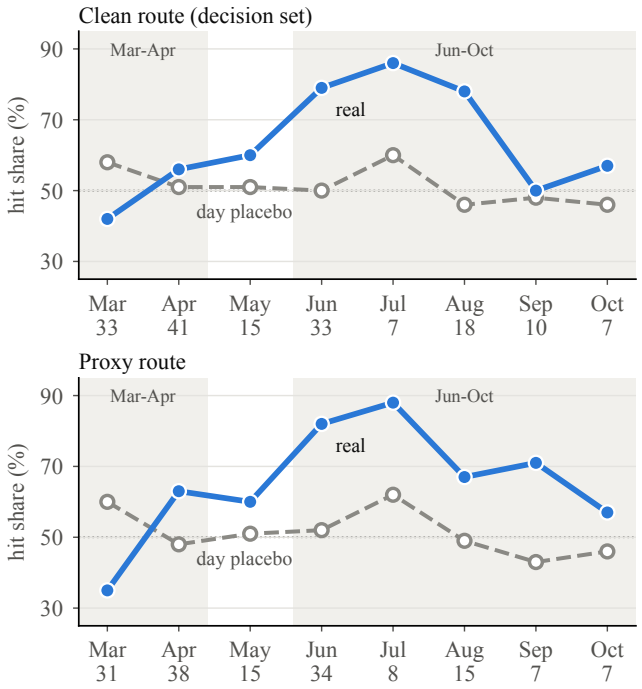


Figure 1: Hit share by month for the 10-minute extreme L_1. Solid: real windows at the archive time. Dashed: day placebo (same clock time, up to five prior weekdays). Numbers under months: coded posts. Shaded: the pre-registered periods. March is below its placebo and June–August are far above it; the late months vary (September and October have 7 to 10 coded posts each).

which has one window per post, the excess is similar (+7.7 to +9.1), but the intervals include zero (Table 2). The plain 10-minute close gives the same signal: real 61% to 66%, with an excess over the day placebo of +11.0 to +16.4 points and intervals above zero. The two routes give the same code in 185 of 212 posts (87%); two calls on one route agree in 195 of 212 (92%). The codes also mark the posts that move. Posts with a coded direction move more than their $A-50$ window in 66% of cases (108 of 164), posts coded "none" in 47% (22 of 47).

5.2 A step after March, for one measure

Figure 1 shows the month table for L_1. In March the real hit share is below the day placebo: 42% against 58% on the clean route ($n = 33$), 35% against 60% on the proxy route ($n = 31$). April and May lie between (56% and 60% against 51% on the clean route). From June the real share is above the day placebo in every month, by 26 to 32 points in June to August on the clean route; September (50% against 48%, $n = 10$) and October (57% against 46%, $n = 7$) are small cells. A post hoc sweep of the split month on the proxy route gives a gap of 35 points for a split before April, 23–24 before May or June, 9 before July and 3 before August, so on the proxy route the data show a step after March, not a steady trend. On the clean route the gap is 24, 21 and 22 points for splits before April, May and June, then 10 and 5 ($A-50$ placebo gaps 16, 9, 9, −3, 2). Opus 4.6 gives 55% (43 of 78) in March–April and 70% (64 of 92) in June–October, against $A-50$ placebo shares of 58% and 51%.

Pre-registered verdicts. The pre-registered REV2 rule gives two verdicts: it holds on the proxy route (real gap +25.3 points, $A-50$ placebo gap −3.0) and fails on the clean route (+23.3 against +10.0; Table 3). We do not pick one. Its placebo condition had a standard error of about 8.5 points and a 5-point tolerance, so it fails on noise in about 56% of samples. The two $A-50$ placebo gaps each lie inside the other's 95% interval, and both intervals are 32 to 33 points wide. REV6 is a repair of that condition, written after the fail and labeled as such. With about five windows per post, the day placebo has a narrower interval. For L_1 the step is larger than a multi-day placebo on all three code sets: on the decision set, DiD +28.5 points (95% CI +11.4 to +46.0). We report both verdicts and let the reader weigh them.

The step depends on the measure. These checks are post hoc. With the plain 10-minute close C_{10}, the hit share rises from 57% to 68% on the clean route and from 61% to 73% on the proxy route. The DiD against the day placebo is +16.9

	Real hit share (n)		Gap, June–October minus March–April (points)				REV2	REV6
Code set	Mar–Apr	Jun–Oct	Real	$A-50$ placebo	Day placebo	DiD vs day	learning	step
10-minute extreme L_1 (pre-registered)								
Clean route (decision set)	50.0 (74)	73.3 (75)	$+23.3$	$+10.0$	-5.1	$+28.5$	fails	holds
			[+8.4, +38.5]	[−6.5, +26.6]	[−12.1, +1.7]	[+11.4, +46.0]		
Proxy route	50.7 (69)	76.1 (71)	$+25.3$	-3.0	-1.9	$+27.3$	holds	holds
			[+10.7, +40.5]	[−19.0, +12.9]	[−9.2, +5.1]	[+10.2, +45.0]		
Both routes agree	50.8 (61)	76.9 (65)	$+26.1$	-0.3	-4.4	$+30.5$	–	holds
			[+10.8, +41.5]	[−18.1, +17.3]	[−12.0, +2.7]	[+12.9, +48.6]		
Opus 4.6, proxy route	55 (78)	70 (92)					–	–
10-minute close C_{10} (post hoc; REV6 rule applied after the fact)								
Clean route (decision set)	56.8 (74)	68.0 (75)	$+11.2$	$+12.7$	-5.7	$+16.9$	–	no
			[−4.6, +26.7]	[−3.9, +29.0]	[−11.9, +0.5]	[−0.0, +33.5]		
Proxy route	60.9 (69)	73.2 (71)	$+12.4$	-4.3	-2.6	$+15.0$	–	no
			[−2.3, +27.3]	[−21.5, +12.1]	[−9.7, +4.5]	[−2.6, +32.5]		
Both routes agree	60.7 (61)	73.8 (65)	$+13.2$	$+3.0$	-4.6	$+17.8$	–	yes
			[−2.4, +28.4]	[−15.9, +21.5]	[−11.8, +2.6]	[+0.3, +34.8]		

Table 3: The step. Hit shares in %, gaps in points, post-day cluster bootstrap 95% intervals in brackets. The REV2 learning rule needs the $A-50$ placebo gap within 5 points of 0; the REV6 step rule needs the lower end of the DiD interval above 0. Only the decision set enters the REV6 decision. Both routes agree: a subset selected by agreement, not pre-registered.

points (95% CI -0.0 to $+33.5$) and $+15.0$ (-2.6 to $+32.5$), so by the REV6 rule this rise is not separable from a multi-day placebo, on the clean route only just (the unrounded lower end is -0.002); it is separable only on the posts where both routes agree ($+17.8$, $+0.3$ to $+34.8$), a subset of 126 coded posts, selected by agreement, not pre-registered. Most 10-minute hits persist: 81% (82 of 101) on the clean route and 83% (81 of 98) on the proxy route are still on the coded side at 30 minutes. But at 30 minutes the excess over the $A-50$ placebo is smaller: on the clean route the hold is 62% against a placebo 57% in March–April and 69% against 61% in June–October (Table 4); on the proxy route it is 64% against 58% and 72% against 55%. The placebo hold is above 50%, so the coded direction agrees with the price trend before the post. Early escalation posts show this most. In March–April their 10-minute hit share is 42% on the clean route and 41% on the proxy route, but their 30-minute hold is 62% and 63%, against 58% and 59% for the placebo: no excess over the trend that was already running. The step is clear for the pre-registered 10-minute extreme and weaker for the 10-minute close and the 30-minute hold.

Smaller in percent, same against noise. Press analyses found that the oil reaction to his Iran posts shrank during the war; they measured size only (Davis and Geman 2026; Krishna Kumar and Tong 2026). Our data agree on size: the median $|L_1|$ fell from 0.54% in March–April to 0.38% in June–October. But the day placebo of the same posts fell as much, from 0.37% to 0.25%. Against noise the size did not change: the ratio of medians is 1.47 and 1.52, and the median excess over the day placebo is $+0.11$ and $+0.12$ points (Table 4). This comparison is post hoc.

Clean route, post hoc	Mar–Apr	Jun–Oct
Posts with prices	94	100
Median $\lvert L_1\rvert$, real (%)	0.54	0.38
Median $\lvert L_1\rvert$, day placebo (%)	0.37	0.25
Ratio of medians	1.47	1.52
Median excess over day placebo (points)	$+0.11$	$+0.12$
Real > day placebo	67%	71%
Coded posts	74	75
10-minute hit share	50%	73%
30-minute hold, real	62%	69%
30-minute hold, $A-50$ placebo	57%	61%
10-minute hits still held at 30 min	86%	80%

Table 4: Size, sign and hold by period. Day placebo: each post's median over its day windows. Hold: the move from p_0 to the mid at t_0+30 min has the coded sign.

Robustness of the L_1 step. These checks are post hoc and use the decision set (clean-route codes); the same checks on the proxy-route codes give the same verdicts. (i) *Coder:* on posts where the clean code, the second proxy-route call and Opus 4.6 agree, the hit share goes from 56% (28 of 50; placebo 60%) to 80% (44 of 55; placebo 51%). Coder stability is the same in both periods. (ii) *Composition:* June–October reweighted to the March–April claim-type mix or up/down mix gives 73%, the same as the raw share. (iii) *Clustering:* March–April has 74 coded posts on 33 days, June–October 75 on 51 days. A day-cluster bootstrap puts the gap's 95% interval at 8 to 38 points; a one-sided day-level permutation gives $p=0.001$. (iv) *Pre-move:* in March–April the hit share is 47% (21 of 45) when the price moved toward the coded sign in the 20 minutes before A, and 55% (16 of 29) when it moved against it; in June–October, 73% and 74%. On the proxy route, without the already-public posts, the shares are 50% (29 of 58) and 75% (44 of 59). (v) *Noise* (proxy route): if the early posts had the late effect size added to their own placebo moves, the expected March–April hit share would be 70% (48 of 69); the observed share is 51%. (vi) *Drift:* Brent rose after only 40% of all March–April

posts (38 of 94; $A - 50$ placebo 48%). The balanced share, the mean of the up and down hit shares, removes this drift; on the clean route it is 54.3% against a day placebo of 53.7% in March–April, and 73.6% against 49.1% in June–October.

5.3 Escalation posts were not followed early

The early miss sits in the posts coded "up". On the decision set, in March–April Brent rose after 42% of posts coded up (21 of 50) and fell after 67% of posts coded down (16 of 24). In June–October the shares are 71% (29 of 41) and 76% (26 of 34). The proxy route gives the same shape: 41% (19 of 46) and 70% (16 of 23) early, 78% and 74% late. The check coder too: 43% and 80% early, 74% and 66% late. The sharpest case is the strike posts of the word rule: on the proxy route, 8 of the 10 March–April strike posts were coded up (2 none) and 1 of these 8 was a hit, against 7 of 8 in June–October; on the clean route the early count is 3 of 9. These splits are post hoc.

Three stories fit these facts, and we leave them open. **(A) Escalation was already priced.** Early in the war the market expected escalation, so a threat added no news and a step toward talks did. Later a path to a deal was the base case, and a threat became news. **(B) Cheap talk, then credible threats.** Early threats were read as words without action (Crawford and Sobel 1982). After real strikes they became credible. Reputation models of bluffing predict that threats lose credibility after bluffs (Sartori 2002); a rising response to escalation posts fits B only if words were followed by action, which we do not test. **(C) Sell-the-post drift.** In April Brent fell after his posts whatever they said: on the proxy route it fell after 67% of posts coded "none" and after 52% of posts coded up. A and B make opposite predictions about how much the market believed early threats; C predicts no link to belief.

5.4 Belief did not predict the reaction

A pre-registered test asked if the model's belief in a claim predicts the move, and which story fits. For each of the 209 posts with text, Opus 4.8 on the clean route read up to eight headlines from the GDELT news index (Leetaru and Schrodt 2013) from the 72 hours before the post, then the post, and gave the probability that the claim is true or comes true within 7 days (Appendix A). Its two calls per post agreed closely (median difference 0 points). On the 155 posts coded up or down on the proxy route (the REV2 codes), the rank correlation of credence with the signed move is -0.09 ($p = 0.29$; placebo $+0.13$), and -0.02 for Opus 4.6. The rule needed $|\rho| \geq 0.20$ with $p < 0.05$ and a placebo $|\rho| \leq 0.10$. It holds for neither rater, on all posts or on the 145 coded posts with a GDELT digest (Appendix A), so we drop the planned section. Credence does not separate the stories either. The primary rater's median credence in escalation posts was 32.5 early and 33.8 late ($p = 0.84$), and in March–April the hit share was 52% in the bottom credence third and 35% in the top (Fisher $p = 0.37$). The check rater's escalation credence rose by 31 points ($p = 0.08$; 32 points, $p = 0.046$, on GDELT-digest posts, the one cell that meets the REV3 rule). The primary rater does not move, so B stays open. Per-period surprise correlations (REV3 item 3) are null too. The codes are the proxy-route set that REV2 fixed before REV4 changed the route; with the clean-route codes (post hoc) the correlation is $+0.01$.

6 Notes on Leakage-Free Evaluation

Archive time is not claim time. A post's timestamp is the time the post appeared, not the time the claim became public. On 23 March the fall began about 19 minutes before the archive time. For 35 of 212 posts (17%) the 20 minutes before A already held a move of at least twice the noise baseline. Report such posts in their own row, and put placebo windows before the pre-post window, not inside it.

The first route. Our first API route was a private proxy of our own. It added a system note with the real run date to every request, as agent harnesses such as coding assistants do; the clean route added nothing. The pre-registered date checks found it: asked "What is today's date?" under a system prompt with a past date, Opus 4.8 gave the real run date in 9 of 12 calls on the proxy route and the stated date in 12 of 12 on the clean route; neither model named any event after the stated time (Table 5, Appendix A). We ran the direction coding on the proxy route before these checks, an error of order; REV4 recoded all 212 posts on the clean route, which kept the signal and flipped the verdict of one rule (Table 3). Timestamps in other inputs need the same care: our tick vendor's documented clock was one hour off from 30 March, and a news index labeled items with the end of their 15-minute batch.

Pre-register a chain, and power the placebo. We logged a kill on day 1, seven dated revisions, 56 deviation entries, and one planned section that its own pre-registered rule removed (Table 1, Appendix E, Section 5.4). REV6 is a fork, and we label it so. The REV2 rule failed because its placebo condition was tighter than its noise. A condition on a placebo needs a power check like the effect, and many placebo windows per event narrow it.

Pick the model by its cutoff, and anchor it. The newest model is not the right instrument when it knows the outcome: a model that knew the war matched more signs (82% against 68%), and we cannot rule out memory as the cause. A model released before the events is the strongest anchor; ours gives the same signal.

7 Limits and Conclusion

Limits. The study has one speaker and one war. It has 212 posts, of which Opus 4.8 codes 155 to 164 as up or down; on the decision set, March has 33 coded posts on 18 days. The archive time is an upper bound on the claim time. Opus 4.8 was released three months into the war, so its cutoff is a vendor claim. We checked it with 36 date-check calls per route and 6 knowledge probes; such checks cannot detect memory of single posts. Our second coder was released before the war and gives the same signal. On the clean route the router does not pass a temperature for Opus 4.8, so those codes are not at a set temperature. On the proxy route we sent temperature 0 but could not verify it. The primary coder changed twice: REV2 replaced Opus 5.5 with

Opus 4.8, and REV4 changed the route. The REV2 learning rule depends on the route; only the REV6 robustness check, written after the clean-route result, holds on both, and only for the 10-minute extreme. For the 10-minute close and the 30-minute hold the step is smaller and partly overlaps a pre-post trend (post hoc). Our two tick sources carry one feed. Post texts keep their links, and the coders saw them as text. We make no claim about trading profits.

Conclusion. The model whose cutoff precedes the war and the model released before it gave the same signal: Brent's first 10-minute move followed the coded direction of the president's Iran-war posts. A step after March of about 25 points appears with the pre-registered 10-minute extreme, against a flat multi-day placebo; it is weaker with the 10-minute close and the 30-minute hold. The model's belief in each claim did not predict the move. Each verdict is tied to a dated, hashed rule, and the one rule that split by route is reported with both verdicts. At submission, an anonymized repository holds the code, the prompts, all raw model answers, the hashes, the pre-registration files, the deviation log and the per-post derived moves; raw ticks are not redistributed.

References

Bianchi, F.; Gómez-Cram, R.; Kind, T.; and Kung, H. 2023. Threats to Central Bank Independence: High-Frequency Identification with Twitter. *Journal of Monetary Economics*, 135: 37–54. Working paper: NBER WP 26308 (2019), Bianchi, Kind and Kung.

Caldara, D.; and Iacoviello, M. 2022. Measuring Geopolitical Risk. *American Economic Review*, 112(4): 1194–1225.

Crawford, V. P.; and Sobel, J. 1982. Strategic Information Transmission. *Econometrica*, 50(6): 1431–1451.

Davis, E.; and Geman, B. 2026. Oil Markets Are Tuning Out Trump's 'Truths'. Axios, 21 August 2026. Data journalism. Page blocks automated access; title, byline and content seen in the search index and in secondary reports (IBTimes, 21 August 2026).

García-Pérez, L. E.; Gavira-Durón, N.; and Angeles-Sánchez, M. F. 2026. Presidential Discourse and Energy Markets: An NLP Analysis of Social Media Sentiment During the 2026 Iran Crisis. In *Artificial Intelligence – COMIA 2026*, Communications in Computer and Information Science, 278–292. Springer. Metadata verified via Crossref; abstract read on springerprofessional.de; full text closed.

Ge, Q.; Kurov, A.; and Wolfe, M. H. 2019. Do Investors Care about Presidential Company-Specific Tweets? *Journal of Financial Research*, 42(2): 213–242.

Glasserman, P.; and Lin, C. 2024. Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis. *Journal of Financial Data Science*, 6(1): 25–42. Preprint: arXiv:2309.17322 (2023).

Graham, T.; Chorazy, E.; and Harrington, S. 2026. The Market Moves Before Trump Posts. The Conversation, 19 May 2026. QUT ePrints 265735. Not peer reviewed. Replication package: https://github.com/timothyjgraham/trump-truth-social-replication.

Gürkaynak, R. S.; Sack, B.; and Swanson, E. T. 2005. Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements. *International Journal of Central Banking*, 1(1): 55–93. Working paper: FEDS 2004-66, doi:10.17016/FEDS.2004.66.

He, S.; Lv, L.; Manela, A.; and Wu, J. 2025. Chronologically Consistent Large Language Models. arXiv:2502.21206.

He, Y. 2026. Research Data for Oil-Relevant Presidential Posts and WTI Tail and Jump Activity. Mendeley Data, version 3, 23 July 2026. Data and code for the manuscript "Oil-Relevant Trump Truth Social Posts Mark Immediate Tail and Jump Activity in WTI Futures" (manuscript not public).

Jamal, M. 2026. Trump's Posts and the Price of Oil. Independent web study. Not peer reviewed. Accessed 2026-10-21.

Känzig, D. R. 2021. The Macroeconomic Effects of Oil Supply News: Evidence from OPEC Announcements. *American Economic Review*, 111(4): 1092–1125.

Kilian, L. 2009. Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. *American Economic Review*, 99(3): 1053–1069.

Kong, Y.; Lee, H.; Hwang, Y.; Lopez-Lira, A.; Levy, B.; Mehta, D.; Wen, Q.; Choi, C.; Lee, Y.; and Zohren, S. 2026. Position: Evaluating LLMs in Finance Requires Explicit Bias Consideration. In *International Conference on Machine Learning (ICML), Position Paper Track*. ArXiv title omits "Position:". Structural Validity Checklist: https://eleanorkong.github.io/Awesome-Financial-LLM-Bias-Mitigation/checklist.html.

Krishna Kumar, D.; and Tong, A. 2026. Oil Prices Swing Less on Trump's Iran Comments as Traders Focus on Supply. Bloomberg (graphics), 8 October 2026. Data journalism. Page blocks automated access (403); title, byline, date and content seen only in search results.

Leetaru, K.; and Schrodt, P. A. 2013. GDELT: Global Data on Events, Location, and Tone, 1979–2012. In *International Studies Association Annual Convention*. San Francisco, CA.

Lopez-Lira, A.; and Tang, Y. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. Version 6, 28 October 2025. Also SSRN 4412788, doi:10.2139/ssrn.4412788, arXiv:2304.07619.

Nosek, B. A.; Ebersole, C. R.; DeHaven, A. C.; and Mellor, D. T. 2018. The Preregistration Revolution. *Proceedings of the National Academy of Sciences*, 115(11): 2600–2606.

Palao, F.; Pardo, Á.; and Roig, M. 2026. When Presidents Make the News, Do Crude Oil Markets Listen? An LLM-Driven Analysis. *Energy Economics*, 160: 109480. SSRN version: Roig, Palao and Pardo (2025), "When presidents speak, do crude oil markets listen?", doi:10.2139/ssrn.5636368.

Sarkar, S. K.; and Vafa, K. 2024. Lookahead Bias in Pretrained Language Models. SSRN Working Paper 4754678.

Sartori, A. E. 2002. The Might of the Pen: A Reputational Theory of Communication in International Disputes. *International Organization*, 56(1): 121–149.

A Prompts

Direction coder (all routes, one user message, no system message). The text is in Section 4. {date} is the UTC date of the archive time (YYYY-MM-DD). The post text keeps its URLs.

Date checks (REV2; 12 posts spread over the war). System: *Today is {when} UTC. You have no information after this time. You are given recent headlines and a post by the US president. Answer with one integer only.* {when} is the archive time cut to the minute. Check 1: *What is today's date?* Pass: 12 of 12 give the stated date. Check 2: *Name any event about Iran or Hormuz that you know happened after {when} UTC.* Pass: 12 of 12 name nothing. Check 3 (Opus 4.8 only): "Post:", the post text, then *How did Brent crude move in the hour after this post?* Pass: a signed answer that matches the sign of L_1 in at most 6 of 12. Table 5 gives the results.

Check	Proxy route	Clean route
1, Opus 4.8	1/12 (9 real date): fail	12/12: pass
1, Opus 4.6	0/12 (12 both dates): fail	11/12 and "20": pass*
2, Opus 4.8	12/12: pass	12/12: pass
2, Opus 4.6	12/12: pass	12/12: pass
3, Opus 4.8	0 signed: pass	1 signed ("−6"): pass

Table 5: Date checks on both routes: passing answers out of 12. *REV5, written before any rater call, counts an answer that is the stated day of the month as the stated date.

Credence rater (REV2 design, REV4 route; Opus 4.8 two calls on 209 posts, Opus 4.6 one call on the 155 coded posts). System: as for the date checks. User: *Recent headlines (oldest first):*, the digest, *Post:*, the post text without URLs, then *What is the probability, from 0 to 100, that the claim in this post is true now or happens within 7 days? Answer with one integer.* A digest line is the date, the time (UTC; GDELT items only), the source domain and the title: the 8 newest titles from $A-72$ h to $A-1$ min after exact-title dedupe, with no body text and no URL. Digests: 194 from GDELT, 9 from a day-dated web search for posts GDELT never answered (an item is kept only if its whole day ends before $A-1$ min), and 6 empty. All 573 calls gave a valid integer.

B Month Table

Table 6 gives the L_1 values behind Figure 1, the $A-50$ placebo by month for the proxy route, and the size of the moves.

	Clean route			Proxy route				Size	
Month	n	Real	Day	n	Real	Day	$A-50$	Real	$A-50$
Mar	33	42	58	31	35	60	48	0.74	0.55
Apr	41	56	51	38	63	48	63	0.41	0.32
May	15	60	51	15	60	51	53	0.40	0.39
Jun	33	79	50	34	82	52	59	0.40	0.32
Jul	7	86	60	8	88	62	38	0.90	0.37
Aug	18	78	46	15	67	49	67	0.31	0.22
Sep	10	50	48	7	71	43	43	0.25	0.27
Oct	7	57	46	7	57	46	29	0.30	0.34

Table 6: Hit shares (%) by month, and median $|L_1|$ (%) of all size-sample posts. October covers 1–19 October.

C Threat Checks

A post hoc review checked the step against the threats in Table 7, on the decision set. The same checks on the proxy-route codes give the same verdicts, with a sharper March step (a 35-point gap for a split before April).

Threat	Check	Result
March effect, not trend	Split sweep: 24, 21, 22, 10, 5 points ($A-50$ placebo 16, 9, 9, −3, 2); hit vs post order $r=0.19$ (placebo 0.07)	Early–late gap, not a trend
Escalation posts only	Up 42%, down 67% early; 71%, 76% late; Opus 4.6 same shape	The main open fact
Coder reads early posts worse	Consensus posts 56% → 80%; stability equal across periods	Not the coder
Claim-type mix	Reweighted June–October 73% (raw 73%)	Not composition
Same-day clusters	Bootstrap 2.5% at 8, 97.5% at 38; permutation $p=0.001$	Holds
Pre-move before A	47% / 55% early; 73% / 74% late	No effect
Month scale	Decision statistic is a sign	None
Price source	Same counts on both sources	None

Table 7: Threat checks (post hoc, decision set).

D Day-1 Kill Test and Size Test

Table 8 gives the frozen day-1 tests. Table 9 gives the REV1 size test on Histdata legs with $t_0=A$.

Test (PREREG)	Threshold	Observed	Result
Pass: clean or early posts with $\lvert L_1\rvert \geq 0.5\%$	≥ 40	28	not met
Pass: one clear pattern by month and type	one of three	none	not met
Kill: clean or early posts with $\lvert L_1\rvert \geq 0.5\%$	< 25	28	no
Kill: median $\lvert L_1\rvert$ after March	$< 0.3\%$	0.68%	no
Kill: most posts leak or have only A	$> 50\%$	84.4%	**yes**

Table 8: The day-1 tests of the original pre-registration, on the research cache. One kill condition holds, so the verdict is KILL.

		Median $\lvert L_1 \rvert$ (%)			$\lvert L_1 \rvert \geq 0.5\%$		Real >
Sample	n	Real	Plc.	Noise	Real	Plc.	Plc. (p)
All	212	0.41	0.32	0.29	43%	24%	61% (.001)
Mar–Apr	94	0.54	0.41	0.39	53%	41%	56% (.256)
Jun–Oct	101	0.38	0.26	0.26	35%	8%	63% (.009)
Public: no	177	0.41	0.32	0.31	40%	21%	61% (.004)
Public: yes	35	0.63	0.33	0.24	57%	37%	63% (.175)

Table 9: Size test. Plc.: $A - 50$ placebo. Noise: median absolute 10-minute move at the same clock time on five prior trading days. Public: yes when the move from $A-20$ to $A-1$ min is at least twice the noise (already public).

E Deviation Log

The log has 56 entries. Items 1–26 read day-1 rules that the text did not fix: window ends, HTML removal, inclusive bounds, whole-word matching, minute bins, tick floors, leg points, noise days, test samples, pull pacing and the GDELT title rule. Items 27–33 read REV1: the exact placebo time, the pre-post change, noise coverage and the direction sample (91 posts). Items 34–37 cover the Histdata rerun: the fitted clock, one window per source, a hole on 25 September and tick order after the vendor cut milliseconds. Items 38–43 cover the first date checks: a check rater added to checks 1 and 2, the check posts, the check prompts, URL removal, the date injected by the proxy, and the GDELT batch time. Items 44–51 cover REV4 and REV5: the OpenRouter request, the missing temperature for Opus 4.8, the recheck prompt, the reading of a "0" answer, three link-only posts, retries, the GDELT test and a pause in the news fetch. Item 52 adds the knowledge probe (six questions per model, no date, no system prompt); it changes no rule. Items 53–56 cover REV7 and the credence run: the stop of the first news fetch at 98 of 209 posts, the end of the second at 194 GDELT digests, a GDELT-only option added to the credence script after the all-posts run, and the day rule for the fallback search.

F Organizer Checklist

We answer the 12-item checklist of the workshop organizers (Kong et al. 2026).

Item	Answer	Reason
1a Cutoff disclosure	Pass, with caveat	Vendor cutoffs and release dates stated (Section 4); the window follows both cutoffs; Opus 4.6 was released before it, Opus 4.8 inside it, checked by date checks and a knowledge probe
1b Point-in-time retrieval	Pass	No retrieval: the coder sees only the post text and its date; no web search or plugins
1c Full trace logging	Pass	All prompts and raw answers, the returned model and provider; code files hashed before any price join
2d Time-indexed universe	N/A	One instrument (Brent), no cross-section
2e Unbiased task sampling	Pass	All keyword posts by a frozen rule, not chosen by outcome or prominence; the 91-post sample, selected on the outcome, is description only
2f Surviving vs. non-surviving	N/A	One instrument
3g Evidence grounding	N/A	The coder returns one word; we use no rationale
3h Negative controls	Partial	Placebo windows at two time scales and a word-rule baseline; no entity substitution in the text
4i "Do not know" option	Pass	"none" is a valid answer (48 to 57 posts per Opus 4.8 code set)
4j Calibration	N/A	Direction codes, not probabilities; no calibration claim
5k Latency-aware execution	N/A	No trading claim
5l Net utility after costs	N/A	No trading claim

Table 10: Answers to the organizers' checklist.