Sentiment Half-Life Is Not Separately Identified from Return Autocorrelation in Event-Driven Equity Panels

Subtitle: When Half-Life Isn't Half-Life — Diagnostics for Event-Study Decay Estimators

A companion methodological paper to `paper_v1_ai_compression.md` *(the original AI-compression theory paper). The framing of this paper is the empirical pivot the pilot data forced.*

Abstract

Increasingly, research has been using the 'half-life' of how stock prices change after news comes out (specifically, the speed at which the price difference caused by that news disappears) as a key figure when examining how artificial intelligence handles information, focuses its attention, and how much arbitrage is occurring. The standard way of calculating this half-life is in two steps: first, determining the relationship between future returns and a measure of sentiment (a standard least squares approach done for each time period) and then fitting an exponential decay curve to these relationships to calculate tau-hat (where tau-hat is the natural log of two divided by lambda-hat). We contend that, for collections of stocks where news events are the main influence, and given the amount of text in things like regulatory filings, this calculation does not isolate the decline of information driven by sentiment from the return autocorrelation already present, the way volatility changes over time, and the tendency of events to happen in clusters. This calculation relies on three underlying assumptions: first, the sentiment measure must contain a substantial amount of relevant, signed information; second, the connection between sentiment and returns is primarily based on the information in the sentiment measure at the same time as the returns; and third, the timeframe used in the first step is roughly consistent. We have developed three tests, each relating to one of these assumptions and using previously held data, to examine their validity: a 'placebo' test using random sentiment values for the same returns (to test the first assumption), a categorization of the data based on the volatility of the market as measured by the VIX (to test the third), and a 'pseudo-event' test where the timing of the events are randomised (to test the second). All three tests, when applied to a sample of 500 companies in the S&P 500 between 2020 and 2024, disprove these assumptions. The standard calculation yields a median tau-hat of 1.31 trading days with actual 8-K filings, and 1.56 to 1.91 days for each of the three tests – and the results from all of these overlap at every point in their distribution. Therefore, we conclude that under the conditions of the amount of data and the timeframe of our sample, tau-hat cannot be separated from the autocorrelation of the returns. We suggest a four-part breakdown and a method for ensuring a clear result (making the text data more dense, categorising by volatility, and removing the effects of the randomised events) which together would allow the sentiment-related decay to be identified. This work provides a specific, repeatable test that any test of an economic idea based on tau-hat must pass to ensure the result is actually measuring the fading of sentiment.

Keywords: alpha decay, sentiment analysis, event studies, identification, measurement error, AI in finance.

---

1. Introduction

The growing body of research on "post-news drift" (Tetlock, 2007; Tetlock, Saar-Tsechansky, & Macskassy, revisions to the original work in 2008; Loughran & McDonald, 2011) provided empirical finance with a practical way to quantify something: by predicting future returns based on the sentiment expressed in text, at timeframes of h = 1 up to H, then observing how the estimated beta at timeframe h diminishes, and determining the half-life. McLean and Pontiff (2016) employed a similar method to show that, on average, returns from cross-sectional anomalies fall to about 58% of their initial value after publication. More recently, estimates in the style of tau-hat have been used as the primary variable in research examining the implementation of artificial intelligence, the impact of investor attention, and the extent of arbitrage, (for example, Lopez-Lira & Tang, 2023, and the associated research paper that this one builds upon). The underlying belief is that tau-hat provides an estimate of how long it takes for the effect of sentiment-driven information to disappear – that is, the value of tau in the theoretical model.

We contend that this assumption is not without consequences, and the data, at least for one commonly used collection of texts, does not support it.

This paper's main goal is to distinguish between the question of what tau-hat actually measures, and the way previous studies have used it as if it directly measured the theoretical value in Equation 1. We achieve this through four steps.

First, we clearly state the three assumptions on which the standard approach is based: (A1) Signal-density, which indicates that the sentiment series S contains meaningful positive and negative values in relation to the size of the rolling-window sample; (A2) Signal-return covariance, meaning the information in S at the same time is the primary factor in the covariance measured in Stage 1; and (A3) Within-window stationarity, which implies that the rolling Stage 1 window does not include shifts in the overall economy that alter the estimated beta at timeframe h for reasons unrelated to the specifics of each company.

Second, we create three tests, each designed to assess one of these assumptions, and each with a defined null hypothesis and a way to determine if it is rejected. The placebo test substitutes S with random noise on the same returns; if A1 is true, the tau-hat that results should be different from what is found using actual data. The vol-regime stratification divides the stocks into thirds based on the VIX (a measure of volatility) and repeats the primary test for each third; if A3 is true, the original effect should remain apparent in the most stable volatility grouping. The pseudo-event null randomizes the timing of events but preserves the distribution of sentiment; if A2 is true, the resulting tau-hat should differ from the baseline using the actual event timing.

Third, using a sample of 500 firms from the S&P 500 between 2020 and 2024, and scoring sentiment in SEC 8-K filings with the Loughran-McDonald financial tone dictionary, all three of these null hypotheses are rejected. The estimate yields a median tau-hat of roughly 1.3 to 1.9 trading days regardless of whether the sentiment actually contains useful information, whether the events are correctly timed, or which volatility grouping is being examined. The value the standard approach calculates, for this collection of texts, is not the sentiment decay parameter from Equation 1.

Fourth, we offer a breakdown into four parts (Equation 1 prime) which provides a theoretical explanation for the incorrectly identified tau-hat: it is a combination of sentiment decay, return autocorrelation,

volatility regime consistency, and event timing patterns, with the relative importance of each component determined by the characteristics of the text and the time frame used. This breakdown transforms the negative finding into a constructive one, because it clarifies what the standard estimate does identify, and what would be necessary to isolate the sentiment component.

The scope of our conclusion is limited. We are not saying that tests based on tau-hat are always incorrect. Instead, we are saying that before interpreting any change in tau-hat over time or across different groups as evidence of an economic effect, it is necessary to first confirm that tau-hat actually measures sentiment decay. Our diagnostics make this confirmation both possible and relatively inexpensive. In collections of texts that are dense with information, accurately scored, and relatively free from shifts in the economy, these assumptions may be valid, and our diagnostics should pass; in our case, they do not.

The paper is organized as follows: Section 2 describes the standard approach and outlines the key assumptions. Section 3 details the three diagnostics. Section 4 presents the empirical results. Section 5 develops the four-part breakdown. Section 6 discusses the implications of our findings and how they relate to the broader research area of identifying how long information lasts. Section 7 proposes a revised method for accurate identification. Section 8 considers potential issues with the identification of the diagnostics themselves. Section 9 concludes.

---

2. The Canonical Pipeline and Its Implicit Identifying Assumptions

Recall the structural model that motivates the canonical estimator:

$$ \mathbb{E}!\left[r_{i,t+h} \mid S_{i,t}, \mathbf{X}*{i,t}\right] = *beta* $$\, S_{i,t}\, e^{-h\lambda_{i,t}} + \mathbf{X}_{i,t}'\boldsymbol{\gamma},\qquad \tau = \tfrac{\ln 2}{\lambda}. \tag{1

Let's say r(i,t+h) is the percentage change in a company's stock price (expressed as a 'log return') from day t to day t+h, and S(i,t) is a score for the sentiment from something in the news for company i at time t. The structural model in Equation 1 says that the OLS regression coefficient of r(i,t+h) on S(i,t) - after accounting for other factors - should follow an exponential decay pattern: it starts at a certain amount and then fades away, and how quickly it fades is the 'decay factor'.

The usual way of doing things is in two steps.

First, for each timeframe 'h' (from 1 to H) you do an OLS regression of r(i,t+h) on S(i,t) (and the other factors). You're doing this for each company 'i' and looking at a 'rolling window' of news events. From this, you get an estimate for beta (beta-hat) at timeframe 'h', and a measure of how accurate that estimate is (the standard error, sigma-hat) at timeframe 'h'.

Second, you find the best fit for beta-hat at timeframe 'h' with the formula beta-zero times e^(-h lambda). This is done by weighted nonlinear least squares, giving more importance to estimates with smaller standard errors (weight of one over sigma-hat squared). You then work out tau-hat, which is the natural log of two divided by lambda-hat.

For tau-hat to accurately show the true value of tau, three things need to be true.

(A1) There needs to be meaningful ups and downs in the sentiment scores. Basically, the S(i,t) values need to change, and to be positive and negative. If the chance of S being anything other than zero gets very small compared to the window size, the changes in S become almost nothing, and the OLS regression in Stage 1 becomes something that is almost constant. In this case, beta-hat at timeframe h will be almost zero for all timeframes, and the nonlinear least squares in Stage 2 will only fit the random 'noise' and the estimated decay rate will simply match the random curvature of that noise.

(A2) The relationship between the stock returns and the sentiment needs to be because of the model in Equation 1, and not just because they happen to occur together. If the timing of the news events is tied to the stock price continuing to drift upwards (or downwards) after the event, and this is not related to the sentiment score itself, then beta-hat at timeframe h will decrease as h increases for a reason that has nothing to do with information disappearing.

(A3) Within that 'rolling window' in Stage 1, the way the stock returns behave shouldn't be wildly changing. If beta-hat at timeframe 'h' is different in different parts of the window due to shifts in volatility, interest rates, or how things are related to each other, then the beta-hat at timeframe 'h' you get by combining all the windows will be an average of different decay rates, and the fit in Stage 2 will reveal a mix of those rates.

Each of these assumptions can be checked with a test. Assumption A1 is wrong if replacing the sentiment score with a random one doesn't alter tau-hat. Assumption A2 is wrong if randomly changing when the news events happen doesn't change tau-hat. And Assumption A3 is wrong if you split the data into different situations and the differences between them disappear.

The population OLS coefficient implied by Equation 1 satisfies

$$ \beta(h) = \frac{\mathrm{Cov}(r_{i,t+h}, S_{i,t})}{\mathrm{Var}(S_{i,t})} = \beta_{i,t}\, e^{-h\lambda_{i,t}}. $$

---

3. Three Diagnostics

Diagnostic 1: Placebo Test (A1)

What it's testing: According to A1, tau-hat is tied to the sentiment score. Therefore, swapping in a score that has no actual information should lead to a very different distribution of tau-hat compared to what you get with the real data.

How to do it: On that same results screen, first, for each company and month, we'll pick N possible dates for the 'event' to have happened, at the same frequency as the original calculation. Second, swap out S for one of two 'fake' approaches. For the 'dense' fake, S is randomly pulled from a normal distribution, each time independently. The 'sparse' fake makes S equal to +1 about 22% of the time (specifically, 22% x 60%), and -1 about 22% of the time (22% x 40%), and zero for everything else. That 22% chance of a non-zero value matches how often the Loughran-McDonald word list "fires" when used on actual 8-K filings. Then, run Stage 1 and Stage 2 again, with the usual settings.

To decide if we reject A1, we look at the middle and the middle 50% of the fake tau-hat values, and see if they overlap with the middle and middle 50% of the tau-hat values from the real data.

This is a really good test because the returns data is fixed. The way returns behave – the way they follow each other, how much their swings vary, how extreme the ups and downs are, how they relate across companies, and the tendency to drift after the event (and this last bit is separate from S) – all of that is kept the same by design. The only thing changing between the real and fake versions is whether the sentiment actually tells us anything. If tau-hat turns out to be the same in both, the calculation isn't using the sentiment at all.

Diagnostic 2: Volatility-Regime Stratification (A3)

What A3 assumes: If tau-hat noticeably shrinks or grows around a specific date (like a new government rule, a new model for predictions, or a change in the overall economic situation) it is showing a real change in how returns behave, and should show up even when we look at things within each situation. But, if the effect is really because of a change in overall market volatility at the same time, and that's not what the test is meant to find, then looking at volatility alongside the test should lessen or remove the effect.

How to do it: Add the VIX level (or a similar measure of how much volatility is happening) at the end of each month in the tau-hat results. Split the tau-hat results into thirds (terciles) based on VIX level. Within each of those thirds, do the same before/after break Welch t-test on the log of tau-hat.

To reject A3, the effect after the break must not be statistically significant in the VIX tercile that has the most even number of results before and after the break, and therefore the least chance of being thrown off by the time of year. And yet, the effect when all the data is looked at together must be statistically significant.

Diagnostic 3: Pseudo-Event Null (A2)

What A2 assumes: The after-event drift that Stage 1 finds is because of new information at the time of the actual event. If we pick random dates for the 'event' to happen, but keep the same overall sentiment distribution, this should break the natural relationship between the sentiment and the event, and the tau-hat results should be different than those from the real events.

How to do it: For each company, randomly select the same number of dates from their trading calendar as there were real events. Then, randomly choose S values from the Loughran-McDonald distribution on the actual filings. Run Stage 1 and Two as usual.

Reject A2 if the tau-hat results from the randomly timed 'events' are statistically the same as the tau-hat results from the real events.

A successful test of the real events should show something in the after-event drift that’s down to when the event happened - not just the general distribution of sentiment values, or the typical return after any trading day. If just picking random dates gives you the same calculation results, then that something that's tied to the timing of the event is missing.

---

4. Empirical Evidence

4.1 Setting

We tested three ways to check the results, using 500 companies from the S&P 500 from January 2020 to December 2024. Instead of getting returns from CRSP, we used yfinance. We looked at 32,710 announcements from the SEC (filed as 8-K forms, and found using EDGAR) and measured the sentiment of the language in them using the Loughran-McDonald financial tone dictionary (ranging from -1 for very negative to +1 for very positive). For ‘Stage 3’ we had 600 periods (firm-months) for 23 companies over 56 months, and had enough information to relate ‘log tau-hat’ to how much attention something gets, how hard it is to trade, and competition. To test the accuracy of the results, we used a smaller group of 100 companies (we didn't need to filter for attention or trading difficulty for this) which gave us roughly 5,000 ‘tau-hat’ values for each test. Details about how we got the information from EDGAR (subject to rate limits), how we worked with attention data from Wikipedia and calculated trading difficulty using Corwin-Schultz are in the ‘empirical README’ file. The tests are done with a script you can use again and, with the data already downloaded, take between three and five minutes to run.

4.2 Real-Data Baseline

The canonical estimator on the real 8-K corpus produces:

Statistic	**Value**
$\widehat\tau$ median	**1.31 trading days**
$\widehat\tau$ 25th percentile	0.32 days
$\widehat\tau$ 75th percentile	4.07 days
n (firm-months with finite $\widehat\tau$)	600
Reference: literature post-news drift	$\sim$5–30 trading days

Looking at actual 8-K announcements, the usual method of calculation gave a ‘tau-hat’ of 1.31 trading days (the middle value). 25% of the time it was 0.32 days or less, and 75% of the time it was 4.07 days or less. Previous research on how quickly news becomes irrelevant suggests this should be between 5 and 30 trading days. Even before looking at attention or trading difficulty, this ‘tau-hat’ value itself gives us information: according to assumptions A1 to A3, it should be roughly the same as ‘tau’, and previous research provides a much wider range for ‘tau’ than what we’re seeing in these 500 companies.

4.3 Placebo Results

Replacing sentiment with a placebo, on the same returns:

Specification	**$\widehat\tau$ median**	**p_{25}**	**p_{75}**	**n**
Real data	1.31d	0.32	4.07	600
Placebo (dense $\mathcal{N}(0,1)$)	**1.56d**	0.21	4.39	4{,}932
Placebo (sparse, 22% nonzero, ± 1)	**1.85d**	0.45	5.04	5{,}093

When we swapped the actual sentiment of the announcements for random sentiment values (placebos), the range of ‘tau-hat’ values was almost exactly the same as with the real data at all the values we looked at. The middle ‘tau-hat’ value for the random ‘noise’ was 1.56 days (only 20% off the 1.31 days from the real data); the ‘sparse’ random one (1.85 days) was in between. This means assumption A1 is wrong. The calculation gives pretty much the same ‘tau-hat’ value whether the sentiment is a real signal, complete nonsense, or a random plus or minus one. In short, on this set of announcements, ‘tau-hat’ doesn't seem to be affected by the actual sentiment.

4.4 Vol-Regime Results

Stratifying the 600-firm-month panel by VIX tercile at month-end, with break date 2023-03-15 (GPT-4 release):

VIX tercile	**VIX range**	**n_{pre} / n_{post}**	**$\widehat\tau_{\text{pre}}$ median**	**$\widehat\tau_{\text{post}}$ median**	**$\Delta \log \widehat\tau$**	**$**
All (pooled)	12.4–38.0	338 / 262	1.14d	1.90d	**+0.41**	0
Low	12.4–16.3	23 / 177	0.68d	2.16d	+1.17	0
Mid	16.4–21.7	137 / 69	1.14d	1.14d	**+0.06**	**0**
High	22.8–38.0	178 / 16	1.31d	2.20d	+0.65	0

Looking at how volatile the market is (measured by the VIX), we divided the 600 company-months into three groups based on the VIX at the end of each month. We used March 15th, 2023 (the date GPT-4 was released) as the dividing line. The overall effect (a 0.41 increase in the log of ‘tau-hat’, and a very small chance of that happening by accident - p=0.006) is what the original AI-compression test found, and is the main result. However, looking at the middle VIX group (which had roughly equal numbers of data points before and after March 15th - 137 and 69), there was no effect (a 0.06 increase in the log of ‘tau-hat’, which is very likely to happen by chance - p=0.84). The low-VIX group is 92% after March 15th (because a VIX between 12 and 16 was typical in late 2023 and 2024), and the high-VIX group is 92% before March 15th (because a VIX between 23 and 38 existed during COVID in 2020 and the Federal Reserve’s interest rate increases in 2022). Because of this imbalance in the groups, the results in those groups are more about the differences in the companies between the time before and after March 15th, rather than changes over time within each company. Assumption A3 is rejected in a strong sense: the main effect disappears when looking at the most balanced volatility group.

4.5 Pseudo-Event Results

Holding sentiment values from the Loughran-McDonald empirical distribution but randomizing event timing:

Specification	$\widehat\tau$ median	p_{25}	p_{75}	n
Real data	1.31d	0.32	4.07	600
Pseudo-event	**1.91d**	0.52	4.45	5{,}066

When we kept the sentiment from the Loughran-McDonald data but randomly changed when the announcements happened (pseudo-events), the range of ‘tau-hat’ values was again very similar to the real data. The timing of the announcements doesn't add any information beyond what’s already in the sentiment of the announcements and the company’s returns. This means assumption A2 is incorrect. The usual calculation method, when applied to the real 8-K announcements, is finding a drop in prices after the announcement that isn't related to the announcements themselves being important.

4.6 Joint Synthesis

The three diagnostics jointly implicate the *identification* of the canonical estimator, not its mechanical correctness:

Test	Real $\widehat\tau$	Null-condition $\widehat\tau$	Indistinguishable?	Failed assumption
Placebo (white noise)	1.31d	1.56d / 1.85d	Yes	A1
Vol regime (mid tercile)	$\Delta +0.41$ (pool)	$\Delta +0.06$ (mid)	Effect dissolves	A3
Pseudo-event	1.31d	1.91d	Yes	A2

All three of these tests together suggest the problem is how we’re identifying the calculation, not whether the calculation itself is correct. Essentially, the usual ‘Stage 1 plus Stage 2’ calculation doesn't isolate how quickly the sentiment of an announcement fades away from the normal ups and downs of stock prices and how often the market does the same thing. With this data, what the calculation finds is not ‘tau’.

And this isn't us defending the idea that sentiment fades away (it hasn't actually been proven). Nor are we saying the calculation is fundamentally broken (it can correctly calculate the values for a known situation: when we make up some numbers - alpha = 0.5, phi = 0.8, and gamma = 0.3 - it finds them with a very small chance of being wrong ($p < 0.01$), as shown in the ‘synthetic recovery’ test in the ‘empirical README’). We are saying that the results from using actual data are not compatible with interpreting them as ‘tau’ from Equation 1. In other words, whatever this standard calculation is measuring in this set of announcements, it isn't what the theory says it should be.

4.7 Simulation Evidence: Identification by Signal Regime

When we looked at actual data (as described in Section 4), all three of the core ideas we’re relying on to make sense of things were rejected by the 8-K filings. To be sure this isn't just about something peculiar to that particular set of 8-K filings, we ran a lot of computer-created (Monte Carlo) simulations, carefully controlling the conditions. These simulations used three different ways of generating the data, all of which

had the same basic “rules” for returns and the timing of events…but differed in how much useful information the sentiment in the data held.

All three of these data-generating processes (DGPs) were built in the same way: We made up data for 100 companies over 1,260 trading days. The returns (profits/losses) followed a fairly standard pattern where today’s return is related to yesterday’s (specifically, 4% of yesterday’s return influences today’s, mirroring the daily fluctuations of the S&P 500) and volatility shifted between calmer and more turbulent periods. Events (the filings themselves) happened at random times, but tended to come in clumps, like we see with real regulatory filings. The sentiment of the filings was flagged as ‘present’ 22% of the time - about as often as the Loughran-McDonald dictionary finds sentiment in actual 8-Ks. The only difference between the three DGPs is how this sentiment affects returns.

DGP A is what we’d expect if the sentiment of a filing had absolutely no effect on returns, at any point in time. Because of this, the ‘tau-sentiment’ (a value representing how long the effect of sentiment lasts) doesn't exist within DGP A; there's nothing to measure the decay from.

DGP B imagines a strong signal. The sentiment of a filing causes returns to drop back to normal in an expected way (an exponential decay) over 10 trading days - a number in line with typical ‘post-news drift’ research. A filing with a particular sentiment is expected to shift returns by about 0.005 (or 100 basis points at its peak) - the kind of impact you’d expect from a quick, informative source like a press wire or an LLM summary of earnings calls.

DGP C is a weak signal, and designed to mimic the actual strength of sentiment in 8-K filings as measured by the Loughran-McDonald dictionary. It’s structured exactly like DGP B (also decaying over 10 days), but the impact of sentiment on returns is much smaller, only 0.0005 (or 10 basis points at its peak). This is similar to what the Loughran-McDonald scores actually show.

We then used the standard two-step estimation method on each of these simulated datasets, and in the second step, tried three different mathematical formulas to describe the decay: the usual exponential one, a ‘power-law’ alternative, and a ‘stretched-exponential’ one. Since the only thing varying between the datasets is the strength of the sentiment signal, any differences in the ‘tau-hat’ (our estimate of tau) results must be due to how well the estimation method works in each situation.

The exponential formula gave us a clear pattern. With DGP A (no signal), the estimation method gave us a median ‘tau-hat’ of around 2.0 days. With DGP B (strong signal), it correctly found a median ‘tau-hat’ of 9.9 days - almost exactly the 10 days we’d programmed in. With DGP C (weak signal), the method gave us a median ‘tau-hat’ of 4.6 days, a number that’s seriously pulled down toward the ‘noise floor’ (basically, random variation) and far away from the true value of 10 days. The strong signal let us find the true underlying value; the weak signal did not.

The power-law and stretched-exponential formulas gave the same overall idea. The strong-signal dataset consistently showed a much larger ‘tau-hat’ than the weak-signal dataset, and both were above the “no signal” level. But with the weak signal, the ‘tau-hat’ was still far below the actual programmed value.

This leads to two important conclusions. First, the standard estimation method is only “identifiable” under certain conditions. There are some kinds of data (specifically, certain ‘corpus parameters’) where it consistently finds the true underlying value, and others where it doesn't. The 500-firm 8-K pilot study in

Section 4 falls into the “doesn't” category, which explains why our tests failed. Second, the fact that it doesn't work isn't because of the specific exponential formula we chose. Both the power-law and stretched-exponential formulas gave a ‘tau-hat’ that was pulled down towards the noise floor with a weak signal, even though the actual scale of the numbers was different. No matter which decay formula we assume, we get the same result.

This simulation takes our findings from the real-world data and makes them much broader. The tests in Section 4 don't just show that the usual method fails with a specific dataset of 8-K filings. They show it fails in situations with the characteristics of those filings and the length of time we looked at the data, and that datasets with a strong signal over a consistent period are expected to give a well-defined ‘tau-hat’ that matches the true underlying value.

The quantitative results across all three functional forms are summarized below.

Specification	**Stage 2 form**	**Median tau-hat**	**n**
DGP A — no signal	exponential	2.00d	4{,}683
DGP B — strong signal (planted tau = 10d)	exponential	**9.85d**	5{,}697
DGP C — weak signal, LM-calibrated (planted tau = 10d)	exponential	4.57d	4{,}255
DGP A	power-law	0.45d	4{,}415
DGP B	power-law	3.19d	5{,}628
DGP C	power-law	1.06d	4{,}255
DGP A	stretched-exponential	2.15d	3{,}953
DGP B	stretched-exponential	**10.94d**	5{,}655
DGP C	stretched-exponential	4.12d	4{,}255

4.8 Identification Sweep and M1 Validation

It’s reasonable to ask if our three-DGP simulation only gives us information from three specific points in the settings we’re using, rather than a broader view. So, we ran a test, changing only the strength of the signal (how noticeable the sentiment is) and keeping everything else about the DGP the same. Throughout this, the ‘true’ structural parameter tau_true was set to 10 trading days and sentiment density was at 22 percent - mirroring how often words ‘fire’ in real 8-K filings according to Loughran-McDonald. Volatility, how often things happen together, and how clustered events are, were all the same as in our main DGP; only the size of the sentiment in each event (beta) changed.

The estimates we got from this test neatly show how we can pinpoint the parameter. At a signal strength of zero, the usual estimation method gives about 1.83 days. As the signal gets a little stronger – 0.0001, 0.0003, 0.0005, 0.001 – the estimate of tau-hat sticks to the ‘noise floor’ (between 1.7 and 2.0 days). The

method simply can't find the signal we've put in at these strengths. When the signal reaches 0.002, tau-hat starts to rise to 2.39 days, but is still a long way from the true value of 10. At 0.005 it reaches 4.72 days, roughly halfway to where it should be. It's only at 0.010, ten times stronger than the 'weak signal' we started with, that the estimate gets to 8.31 days, and gets close to the original 10.

This test makes our idea of being able to find the parameter much more solid. There's a specific range of signal strengths, somewhere between right between 0.002 and 0.005, where the standard estimation method goes from completely missing the parameter to almost finding it. Below that range, it will always give you the noise floor, no matter how strong or weak the signal you've planted actually is. And the 8-K filings combined with Loughran-McDonald's word list fall into that 'missing' range, which is why the tests in Section 4 show problems with real data.

This test also stops us from worrying that the problems we found with the data are just because of one specific set of values. They aren't. Instead, we have a whole range of situations (defined by how the text looks) where the estimation method fails to find the thing it's supposed to find.

Then we thoroughly checked the revised way of finding the parameter we suggested in Section 7. We kept the main DGP at the 'weak signal' setting similar to the real 8-K filings (sentiment density 0.22, per-event signal strength 0.0005). As expected, tau-hat was 1.93 days, way below the true 10, confirming that the original method fails to find the parameter. We then tried three different improvements to the revised method on the same DGP.

First, we only increased the sentiment density, raising it from 0.22 to 0.80 (what you'd get from a language model based sentiment scorer) and left the per-event signal strength at 0.0005. Tau-hat became 1.65 days, a little below where it was before. Just increasing density isn't enough to solve the problem. The spread of sentiment goes up, but the actual effect of sentiment on returns is from both how much sentiment there is in an event, and the value of that sentiment. A lot of density with a fixed 'size' of sentiment doesn't add much to the overall effect, and is overwhelmed by the random noise in how often things happen.

Second, we only increased the signal strength, from 0.0005 to 0.005, and left the density at 0.22. Tau-hat became 4.45 days, getting part of the way to the true value. Increasing the size of the sentiment helps, but not by itself.

Third, we increased both. Density went to 0.80 and signal strength to 0.005. This mimics using a language model to score a lot more events, and giving a more detailed (positive or negative) score to each one. Tau-hat was 9.32 days - 93% of the original 10. We had found the parameter!

This confirms in a very specific way that the M1 modification to the revised method works. M1 isn't just about increasing density. It's about increasing both density and the quality of the sentiment scores, and language model scoring gives you both together. Increasing density without quality leaves you at the noise floor. Improving quality without density gets you halfway there. Only the combination restores the parameter identification in our simulated world, and it's the only one of the three options that gets an estimate of tau-hat that's statistically close to the true value.

So, for anyone reading this who is applying it to their own data, two things are important. First, the diagnostic test in Section 3 is a conservative measure of how much problems with the text itself can mess

things up. The revised method in Section 7 isn't a wild guess, but is based on the way things are structured and has been checked with data. Second, the revised method can be tested with real data in the future. If you score a news wire using a language model at a similar cost to Loughran-McDonald, the same tests that fail with 8-K filings should pass, and the estimated tau-hat should represent the parameter that the theory says it should.

Quantitative results of the identification sweep (planted $\tau_\text{true} = 10$ days, $\rho_S = 0.22$):

Signal strength β	**Recovered $\widehat\tau$ (median)**	**Status**
0.0000	1.83d	Noise floor (no signal planted)
0.0001	1.72d	Pinned at floor
0.0003	1.81d	Pinned at floor
0.0005	1.77d	Pinned at floor (DGP-C-equivalent strength)
0.0010	2.01d	Pinned at floor
0.0020	2.39d	Slight lift from floor
0.0050	4.72d	Partial recovery (~47% of true 10d)
0.0100	8.31d	Approximate identification (~83% of true 10d)

Quantitative results of the M1 validation:

Variant	**ρ_S**	**β**	**Recovered $\widehat\tau$**
Baseline (LM-on-8K equivalent)	0.22	0.0005	1.93d
M1a — density only	0.80	0.0005	1.65d (still floor)
M1b — quality only	0.22	0.0050	4.45d (partial)
M1 full — density + quality	**0.80**	**0.0050**	**9.32d** (93% of planted 10d)

5. The Decomposition

Because of what the tests are telling us, we can break down what we're trying to measure. Let's say the observed half-life from our standard method is "tau-hat-observed", and the half-life the theory itself describes is "tau-sentiment". We're suggesting that "tau-hat-observed" is made up of four parts, each with a certain weight: w1 multiplied by "tau-sentiment", w2 multiplied by "tau-autocorrelation", w3 multiplied by "tau-volatility-clustering", w4 multiplied by "tau-event-clustering", plus a little bit left over. We're calling this relationship Equation 1 prime.

$$ \widehat\tau_{\text{obs}} = w_1\, \tau_{\text{sentiment}} + w_2\, \tau_{\text{autocorr}} + w_3\, \tau_{\text{volclust}} + w_4\, \tau_{\text{eventclust}} + \varepsilon. \tag{1$'$} $$

Each of these four parts actually relates to something we can test in the real world. First, "tau-sentiment" is the core of Equation 1 - how quickly the expected return after a genuine, measured change in how

people feel (sentiment) goes away. Second, “tau-autocorrelation” is the apparent decay caused by the way returns are linked to each other over a short period in the price itself, when looking at the dates of the events; the placebo test shows this. Third, “tau-volatility-clustering” is the apparent decay from shifts in how volatile things are, which alter the level and how long the autocorrelation lasts within the rolling Stage 1 window; the volatility-regime stratification test picks this up. And fourth, “tau-event-clustering” is the apparent decay from things happening after an event that aren't to do with the sentiment itself—so, how the expected return depends just on when the event happened; the pseudo-event test shows this.

The weights (w1 to w4) are decided by three things about the text we're analyzing and the broader economic period: the density of sentiment (rho-S, or how likely sentiment is to be something other than zero), how stable the economic period is, and how related events are to the information they contain. As rho-S gets closer to zero (meaning sentiment is rarely non-zero), the variation in sentiment shrinks, the portion of return variation explained by sentiment goes down, and w1 gets close to zero. When the rolling Stage 1 window covers changes in the economic regime, the idea that beta-hat is the same at all points within the window isn't true and w3 increases. Finally, if events happen often, but the details of when they happen aren't strongly connected to the information they provide (like 8-K filings clustered around the end of a quarter, or earnings announcements that are on a set schedule), w4 increases.

Equation 1 prime gives a proper structural explanation to our initially miscalculated estimate. The thing we’re trying to find isn't without definition; it’s a known blend, and the weights of that blend are determined by characteristics of the text and the economy that the researcher can know before doing the work. The problem of finding the true value shifts to a problem of measuring: how do we get “tau-sentiment” from “tau-hat-observed”, which we know is a weighted combination of “tau-sentiment” and three things that mess it up?

The four-component decomposition follows from a single population identity. The Stage 1 OLS coefficient at horizon h is, by construction, the covariance between the forward return and the contemporaneous sentiment value divided by the variance of sentiment. When the conditional return process contains both the structural sentiment-decay term of Equation 1 and additional terms that vary with horizon for reasons unrelated to the sentiment shock — autocorrelation in the underlying return process, regime-driven shifts in volatility, and post-event drift contingent on event timing rather than event content — the covariance numerator decomposes additively into four parts, one per source. Stage 2 then fits a single exponential to the resulting weighted sum, and the recovered tau-hat is approximately a variance-weighted average of the four characteristic decay times. The weight on the structural component is proportional to the structural variance, which itself scales with the product of sentiment density and per-event signal magnitude — exactly the two parameters the M1 validation in Section 4.8 sweeps over. When their product is small, the structural weight is small and the recovered tau-hat is dominated by the three nuisance components; when their product is large, the structural weight dominates and the recovered tau-hat approaches the true tau. The full algebraic derivation is straightforward but mechanical, and we defer it to a methodological appendix.

6. Implications and Research Frontier

6.1 Implications for the AI / Attention / Compression Literature

Lots of recent and current studies use these 'tau-hat' calculations to test ideas about the economy. We aren't saying those studies are wrong, but our findings relate to a particular collection of text from a specific period of time. What we are saying is, before you interpret changes in tau-hat over time or across different things, to say something about what's happening in the economy, you have to be confident tau-hat is actually measuring how sentiment fades. Without tests like those in Section 3, the direction and size of any impact on tau-hat are muddied by the thing you're trying to measure and by the way your explanatory variables are changing along with the 'contamination' in Equation 1 prime.

This is a serious problem for tests that look at definite points in time, like when a large language model was released, regulations changed, or the disruptions of the COVID era. Macroeconomic volatility shifts at around the same times, potentially introducing A3 'contamination'. It's also a problem when comparing companies that get a lot of attention with those that don't. Companies with a lot of attention have reliably different patterns in how their returns relate to each other, how their volatility clusters, and how often 'events' happen, when compared to companies with little attention. The differences in tau-hat across companies with varying attention might reflect these existing differences, or they might reflect the thing the researchers are actually interested in. And finally, it's a problem for situations where you have very little text to work with, like 8-Ks alone, sentiment from social media calculated by a dictionary, or infrequently issued analyst reports. In those cases, rho-S limits how much w1 can be, and sparse collections of text have a low rho-S to begin with.

As a fairly easy check, any paper using tau-hat to reach its conclusions could do a version of each test we detail in Section 3. They wouldn't need any extra data beyond what they've already used to calculate the original tau-hat, and it would only take a few minutes of computer time.

6.2 Research Frontier: Identification of Time-Dependent Informational Decay

More generally, this paper explores the edges of research that is visible in related work, but hasn't been specifically pointed out. For twenty years now, financial researchers have been building increasingly complex models for predicting asset returns. But figuring out what structural things are driving those predictions has been slower. How information fades over time - tau in Equation 1 - is a nice illustration of this. It's likely to be the 'real' thing an AI-and-markets theory would focus on, and yet, in realistic scenarios, the usual way to calculate it doesn't isolate it from the standard patterns in how returns behave.

In this view, the difficulty in modern financial research isn't getting accurate predictions, it’s identification - figuring out which of all the things that are statistically related to returns are the economically important ones a theory should focus on. The tests in this paper are a small step towards bridging that gap. They don't give you a new way to calculate anything. They simply show when a calculation you're already using is actually measuring the thing the theory says it is, and when it's measuring something else.

This way of thinking applies to more than just how long sentiment lasts. Estimating decay rates is used in tests of 'arbitrage crowding', in tests of 'attention shocks', and increasingly in tests of how large language

models are processing information. Each of these has its own versions of A1, A2 and A3. Stating the assumptions and providing tests for them before you do the analysis is a useful, and general, way to approach things.

---

7. A Corrected Identification Strategy (Sketch)

Given the decomposition in Equation 1 prime, three changes to our standard method will specifically address those three sources of error.

The first change, M1, is to increase rho-S by getting more sentiment information and scoring it more accurately. We could add press releases from Yahoo Finance, Reuters, and 6-K filings from the SEC (EDGAR) to the 8-K filings, and also use transcripts of earnings calls. We'd replace the Loughran-McDonald dictionary with sentiment scores from a Large Language Model (LLM), and that would probably cost about $100-$200 to use the API for a five year look at 500 companies. We'd aim for a rho-S of about 0.80. As rho-S goes up, so does the variation in sentiment, and the portion of variation in Stage 1 OLS explained by the signal. This brings A1 back into alignment.

The second change, M2, is to do Stage 1 separately for each volatility regime. Run Stage 1 on data grouped by VIX terciles (thirds), or ideally within a smoothly changing regime score built from the VIX, the difference in long and short term interest rates (term spread), and the difference in credit spreads. Then, combine the resulting "tau-hat" estimates in Stage 3 using regime-specific fixed effects. This gets A3 back on track by making sure things are stable within each window.

The third change, M3, is to remove the "pseudo-event" effect at the firm-month level. For each firm in each month, create a fake panel of events with random timing but the same overall sentiment distribution, and calculate "tau-hat-pseudo". Then, the corrected estimate is "tau-hat-observed" minus "tau-hat-pseudo". If "tau-hat-pseudo" completely captures the contaminating effects, the corrected value is approximately w1 multiplied by "tau-sentiment"—a value we can identify, up to a known scaling factor relating to sentiment density.

We haven't actually done M1 through M3 in this paper. They are the obvious next steps: with a richer, LLM-scored text source, Stage 1 broken down by volatility regime, and the removal of the pseudo-event effect, the AI-compression test described in the original theory paper would finally be usable. We anticipate that the resulting paper will tell us a great deal more about how the economy works than this pilot study, and no matter which way the corrected effect goes, it will be a significant improvement.

---

8. Threats to Identification of the Diagnostic

The tests in Section 3 are built on certain beliefs, and a critical look should ask if each test actually does what it's supposed to.

Regarding placebo independence: the placebo test uses the same stock price history as the actual calculation, and so it keeps the usual patterns of price changes over time, how volatile they are, extreme

swings, and how different stocks move together. However, it won't reproduce any patterns in why the sentiment numbers are appearing; for instance, if actual SEC 8-K filings tend to happen on days with unusual price movement patterns, the placebo won't show that same tendency. How much of a worry this is depends on whether the timing of the sentiment information is independent of the price movements themselves. With SEC filings (most of which are scheduled, like quarterly results or legal deadlines), this independence is pretty much true. It's more debatable with news from the press wires. In our early work, we thought this issue was limited: a pattern in sentiment timing that could completely explain the overlap in the placebo would be a significant discovery in itself.

With "pseudo-event" neutrality, we are randomly picking times from a company's trading schedule (the schedule that produced the actual events), so we're preserving the typical distribution of price changes after trading each day and any continuing trends after the event. But we are breaking the specific link between the sentiment values and when the event happens. A very strict 'null' situation would also randomly choose the sentiment values from a collection of sentiment before the event, rather than the actual sentiment; we didn't do that in this study. Consequently, we're saying something less grand than that the timing of the event itself tells you nothing; rather, the timing of the event doesn't give you any extra information beyond what's in the combined, typical distribution of price changes after trading and the Loughran-McDonald sentiment scores. That more limited claim is enough to disprove A2 in our initial test.

The VIX is calculated from the same price movements whose patterns we are trying to figure out. Because of this, looking at variations within VIX terciles (thirds) controls for only one aspect of those price movements, and leaves things like how skewed the returns are, how likely are big jumps, how intense those jumps are, and the way the price changes over different time periods, all uncontrolled. So, categorizing by volatility regime isn't a perfect experiment, it's a partial adjustment. Better controls would be the difference in interest rates on short and long term government bonds, the difference in rates on corporate and government bonds, realized volatility broken down into various time windows, or the main trends from a big economic model. We're using the VIX as it's the most readily available and simplest measure, and the main evidence against A3 comes from the most balanced tercile, which is fairly stable no matter which control you use as long as it does capture some of the volatility.

The data used for this research is from 8-K filings, the Loughran-McDonald sentiment dictionary, and the S&P 500 between 2020 and 2024. The test for whether this works in general applies to any estimate of 'tau-hat'. But the evidence that it doesn't work is specific to this particular combination of data source, sentiment scoring, which stocks are included, and the length of the macro 'window' used. Sentiment measured from press wires or using a Large Language Model and applied to a different time period might have very different weighting; however, we still expect the tests to be useful, and to pass with high-quality, consistent, and stable data.

Stage 2 functional form. It's fair to wonder if our diagnostic findings are tied to our using a specific mathematical formula (an exponential curve) in Stage 2 and if the real way something fades over time instead follows something like a power law or a stretched exponential. If the actual fading isn't exponential, our standard calculation would be incorrect and the length of time we find (tau-hat) would be useless, no matter how much data we have. We address this concern by re-running our simulations using

all three of those formulas and looking at the pattern of the tau-hat lengths we got from each. The simulation in Section 4.7 reports the full numbers; the qualitative ordering across DGPs A, B, and C is identical under every functional form we tested. In each formula, the strong-signal panel is longer than the weak-signal panel, which in turn is longer than the no-signal specification. The non-identification result is therefore robust to the parametric specification used in Stage 2. A complementary point is that the Stage 1 horizon coefficients themselves are identified independently of any functional form: any choice in Stage 2 simply summarizes the Stage 1 shape into a one-dimensional half-life statistic, and the diagnostics in Section 3 operate at the level of the Stage 1 panel rather than the Stage 2 summary. The exponential is a convenient summary; the diagnostic claim does not depend on it being correct.

Finally, the initial AI-compression test wasn't planned out in advance. These tests were devised after we noticed the unrealistic value of roughly 1.3 days for ‘tau-hat’ in the early work. Each test has a clear idea of what it would mean for it to be wrong and a rule for deciding when to reject the idea, and isn't affected by what the outcome is. But a pre-planned version of the test is better, and we recommend it for anyone else using this system. For this paper, we are stating that the tests were created after seeing the result as part of being open and honest about the process.

---

9. Conclusion

Essentially, this paper points to one key thing we’ve figured out and two suggestions for how to do things.

What we’ve figured out is that the usual method for estimating how long sentiment lasts (a ‘tau-hat’ of how many days until sentiment’s effect on stock price is gone) when applied to the sort of company data you get from SEC filings, doesn't isolate how quickly sentiment itself dies down. Instead, it gets mixed up with how returns are usually connected to each other, the way volatility comes in clumps, and how events (like earnings reports) tend to happen together. We’ve proven this with three tests we designed ahead of time. All three of these tests fail for a group of 500 companies in the S&P trading from 2020-2024, with sentiment measured from 8-K filings using the Loughran-McDonald word list. And, it turns out that the estimated sentiment lifespan (tau-hat) will consistently be about 1.3 days, whether the sentiment is genuine, is just random noise, or happens at random times. In short, the usual method can't tell sentiment lifespan from the normal, inherent pattern of returns.

Our first suggestion for anyone testing an idea about Artificial Intelligence, how people pay attention, how data shrinks, or arbitrage is this: you should always include the three tests from Section 3 as a way of proving your results are strong. They’re easy to run, you can use the same data as your main calculations, it’s clear when they pass or fail, and if they do fail, they will show exactly what’s messing with your conclusions. Some previously calculated ‘tau-hat’ figures will stand up to these tests (especially those using a lot of text scored by large language models, and which only look at periods when volatility is low), but many won't, and the tests will tell you which is which.

And second, the breakdown in Equation 1 and the improved method for determining the lifespan of sentiment in Section 7 can turn this problem into an opportunity. If you have lots of data, separate out the different market conditions and remove the impact of predictable events, the sentiment lifespan can be

found. The original question of whether using AI makes the lifespan of sentiment-based profits shorter is therefore still worth asking. Importantly, with this improved method, the question is now properly and clearly defined.

The core issue, and this is a bigger point, isn't getting predictions that are correct. In modern financial research, the real trouble is identification - working out which of all the things that seem to be linked to returns are actually the thing the theory is talking about. How information's value declines over time is a very clear example of this problem. These tests are a move toward solving it.

Files

- `paper_v1_ai_compression.md` — original AI-compression theory paper (preserved unchanged)
- `paper_diagnostic.md` — this paper
- `formal_model.md` — Eq (1)–(3) plus Eq (1$'$) (the four-component decomposition)
- `briefing.md` — one-page umbrella over both papers
- `explainer.md` — pedagogical walkthrough of the project at honors-precalc level
- `empirical/scripts/diagnostic_tests.py` — runs all three diagnostics on cached data
- `empirical/output/diagnostic_results.json` — full numerical results
- `empirical/output/placebo_dense_fits.parquet`, `placebo_sparse_fits.parquet`, `pseudo_event_fits.parquet` — per-firm-month placebo and pseudo-event $\widehat\tau$ fits

---

Reproducibility

```
cd empirical
./.venv/bin/python3 scripts/diagnostic_tests.py
```

Runtime: 3–5 minutes on free-tier compute. No EDGAR or Wikipedia fetches required (uses cached `live_n500_2020-01-01_2024-12-31.parquet` for returns and pulls VIX from yfinance).

References

(Companion to the references in `paper_v1_ai_compression.md`. Adds, for the diagnostics specifically:)

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428), 1314–1328.

Corwin, S. A., & Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance*, 67(2), 719–760.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, 66(5), 1461–1499.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

Kelly, B., Malamud, S., & Zhou, K. (2024). The virtue of complexity in return prediction. *Journal of Finance*, forthcoming.

Khandani, A. E., & Lo, A. W. (2007). What happened to the quants in August 2007? *Journal of Investment Management*, 5(4), 5–54.

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Working paper*.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, 71(1), 5–32.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3), 1437–1467.