Cross Asset Systematic Highlights

With a Little Help from My Friends: Agentic AI for Market Regimes and Asset Allocation

- We are enthusiastic about the possibilities of agentic AI, even as we are wary to hand off asset allocation decision-making to an agent.
- Given that we know how different asset classes behave in different market regimes, we ask how well an AI agent can do identifying the current macro regime.
- We strongly caution against uncritically accepting what amounts to in-sample, overly confident answers of AI. Agentic AI needs to be grounded in a well thought-out asset allocation process, rather than naively assuming the agent can be the source of the domain knowledge.
- We quantify the degree to which our AI agent is in-sample or unfairly looking ahead, and we offer corrections to mitigate potential cheating in the results.
- Overall, we are pleased with the result, meaning the AI agent can be set up with a process to be empowered to make decisions under uncertainty, producing outperformance vs a reasonable benchmark. Here the specific application was market regime identification, but so long as there is a deterministic process for our decision-making, we can give the agent the advice it needs. And we can broaden our agentic efforts to other asset allocation questions.
- For our legacy Market Regimes process, see the Market Regimes Dashboard updated monthly. For seeing what's out of whack with regard to valuation, positioning, and momentum, please see the Market Dislocations Dashboard, updated weekly.

Cross Asset Systematic Strategy

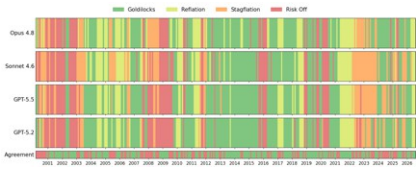
Thomas Salopek AC
(1-212) 834-5476
thomas.salopek@jpmorgan.com
J.P. Morgan Securities LLC

Sandra Djambazovska AC
sandra.djambazovska@jpmorgan.com
J.P. Morgan Securities plc

Dobromir Tzotchev, PhD
(44-20) 7134-5331
dobromir.tzotchev@jpmorgan.com
J.P. Morgan Securities plc

Figure 1: We build an AI agent to identify macro regimes, and the results are fairly consistent across LLMs: Opus, Sonnet, and ChatGPT

AI Agent identified regimes: Goldilocks, Reflation, Stagflation, Risk Off



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Figure 2: Asset allocation based on the AI agents' regime identification boosted the Sharpe by ~0.3 vs the 60/40

Cumulative pnl scaled to same risk target including t-costs. BM is 60/40 or Rule-Based via CPI & PMI



Source: J.P. Morgan Research, Bloomberg Financial L.P.

See page 19 for analyst certification and important disclosures.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Market Backdrop

We have seen descriptions of Agentic AI for asset allocation that have made things sound too easy, paraphrasing "we have an agent for choosing expected returns, portfolio risk methodologies, portfolio constructions, and any other asset allocation task, and the agents will talk among themselves to produce the best answer." While we would trust that the output would *look* reasonable, as it is finely tuned to a full view of the history, we would be much less confident that this would work out-of-sample.

We will jump into the methodology shortly, but it is worth scrutinizing the agentic model's assessment of where we are now in terms of market regime. For this exercise, we do not go into the detail of our usual market regime model, which looks at every input (GDP, inflation, etc.) in terms of level, change, and curvature. Instead, we collapse the world into the familiar market regimes of Goldilocks, Reflation, Stagflation, and Risk-Off. So to begin with, the current period is identified as Reflation (in line with our legacy Regimes model), which may not line up chronologically with earlier periods of Reflation, many of which occurred after bad times, where the asset class performance is reflective of cyclical assets that have been hammered during a recession, which are now bouncing. In fixed income, government bond Sharpe ratios during Reflation are decent, with steep yield curves producing rolldown PNLs with no rate hikes in sight so early in the business cycle.

So is it fair for our agent to categorize the current period as Reflation? We are used to thinking of reflation as periods where policy is trying to boost the economy or counter deflation. So this may come in the form of lowering interest rates or boosting fiscal spending. If we are in a world transitioning from cutting to hiking rates, reflation doesn't obviously fit; however, if central banks are not hiking enough to deal with the last stubborn bit of inflation, this is effectively a form of passive easing. And fiscal spending is already high, with elevated capex spending acting as another support. We considered this question in our Value vs. Growth model and come to the following conclusion: if one looked purely at dispersion of fundamentals (e.g., z-score of standard deviation of FCF yield now vs history), one would come to the conclusion that based on the low dispersion alone, one should tilt to Growth, whereas the supportive monetary and fiscal picture suggests a Value tilt. In any case, the history does not suggest that stocks can crack with such a supportive backdrop, so Defensives are in last place.

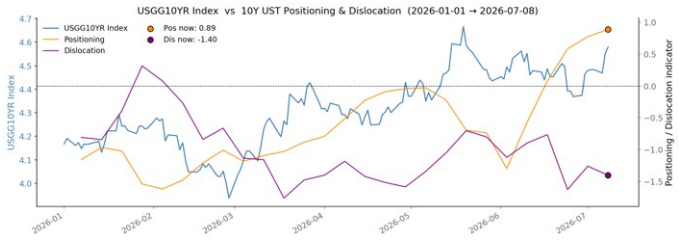
Ultimately, we feel that the worry here is stocks correcting if and when bond yields get to problematic levels. The problematic levels for treasury yields are somewhere between 4.7-5%, which is not that far away. This has been the zone where the market in recent years has been worried about "boil the frog" and overheating. The US faces a very difficult juggling act, trying to boost a demoralized consumer ahead of midterm elections, either by boosting fiscal or lowering interest rates. But doing the former will be inflationary, making it harder to do the latter. And looking across business cycles, by the point where unemployment is well off the highs, so is the deficit, but not so in this case, leaving us with a higher sensitivity to higher yields. So viewing this from the perspective of the Central Bank Trilemma and the Impossible Trinity, we can see higher interest rates and yields, and assuming we do not get enough of an adjustment here, a weakening dollar can be the release valve. This mirrors the difficult choices of the BOJ, although less extreme.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

So how much of a threat to the assumed Reflation regime do higher yields present? So the biggest driver of the 20bp bounce in Treasury yields has been stronger employment numbers, suggesting that the Fed will have to be more restrictive. Before the Iran flare-up in the past few days, fading geopolitical risk should suggest lower bond yields and a flatter term premium, with the assumption that peak hawkishness was behind us. This has reversed somewhat, although let's keep things in perspective; for example, Brent is only $6 off the recent lows, VIX is only 1 point above the recent lows, and Hormuz crossings are at 14 vs ~20 recently. So we would expect the rising bond yields should revert as tensions calm assuming a "mow the lawn" strategy for the US brings tensions down.

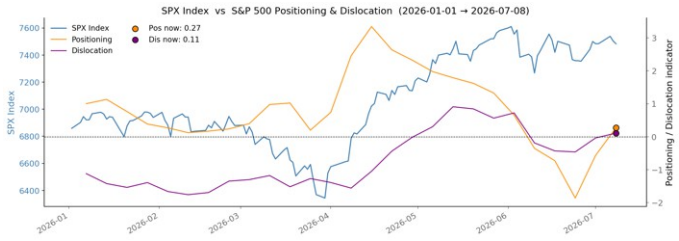
Decomposing the drivers of the bond yield moves through the lens of our Market Dislocations Dashboard, let's zoom in on US Treasury positioning. For Treasuries, we define the hedging pressure as the net long positions of asset manager (AIN) scaled by open interest (not the usual Non-Commercials). This is due to the fact that asset managers are more likely to use treasury futures to express duration views, while leveraged funds are more likely to implement basis trades. Earlier in the year, positioning was ~1 z-score short, and this was neutralized in early May, although the shorts were reinstated in June after yields came down ~20bps off the mid-May peaks. Compared to the more muted yield changes, look at the more dramatic positioning shift from -1z to +1z currently. This positioning shift tells us that even modestly higher yields will scare away the shorts and bring in duration buyers, limiting upside for bond yields. At the same time, stock positioning has gone from 3z long to -2z short, and only recently neutralized. The geopolitical flareups can keep stocks under-owned, although the stock vs bond risk premia tradeoffs are supportive, and the relative positioning behavior confirm to us that asset class moves associated with Reflation (better for stocks, but not necessarily bad for bonds) are achievable. Looking past midterms, rising yields are the big worry given elevated an elevated debt/GDP, and the calming of the recent geopolitical tensions should have a normalizing effect on both the term premium and cross-asset positioning.

Figure 3: Treasury positioning has gone from -1z short to +1z long, indicating there are buyers at higher yields, slowing a further rise



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Figure 4: S&P500 position was heavy at +3z and has since neutralized. Stock vs Bond risk premia tradeoffs are still favorable as is a Reflation regime, and the associated position shifts will be supportive



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Introduction

Using **LLMs as agents inside investment processes** has become an active area of research in quantitative finance. A steady stream of papers and open-source projects now hand an LLM some market or macro context and ask it to reason about the world and take a position, make regime calls, stock selection, news-driven signals, portfolio construction. The appeal is genuine: an LLM can read unstructured information, weigh many signals at once, and articulate a coherent narrative. **The catch: long-history LLM back-tests are rarely truly out-of-sample.** Evaluating agentic models out-of-sample is extremely hard. A back-test is only honest if, at each historical date, the strategy uses only information available on that date. With classical models, this is a matter of discipline: lag your data, avoid restatements, freeze your parameters. With an LLM it is fundamentally harder because the model was trained on text that already describes the period you are testing on. If we ask the LLM model "what regime is it?" as of March 2020 and it does not need to reason from the data, it has read thousands of articles of the COVID crash. The knowledge of what happened next is baked into the model weights. This is look-ahead bias arriving not through your data pipeline but through the model itself, and standard back-testing hygiene does nothing to remove it. The community is well aware of this, and a lot of effort has gone into mitigating it, restricting tests to the period after a model's training cutoff, stripping identifying details from prompts, probing models for memorization, or building synthetic scenarios the model cannot have seen.

One solution would be a **point-in-time model**. For each back-test date, an LLM whose training data is truncated at exactly that date (or once per year), so it knows nothing about the future. Running a 2006–2026 monthly back-test this way would require re-training (or at least re-cutting the knowledge of) a frontier model at every rebalance, hundreds of times. This is not currently possible in our case as LLM models are enormously expensive to train, and their exact training corpora and cutoffs are opaque.

Agentic regime model

This note is our first attempt at building a **simple agentic model** for regime classification. **We acknowledge up front that we cannot guarantee the results are fully out of sample.** We begin with a simple **anonymized prompt** in which we strip out the as-of date. The model receives the macro and market data, growth, inflation, policy, and market/risk state but is never told which month it is looking at. It must therefore determine the regime from the numbers rather than recall it from the date. This does not completely solve the look-ahead problem as a distinctive enough data fingerprint (equities down 30%, volatility spiking, credit spreads blowing out) can still resemble a crisis the model has read about. But it removes the most direct channel through which hindsight leaks in, the date.

Prompt design

In this section we present the full prompt and show each one of the blocks and the design choice behind it.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

J.P.Morgan

ROLE. We cast the model as a senior global-macro strategist and give it a single, tightly-scoped job, **estimate the probability of each regime for the upcoming month**. Crucially, it is explicitly forbidden from offering any asset-class, trade, or portfolio advice, and it is told never to invent data. The model can influence the portfolio only through the regime label it assigns, so it cannot fabricate a return or talk itself into a position.

> *ROLE*
>
> *You are a senior global macro strategist on a multi-asset team. Your ONLY task is to estimate the probability of each of the four macro regimes for the UPCOMING MONTH. DO NOT provide asset-class, investment, trade, or portfolio recommendations of any kind. Communicate like a desk strategist: concise, specific, evidence-led, never invent data.*

REGIME DEFINITIONS. We give the four regimes in the standard growth × inflation framing – Goldilocks, Reflation, Stagflation, Risk-Off – each with its growth/inflation/policy signature and a one-line note on typical winners. This anchors the model to a well-established macro map rather than letting it invent its own taxonomy, and it fixes the label set so that outputs are comparable across dates, models, and reasoning levels.

> *REGIME DEFINITIONS*
>
> *Given a point-in-time context pack for a single as-of date, estimate a PROBABILITY DISTRIBUTION over all four macro regimes (non-negative, sum to 1):*
>
> *- "Goldilocks" — growth above trend, inflation falling toward target, policy supportive/neutral; risk assets benefit, duration neutral.*
>
> *- "Reflation" — growth accelerating AND inflation rising, policy tightening behind the curve; cyclicals/credit benefit, long-duration govt suffers.*
>
> *- "Stagflation" — growth weak/falling while inflation high/sticky, policy constrained; defensive govt and real assets favoured, equities/credit lag.*
>
> *- "Risk_Off" — growth slowing sharply or financial stress dominant (widening credit spreads, equity drawdown, vol spike); long-duration govt (US/JPY) is the haven, HY and EM under pressure.*

OUTPUT FORMAT (STRICT). We require a single JSON object with a fixed schema, a probability for each of the four regimes, a confidence in [0,1], a short list of drivers, and a brief rationale. Demanding strict JSON format makes every monthly reply machine-parseable and turns free-form commentary into structured [illegible] test and audit. The confidence and drivers fields also give us a wind[illegible] sure the model is and why.

> *OUTPUT FORMAT (STRICT)*



Thomas Salopek AC
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Return ONLY a single valid JSON object, with no surrounding text, no markdown, and no code fences. Use this exact schema:

{

"probabilities": {

"Goldilocks": <float>,

"Reflation": <float>,

"Stagflation": <float>,

"Risk_Off": <float>

},

"confidence": <float in [0, 1]>,

"drivers": ["...", "..."],

"rationale": "2-3 short sentences (<= 60 words) grounded in the provided inputs."

}

CONSTRAINTS

- each prob in [0,1]; all four keys present; the four SUM TO 1 (+/-0.01); confidence in [0,1] (your overall confidence in the call); 2-5 non-empty drivers (NEVER more than 5); rationale 20-400 chars. If you cannot comply, still return valid JSON with your best estimate.

- Output the JSON object only — nothing before or after it.

Context pack. Finally, macro snapshot and market-state block are injected in place of {macro_block} and {market_block}. This is the only place data enters the prompt. The macro snapshot groups inputs by economic meaning, Growth (PMI), Labor (unemployment), Inflation (CPI YoY vs the 2% target), Policy rates , and a Rates/FX/ Commodities/Risk panel (curve, FX, commodities, VIX/credit gauges). Each with its level, multi-horizon changes (1m/3m/6m/12m) and a 36-month z-score, plus a few derived flags (yield-curve sign/direction, VIX regime, HY/HG spread states). The market state block lists each asset's past returns (1m/3m/6m/12m), 6-month realised volatility, and 12-month drawdown, followed by cross-asset risk proxies (e.g., stock-bond correlation) that capture the risk-on/off tone.

 # Context pack

Macro snapshot

{macro_block}

Thomas Salopek AC
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Market state

{market_block}

Comparing different models and levels of thinking

We ran the prompt above using four different models and two levels of thinking. Two Anthropic (**Opus 4.8**, **Sonnet 4.6**) and two OpenAI (**GPT-5.5**, **GPT-5.2**). For each model we run two reasoning efforts, **low** and **high**. "Low" is closer to a fast, direct answer; "high" lets the model deliberate at greater length before committing to a distribution. This lets us ask a question that matters both for cost and for interpretation: does more reasoning actually produce better regime calls, or is a clean read of the context pack already enough?

In the plot below we show the regime identified by each model over time, taking the regime with the highest probability as the most likely. The bottom strip shows the agreement among the four models, green if all four forecast the same regime, red otherwise. Overall all four models pick the same regime to be the most likely ~60% of the time (62% in low thinking and 60% at high thinking mode). Notably, extra reasoning does not buy more consensus, suggesting the broad regime story is already legible from a direct read of the context pack. Agreement is also uneven across regimes, the models agree the most in Goldilocks and Risk off, where the risk state is clearest, and disagree most in Reflation and Stagflation. To look beneath the single-most-likely label, the figure below plots the full probability each model assigns to each regime over time, one panel per regime (in low reasoning mode). The four lines are closely aligned across all regimes. Same as with the agreement, the alignement is most tight in Risk-Off and Goldilocks, and most loose in Stagflation. This confirms the models agree on the whole distribution, not merely the top pick.

Despite the differences, the four models tell a similar broad story that lines up with the historical records:

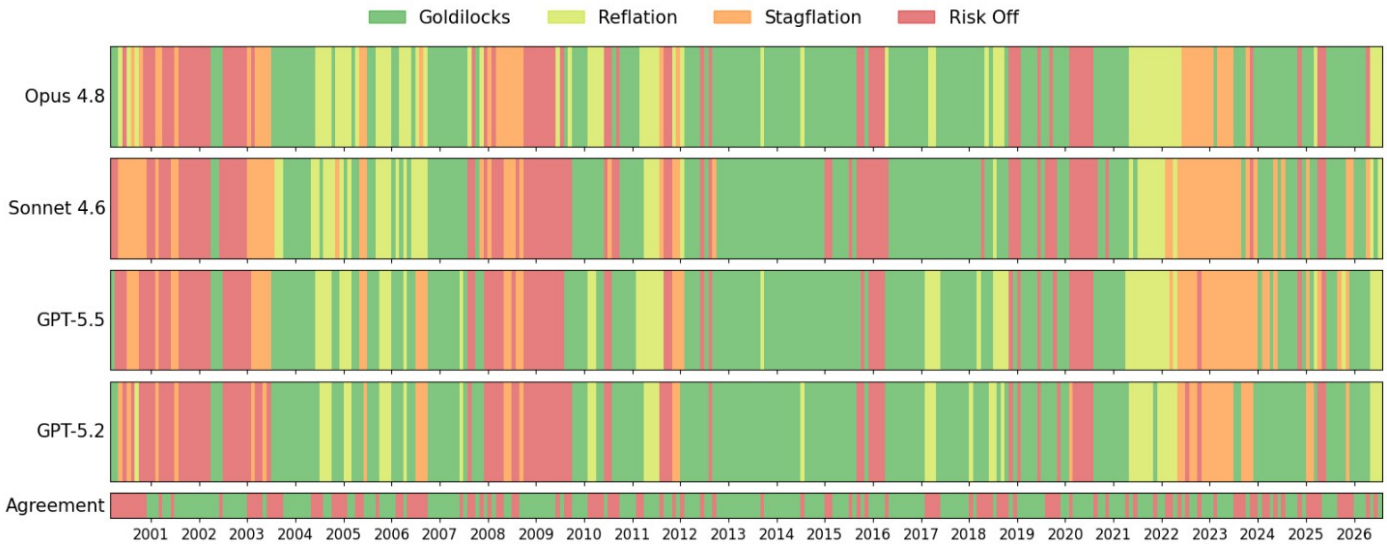
- 2001-2003 is dominated by red (Risk Off) and orange (Staglation), reflecting dot-com
- 2004-2007 is green (Goldilocks) and yellow mostly (Reflation)
- 2008-2009 shows a shared risk-off block, with Opus flagging Stagflation before rotating to Risk off
- 2010s are mostly Goldilocks and see the highest agreement among the models
- 2020 produces a clean Risk off spike followed by Goldilocks, Reflation, and finaly Stagflation
- 2026 starts in Goldilocks and transitions into Reflation

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

J.P.Morgan

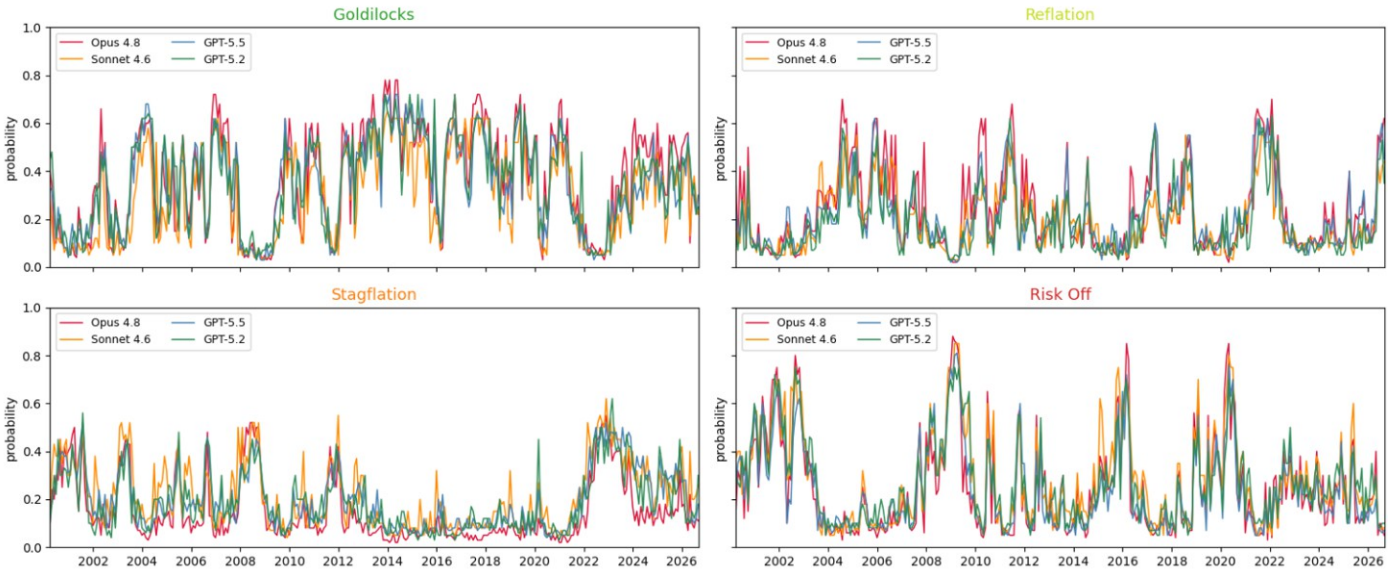
Figure 5: Most likely regime over time, by model (low reasoning)

Low thinking mode.



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Figure 6: Per regime probability across all models in low reasoning mode

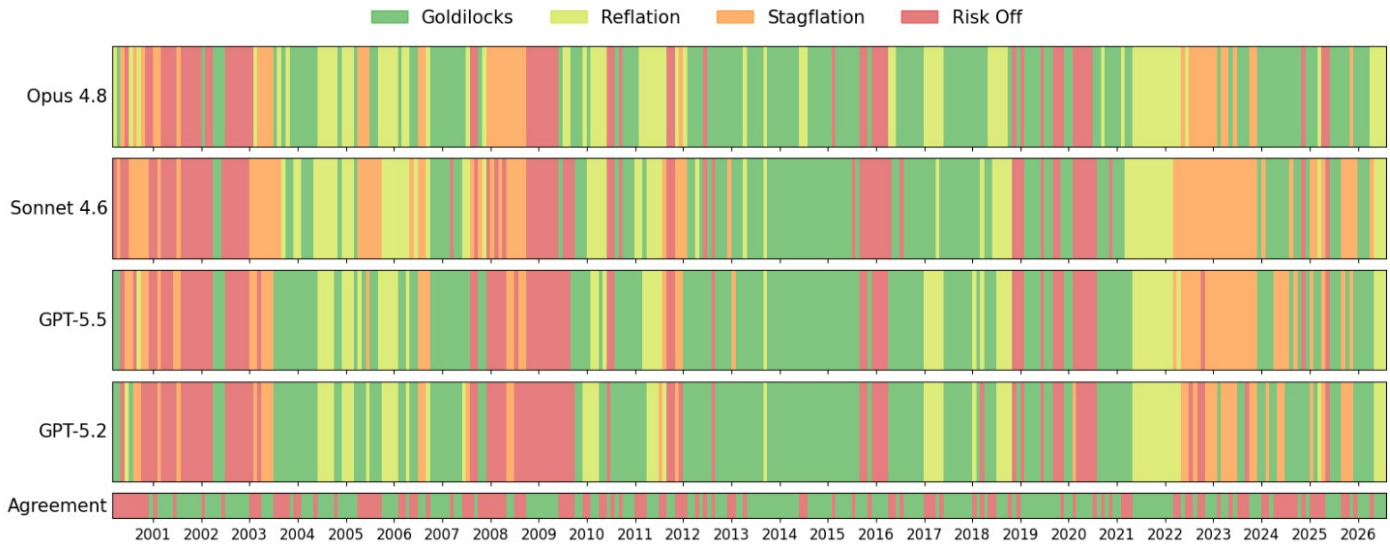


Source: J.P. Morgan Research, Bloomberg Financial L.P.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Figure 7: Most-likely regime over time, by model (high reasoning)

High thinking mode.



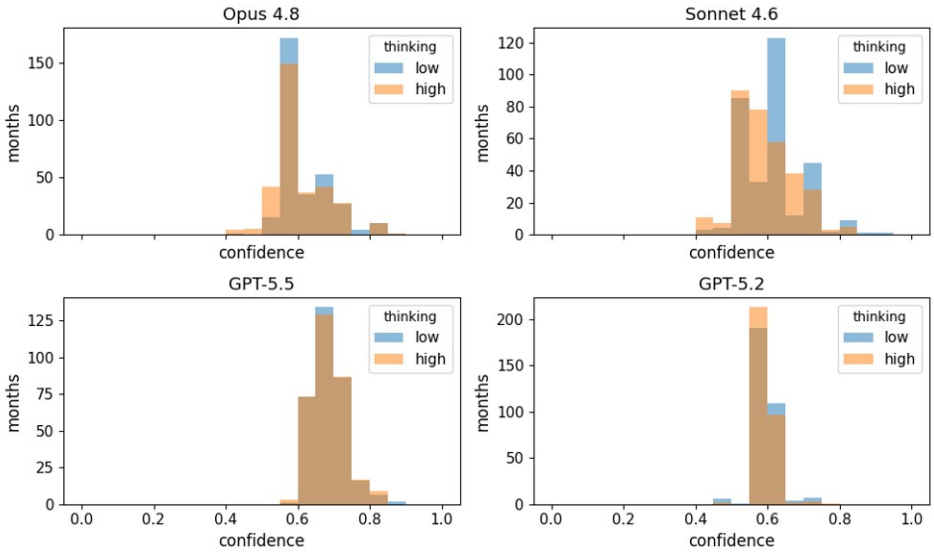
Source: J.P. Morgan Research, Bloomberg Financial L.P.

In the plot below we show the distribution of the model's self-reported confidence in its regime probability distribution, one panel per model, overlaying low and high reasoning. Two patterns stand out, one about the level and another about the spread. The GPT models produce much tighter confidence distribution than the Anthropic models. This means GPT models are the most decisive, which is often one of the reasons why they are used as judges. The distributions are also centered differently; GPT 5.5 is the most confident on average.

Across all experiments we find that confidence is an informative relative signal, but it is not comparable across models. Some families (e.g., GPT) systematically report tighter confidence compared to others.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Figure 8: GPT models produce tighter self-reported confidence distributions than Opus/Sonnet



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Another way to evaluate the regime identification is to calculate the realized Sharpe in each of the regimes. Below we show the realized Sharpe ratio for each model and regime for US Equities and Bonds. Broadly, all the models roughly match our view and the pattern is consistent across models and across low/high reasoning. Equities are rewarded in risk-on regimes. Equities Sharpe is the highest in Goldilocks and remains positive in Reflation. The Sharpe drops in Stagflation and Risk Off. Bonds Sharpe is the highest in Risk off and Stagflation.

Figure 9: Equities peak in Goldilocks, while Bonds peak in Risk Off and Stagflation

Realized annual Sharpe ratio in each regime

		Goldilocks		Reflation		Stagflation		Risk_Off	
		Equities US	Bonds US	Equities US	Bonds US	Equities US	Bonds US	Equities US	Bonds US
Model	Thinking								
Opus 4.8	Low	0.96	0.38	0.20	0.66	-0.04	0.94	0.36	1.54
	High	0.67	0.44	0.55	0.54	0.41	1.33	0.30	1.29
Sonnet 4.6	Low	0.76	0.26	0.83	0.90	0.07	0.84	0.43	1.41
	High	1.03	0.10	0.38	1.28	0.10	0.83	0.42	1.25
GPT-5.5	Low	0.78	0.45	0.34	0.59	0.49	1.02	0.25	1.27
	High	0.99	0.45	0.24	0.40	0.08	1.16	0.30	1.23
GPT-5.2	Low	0.92	0.56	0.42	-0.32	-0.59	1.30	0.39	1.36
	High	0.90	0.49	0.40	0.22	0.41	1.79	0.14	1.04

Source: J.P. Morgan Research, Bloomberg Financial L.P.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Figure 10: Regime frequency by model

Number of months each model assigned to each regime.

Model	Thinking	Goldilocks	Reflation	Stagflation	Risk_Off
Opus 4.8	Low	156	63	36	63
	High	143	77	40	58
Sonnet 4.6	Low	145	37	57	79
	High	114	62	70	72
GPT-5.5	Low	152	49	50	67
	High	148	50	52	68
GPT-5.2	Low	171	40	34	73
	High	155	52	36	75

Source: J.P. Morgan Research, Bloomberg Financial L.P.

The table below shows each model's **forecast for July 2026**, including its confidence, key drivers, and rationale. All four models forecast the same most likely regime, **Reflation**. Reading the same data, the models converge on the same story, above trend growth alongside re-accelerating inflation, in a calm risk environment. They all assign the lowest probability to the Risk Off regime. The shared drivers are:

- **Growth is accelerating** – all three PMIs are above 50, with the US PMI at 54 and rising
- **Inflation is re-accelerating, not falling** – US CPI at 4.2% YoY and climbing, the signal that tips the call from Goldilocks to Reflation
- **Risk conditions are benign** – VIX is calm, HY and HG spreads near multi-year tight and risk-on equity momentum.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

Figure 11: A unanimous Reflation call

As of 30th June 2026 with low thinking mode.

Model	Regime	Confidence	Key drivers	Rationale
Opus 4.8	Reflation	62%	• US ISM 54 rising (+1.6 3m), PMI breadth 3/3 >50 • US CPI 4.2% and accelerating (+1.8 3m, z +2.68) • Fed on hold at 3.75 despite reheating inflation - behind curve • Tight HY/HG spreads, calm VIX, risk-on equity momentum • Curve flattening with rising 10y real yields	Growth is accelerating across DM while US inflation reaccelerates sharply with the Fed on hold, the classic reflation setup. Falling UK/moderating EZ inflation and supportive risk conditions leave some Goldilocks weight; low stress keeps Risk_Off small.
Sonnet 4.6	Reflation	58%	• US ISM at 54 with +6pt 6m acceleration signals above-trend growth momentum • US CPI at 4.2% YoY, +180bps 3m delta — inflation re-accelerating well above target • ECB still tightening (+25bps 3m) while Fed on hold; policy divergence with inflation sticky • Curve flattening sharply (-40bps 6m) with 10y real yield at 93rd percentile — financial conditions tightening • Credit spreads near multi-year tights (HY OAS 15th pct, HG 6th pct) and VIX calm — no imminent stress	Growth is clearly above trend globally but inflation is re-accelerating, particularly in the US, tilting the regime toward Reflation over Goldilocks. Policy is constrained — Fed on hold with inflation rising again — echoing a behind-the-curve dynamic. Stagflation risk is non-trivial given elevated real yields and sticky inflation, but PMI breadth and tight credit argue against near-term growth collapse.
GPT-5.5	Reflation	72%	• PMI breadth is strong at 3/3 above 50 with positive 3m momentum across US, EZ, and UK • Inflation is above target and rising over 3m in the US and EZ, led by US CPI at 4.2% • Policy is not aggressively tightening overall, with Fed and BoE flat over 3m and lower YoY • Financial stress is contained: VIX calm, HY/HG spreads tight, and 3m risk-on momentum is positive	The dominant setup is reflation: broad above-50 PMIs and firm risk appetite alongside renewed US/EZ inflation pressure. Goldilocks is capped because inflation is not falling broadly; Risk_Off remains low given tight spreads and calm vol.
GPT-5.2	Reflation	62%	• PMI breadth strong (3/3 >50) with positive curve • Inflation re-accelerating (US CPI 4.2% YoY; EZ HICP 3.2% YoY, both rising over 3-6m) • Policy looks only mildly restrictive/slow to tighten vs inflation impulse • Financial conditions/risk pricing still benign (VIX calm; HY/HG spreads tight)	Growth indicators are expansionary while inflation is re-accelerating, tilting the next month toward reflation over goldilocks. Risk pricing remains calm with tight credit spreads, limiting risk-off odds. Stagflation risk is present but capped by PMI breadth and stable unemployment.

Source: J.P. Morgan Research, Bloomberg Financial L.P.

Reproducibility – same prompt, three independent runs

LLM outputs are not deterministic, so the response to an identical query can differ from one run to another. To investigate this difference, we ran the same prompt three times with Opus 4.8 and GPT 5.5 using low thinking mode. We first compare the most-likely regime, the one with the highest probability, across runs.

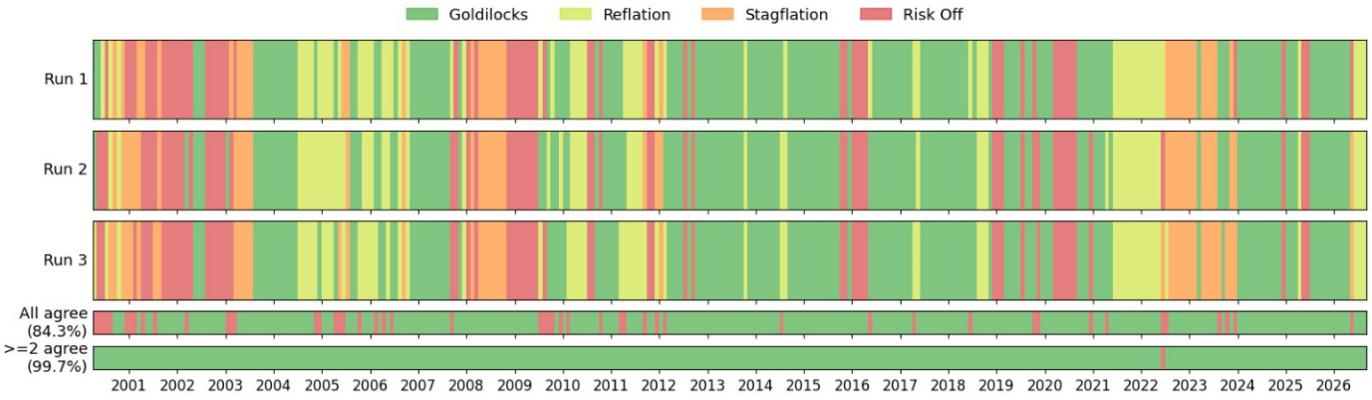
The figures below show the monthly most likely regime for each of the three runs, with two agreement strips beneath, the months where all three runs pick the same regime and the months where at least two out of three agree. The calls are highly stable. All three runs select the same regime around 84% of the time, and at least two out of three almost always agree (~97.7% of the time). We draw the same conclusion with Opus 4.8 and GPT 5.5. The disagreements are not random. They cluster at genuine regime transitions, where the two leading regimes are close in probability and a small sampling perturbation is enough to reorder them. The stability is even clearer at the distribution level. Plotting each regime's probability path for all three runs, the run lines sit almost perfectly on top of one another across the full sample. The model reproduces not just the top-pick regime but the entire probability vector from run to run, with visible separation only at the occasional turning point where two regimes are genuinely close. These results suggest that the run-to-run sampling noise is relatively small.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

J.P.Morgan

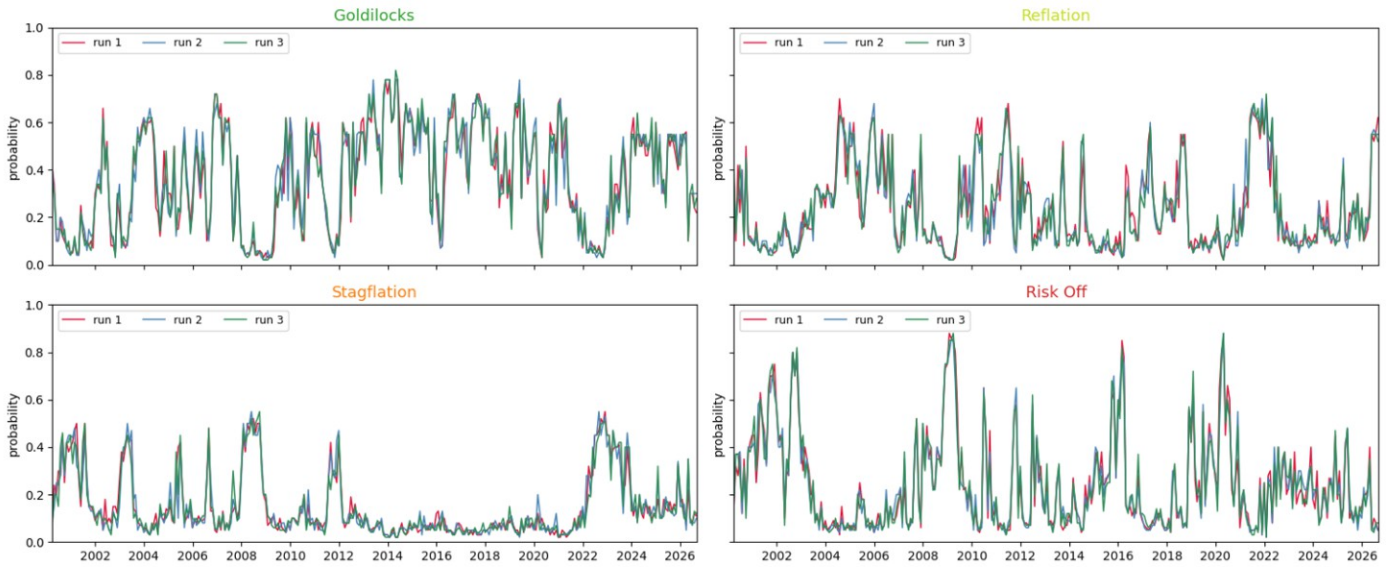
Figure 12: Opus 4.8 – sampling noise is very small, three runs agree ~84%

Monthly most likely regime (highest probability) for three independent runs using the same prompt



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Figure 13: Opus 4.8 – per regime probability is almost identical across three runs

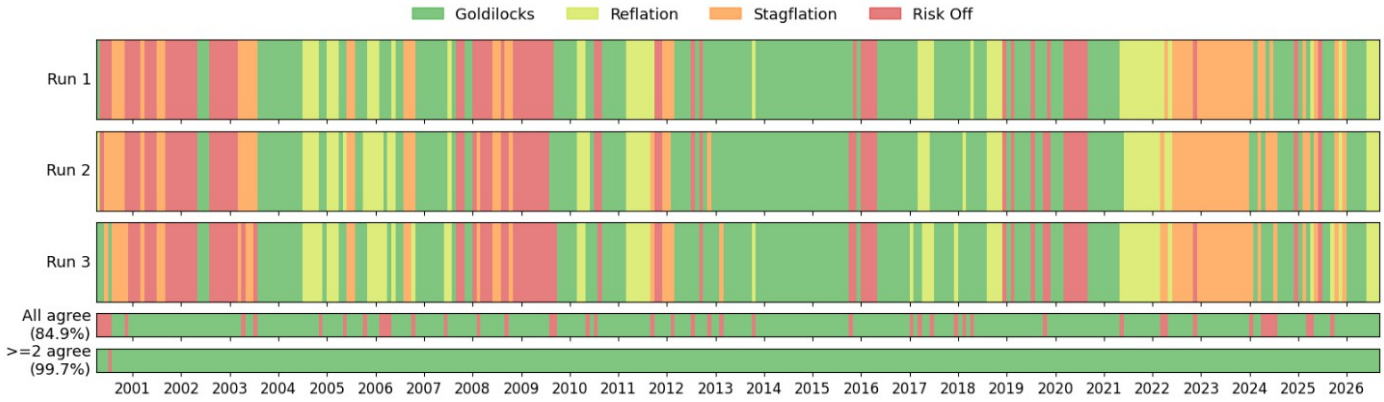


Source: J.P. Morgan Research, Bloomberg Financial L.P.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

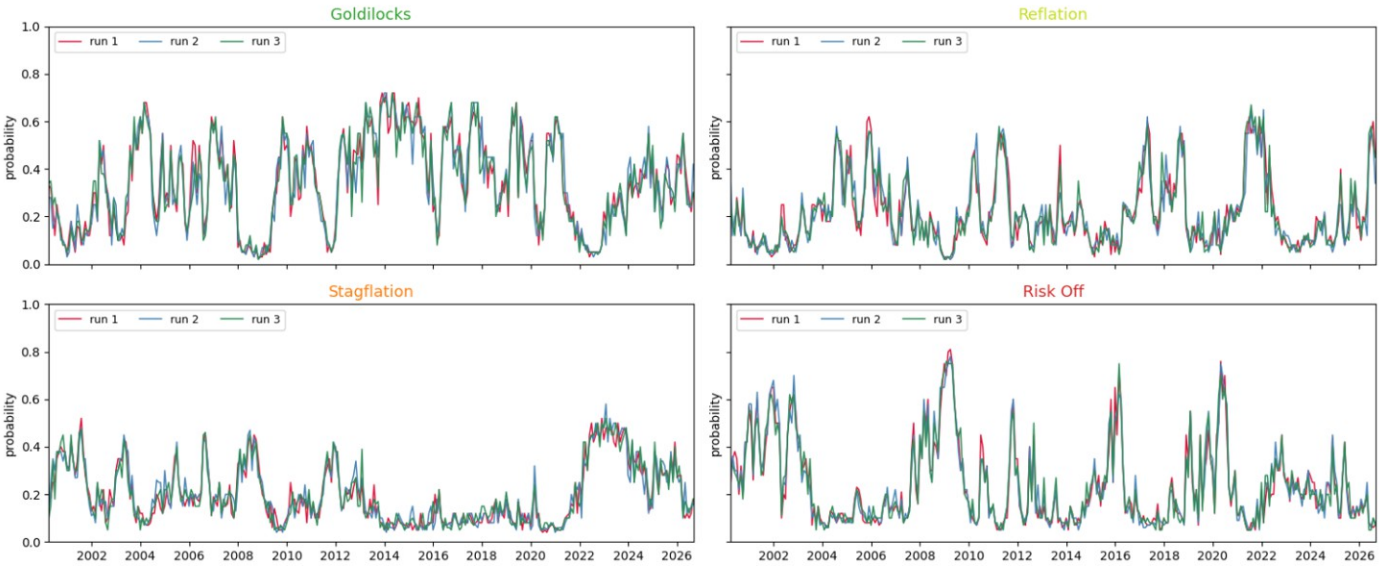
Figure 14: GPT 5.5 – sampling noise is modest, three runs agree ~ 84%

Monthly most likely regime (highest probability) for three independent runs using the same prompt



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Figure 15: GPT 5.5 – per regime probability is almost identical across three runs



Source: J.P. Morgan Research, Bloomberg Financial L.P.

No context, anonymized vs de-anonymized

A central concern for any LLM-based regime classifier is whether the model is genuinely reasoning from the data or simply recalling what happened on a given historical date. Because the models were trained on text that post-dates our sample, a model told only the calendar date could reproduce the known outcome from memory, a form of look-ahead that would not be available in live use. To separate these channels, we evaluate three different prompts on the identical task and output schema, for two models (Opus 4.8 and GPT-5.5) at low thinking:

- **No context** – the model is given only the as-of date (as well as the *regimes definitions* and *output format*) and must name the regime from memory. The prompt

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

is deliberately minimal:

> Given the four macro-regime definitions below, which regime was the global economy in during the month of {as_of}? You are given ONLY the date — no macro or market data. Answer from your own knowledge; if unsure, keep the distribution closer to uniform.

- **Anonymized** – the model is given the point-in-time context pack (macro and market data) with the **date removed**. This is the specification used throughout the paper.
- **De-anonymized** – same prompt as anonymzsed; we just add the date, e.g., *Context pack as of 2008-03-31*

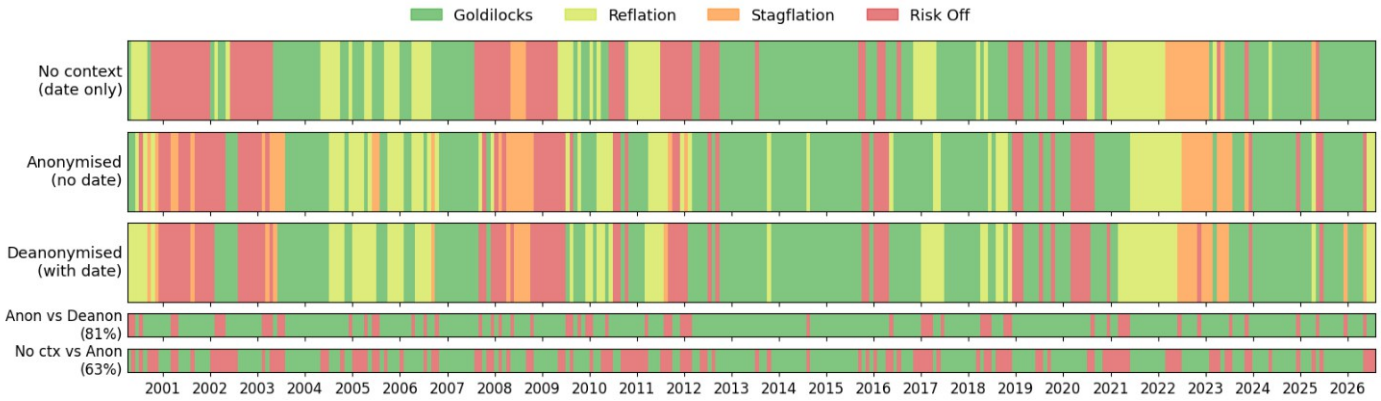
The figures below show the monthly most likely regime and the pairwise agreement strips below. In the first one, we compare the agreement of Anonymized vs De-anonymized prompt, and in the second one we compare the No context vs Anonymized prompt.

Adding context data (market and macro snapshot) materially changes the call; no context and anonymized agree only ~60% of the time.

By contrast, once the data is present (macro and market snapshot), adding the date changes very little. The anonymized and de-anonymized rows agree in ~ 81% of the months for Opus and ~ 87% for GPT 5.5, and the disagreements are concentrated at the regime boundaries rather than being systematic. Given only the date, the model turns defensive far more often than with the context prompt, labelling 81 months Risk Off for Opus and 100 for GPT-5.5, versus just 63–67 under the anonymized (context-only) specification. It also does so sooner; at the onset of the 2008 crisis the memory-only model flags Risk Off in July 2007, a month ahead of the context versions (August and September 2007).

Figure 16: Opus 4.8: Is the model reading the data or remembering the date ?

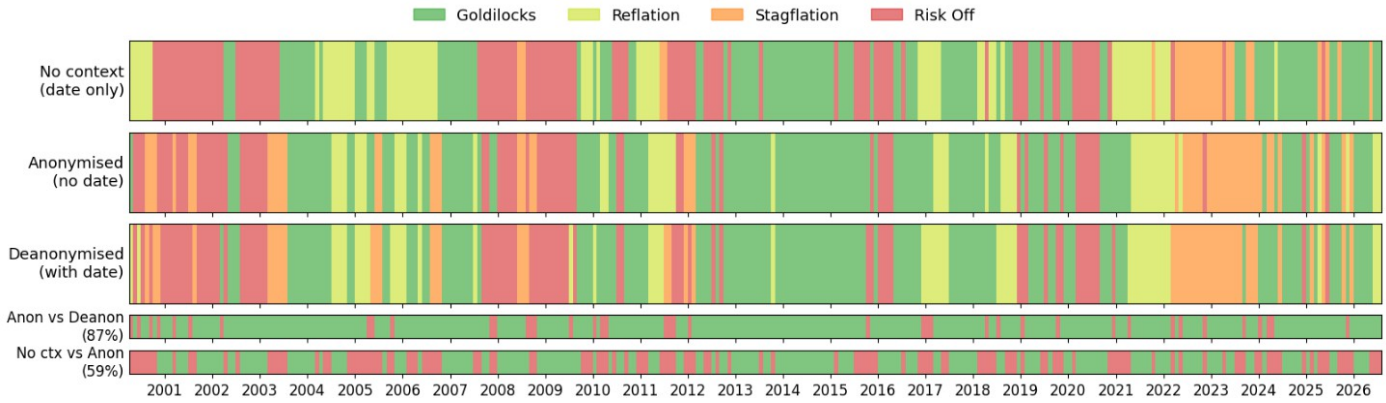
W



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

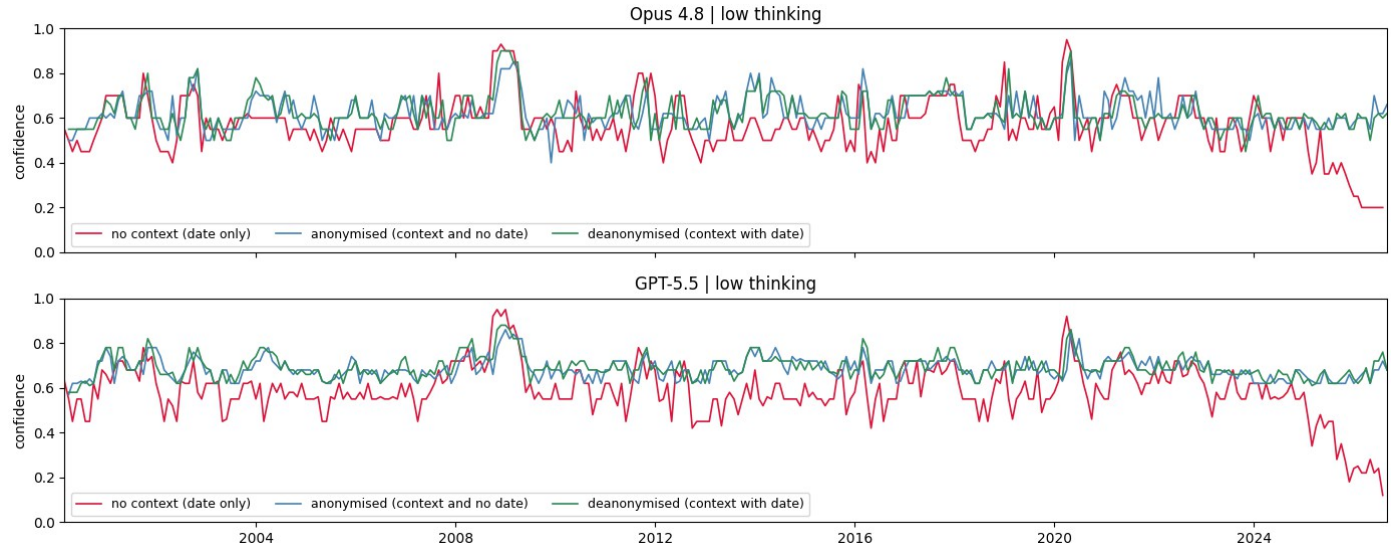
Figure 17: GPT 5.5: Is the model reading the data or remembering the date ?



Source: J.P. Morgan Research, Bloomberg Financial L.P.

The anonymized and de-anonymized prompts report similar stable confidence, with higher confidence usually during Risk off calls. The confidence of the date-only (no context) prompt tends to be lower and more volatile. The one exception is Risk Off periods, where the confidence increases before the context models and is usually higher. A model that "remembers" a crisis year commits to the risk-off call before the data would warrant it. As expected, the confidence of the no-context prompt drops sharply recently as the models have not been trained on that data, while the context prompt confidence stays stable, which is encouraging.

Figure 18: Memory-only confidence collapses out-of-sample, while data-driven confidence holds



Source: J.P. Morgan Research, Bloomberg Financial L.P.

Back-test

To assess the performance of the agentic regime models, we run a simple back-test. The mapping from regime to portfolio is deliberately naïve, and we do not optimize weights on returns. Instead, for each of the four regimes we fix a static regime playbook, a single set of portfolio weights, shown in the table below. Each month we use the most likely regime, provided by the LLM agent, and use the portfolio weights for that regime. This

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

isolates the value of the regime classification. Positions are rebalanced monthly, and we compare the results with two benchmarks, 60/40 and a rule-based benchmark.

The rule-based benchmark applies the same regime playbook weights but derives the regime deterministically from CPI and PMI at each month-end and carries that regime into the next month (i.e., it classifies the current environment and assumes it will persist), rather than forecasting the upcoming month. At each month-end, we look at the growth (PMI) and inflation (CPI YoY) series, decide whether each is rising or falling, based on its 6-month trend (3-month-smoothed level vs 6 months ago), and map that growth×inflation combination to the regime. This is similar to our Market Regime Model; the level is the 3-month moving average and the trend is the 6-month change of the level. Finally, the four regimes are defined as **Goldilocks** (growth up, inflation down), **Reflation** (growth up, inflation up), **Stagflation** (growth down, inflation up), and **Risk-Off** (growth down, inflation down).

Figure 19: Fixed portfolio weights in each Regime

	Goldilocks	Reflation	Stagflation	Risk_Off
Equities US	26 %	22 %	8 %	6 %
Equities EU	12 %	12 %	4 %	2 %
Equities UK	9 %	10 %	4 %	2 %
Equities EM	15 %	18 %	4 %	0 %
Equities total	**62 %**	**62 %**	**20 %**	**10 %**
Bonds US	5 %	3 %	23 %	35 %
Bonds EU	4 %	3 %	19 %	25 %
Bonds UK	4 %	2 %	19 %	20 %
Bonds total	**13 %**	**8 %**	**61 %**	**80 %**
US HG	6 %	5 %	8 %	5 %
EU HG	4 %	3 %	6 %	3 %
US HY	9 %	13 %	3 %	1 %
EU HY	6 %	9 %	2 %	1 %
Credit total	**25 %**	**30 %**	**19 %**	**10 %**

Source: J.P. Morgan Research, Bloomberg Financial L.P.

Below we show the performance of all the agentic models and the benchmarks. The results are robust; all eight models beat the benchmarks. Consistent with the previous results, the low thinking mode runs the same or beats the high thinking counterpart for most of the models. We specify again that **these results are not guaranteed to be genuinely out-of-sample** because of the look-ahead bias introduced by the agentic component. Although the data is lagged and the prompt is date-anonymized, the LLM models are still trained on data after the cut-off point and may implicitly recall the outcome of recognizable historical episodes (e.g., 2008, COVID, etc.).

Thomas Salopek [AC]
(1-212) 834-5476
thomas.salopek@jpmorgan.com

J.P.Morgan

Figure 20: Performance summary

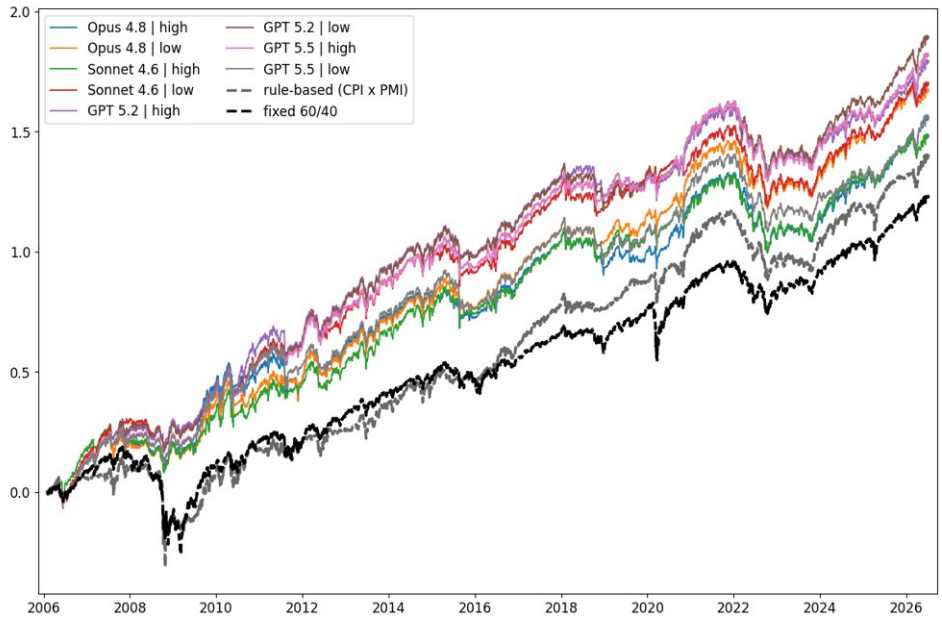
2.5bps transaction cost is included

model	Ann. Vol.	Ann. Ret.	Sharpe	Max DD
GPT 5.2 | low	6.6%	6.2%	0.95	-19.8%
GPT 5.5 | high	6.5%	5.9%	0.91	-21.0%
GPT 5.2 | high	6.7%	6.0%	0.90	-19.4%
Sonnet 4.6 | low	6.2%	5.2%	0.85	-20.1%
Opus 4.8 | low	6.9%	5.7%	0.84	-19.0%
Opus 4.8 | high	7.0%	5.4%	0.78	-22.4%
GPT 5.5 | low	6.7%	5.2%	0.78	-21.0%
Sonnet 4.6 | high	6.3%	4.6%	0.74	-20.0%
rule-based (CPI x PMI)	7.0%	4.8%	0.70	-29.6%
fixed 60/40	9.4%	5.5%	0.61	-37.0%

Source: J.P. Morgan Research, Bloomberg Financial L.P.

Figure 21: Cumulative PnL

The PnL series are scaled to the same volatility in this plot.



Source: J.P. Morgan Research, Bloomberg Financial L.P.