

# Sustainability is not a conscience. It is P&L you have not measured yet.

Does AI Trading Have an ESG Problem? A practitioner's view from Hands-On AI Trading with Python, QuantConnect, and AWS

**+90-95%**

DRAM contract price, QoQ Q1 2026  
TrendForce forecast

**~\$800B**

Hyperscaler capex 2026  
Analyst consensus, Goldman Sachs

**~3×**

Singapore vs Johor colocation  
\$330-475 vs \$110-150 per kW/month

Jiri Pik, CFA · Co-author · Founder & CEO, RocketEdge.com

Source: [TrendForce](#), [Goldman Sachs](#), [Coloprice](#)

# Four reasons to care. All of them material.

## COST

Footprint is the compute bill. Every wasted kWh, GB and byte is alpha given away.

+90–95% DRAM contract price, Q1 2026 · H100 1-year rental \$1.70 → \$2.35 per hour, +40% in five months

## CAPACITY

Compute is rationed and the biggest buyers are served first. The efficient desk runs more experiments per allocated unit. Research velocity is the edge.

AI data-centre electricity +50% in 2025 vs +3% global demand · On-demand H100 capacity sold out across providers

## CAPITAL AND COUNTERPARTIES

The people who allocate to you, lend to you, clear for you and license you are already asking.

60% of LPs have turned down a deal on ESG grounds (Preqin 2024) · EBA ESG-risk rules apply from Jan 2026 · MAS transition plans from Sep 2027

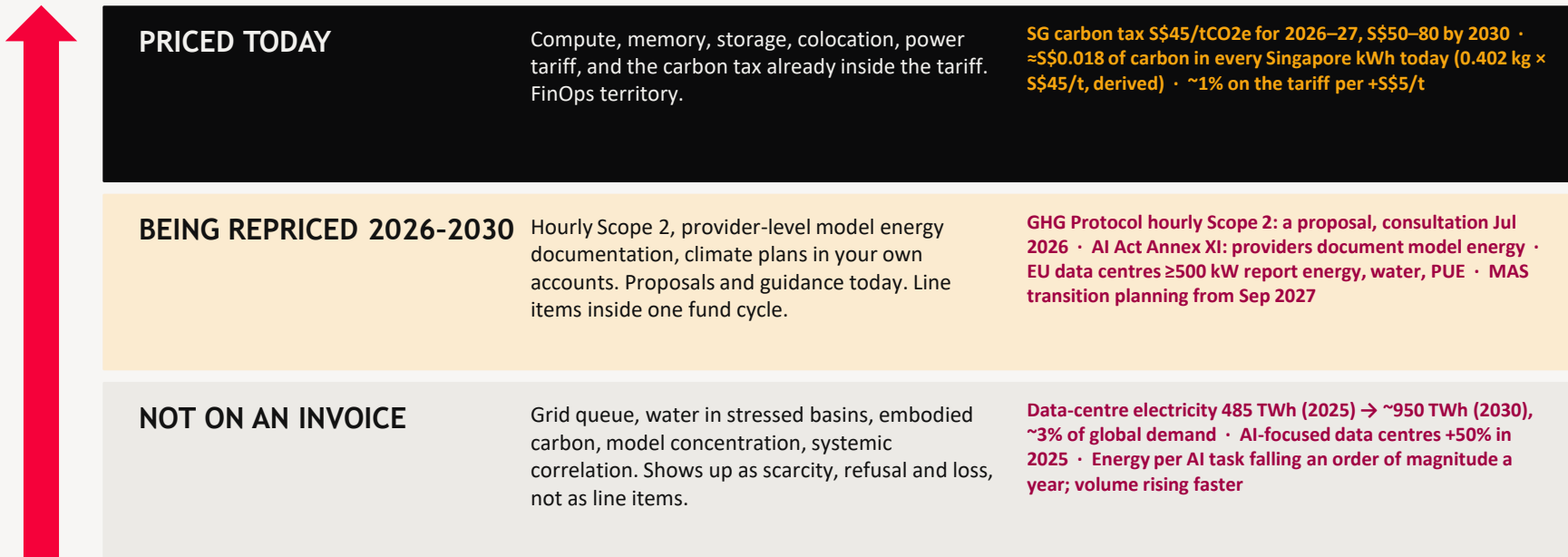
## CONTAGION

Same foundation model, same region, same prompts equals one trade wearing many logos. When it unwinds, it unwinds together.

Two providers = 67% of enterprise LLM spend (Dec 2025) · CFA Institute: "algorithmic monoculture" is systemic

You do not need to believe in ESG to need these numbers. The people who sell you compute, allocate you capital, clear your trades and regulate your licence are already asking — and every one of them is a line on your own P&L.

# No. Cost is where a desk starts. The forward curve is where it ends.



The arrow points up. Manage only the top band and you are hedging the 2024 version of your cost base.

# The tension is already inside CFA Institute

- 2025** CFA Institute softens the ESG-alpha claim in its own certificate curriculum.
- 2025** First negative year for global sustainable-fund flows since 2018: -US\$84bn.
- 2025** 38% of firms report AI in portfolio construction or trading, up from 15%. Adoption, not a washing rate.
- 2026** Q2: US sustainable funds take in nearly \$3bn after 14 quarters of outflows. \$3.1bn of it went into one grid ETF.
- 2026** "Algorithmic monoculture" named a systemic risk in the Algorithmic Market Hypothesis.

Sustainable investing is under pressure to prove it is decision-useful.

AI is becoming the dominant interpretive engine of markets.

Those two facts collide on a trading desk. That is why this session is here and not in a technology track.

Mental model: base-rate shift

# Three columns, one evidence rule, and one number to hold

## Footprint (E)

• • •

## People (S)

• • •

## Control (G)

• • •

**EVIDENCE RULE, UNDER ALL THREE:** model + version · data known at decision time · every trial counted · out-of-sample statistic

**0-5**

Framing ·  
Jiri Pik

**5-20**

What AI actually does in the loop ·  
Philip Sun

**20-35**

Live multi-agent research demo · Jared  
Broad

**35-50**

The ledger and six questions · Jiri Pik

**50-60**

Q&A

Hold this arithmetic for Jared's demo: **40 calls × 3.91 Wh ≈ 156 Wh per agent research loop.** Assumed, not measured. Watch whether his trace gives you a real one.

# Sustainability is not a conscience. It is P&L you have not measured yet.

## ALREADY ANSWERED

Allocated cloud emissions (AWS CCFT: a monthly allocation, not a per-run meter)

Auditable model trail (Git, MLflow)

Point-in-time data lineage

## STILL DUCKED

**No receipt.** Wh per accepted decision. Not disclosed by the models your desk calls.

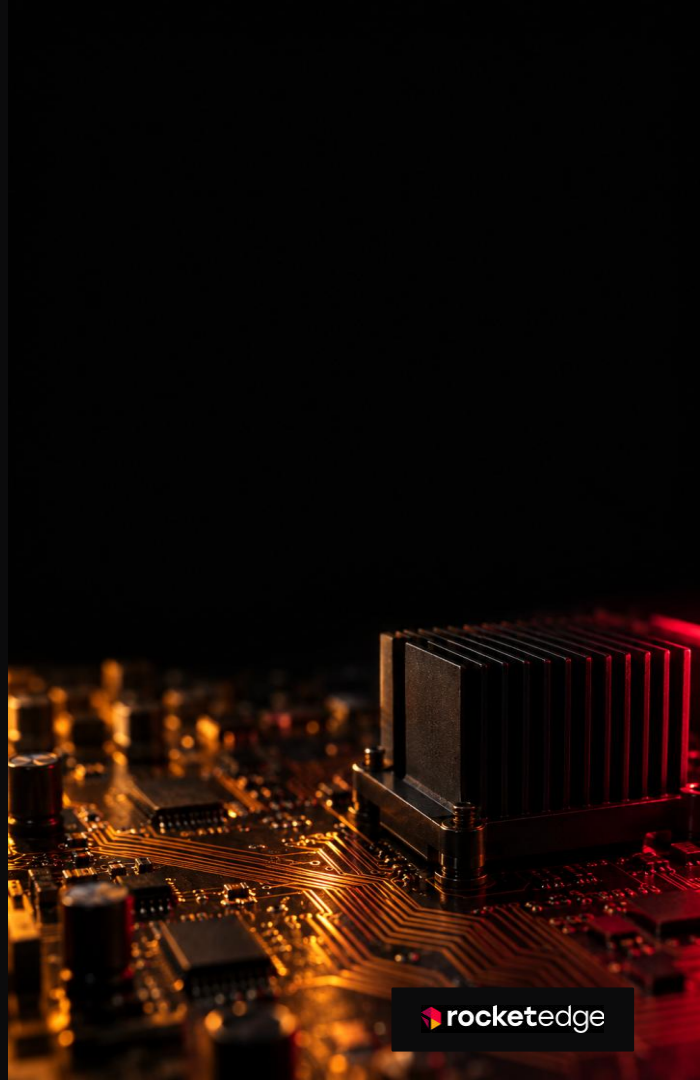
**No person.** Human minutes per accepted decision.

**No stop.** One named owner who can halt it at 2 a.m.

**If a desk cannot answer the right-hand column, that is a diligence signal, not a moral one.**

Mental model: inversion

Source: [AWS CCFT methodology](#)



# FinOps and ESG are cousins, not twins

| Lever                                       | \$ effect per workload                                     | E effect per workload                                                                   | Verdict       |
|---------------------------------------------|------------------------------------------------------------|-----------------------------------------------------------------------------------------|---------------|
| ARM (Graviton) vs Intel                     | ≈20% lower hourly cost; up to 40% better price-performance | Lower kWh per workload                                                                  | Aligned       |
| Lean code, less RAM                         | Lower RAM bill; hedge against DRAM that doubled in Q1      | Fewer instances, lower kWh                                                              | Aligned       |
| Hot S3 to Glacier                           | Multiples off storage cost                                 | Idle disks, lower kWh                                                                   | Aligned       |
| Cheaper model tier: Sonnet 5.5 vs Fable 5.1 | Price ratio 5× (\$2/\$10 vs \$10/\$50 per MTok)            | Energy ratio not published. 5× the price is not 5× the watt-hours.                      | Not disclosed |
| Cheaper, dirtier region                     | Lower \$/kWh                                               | Peninsular Malaysia 0.740 vs Singapore 0.402 kgCO2/kWh = 1.84× (regulator prints, 2024) | Divergent     |
| Bigger model for headline Sharpe            | Higher compute cost                                        | More kWh; more monoculture exposure                                                     | Divergent     |
| Real-time inference for a slow signal       | Higher \$ per query                                        | More kWh; peak load on a rationed grid                                                  | Divergent     |

A DDQ that measures cost per workload but not watt-hours per workload is diligencing half the manager. • Mental model: co-benefits vs trade-offs

# Compute is a line item, not a virtue

~\$800B

Hyperscaler capex 2026, analyst consensus

+90-95%

DRAM contract price, QoQ Q1 2026 (forecast)

+40%

H100 1-year rental, Oct 2025 → Mar 2026 (\$1.70 → \$2.35/hr)

1.84×

Grid CO2 per kWh, Peninsular Malaysia vs Singapore (0.740 vs 0.402 kg)

| Layer            | Decision                                                                                              | Cost lever                                                         | E lever                                                                 |
|------------------|-------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------|-------------------------------------------------------------------------|
| <b>Silicon</b>   | Intel/AMD vs ARM                                                                                      | ≈20% lower hourly; up to 40% better price-performance              | Lower kWh                                                               |
| <b>Code</b>      | Python loop vs vectorised / JIT / Rust                                                                | Fewer GB bought into a market that doubled                         | Fewer instances                                                         |
| <b>Storage</b>   | Hot S3 vs Glacier                                                                                     | Multiples off                                                      | Idle disks                                                              |
| <b>Geography</b> | Singapore (premium, green-regulated) vs Johor (cheaper colo; MY data-centre tariff +10–14%, Jul 2025) | Colo ~3× cheaper; power \$0.281 vs \$0.135 per kWh (market quotes) | 1.84× CO2 per kWh; hourly Scope 2 would make region and hour reportable |

The most sustainable thing a PM can do this quarter is not buy an ESG label. It is halve the memory footprint of the research loop — also the biggest hedge against the DRAM shock. • Mental model: Jevons paradox

# The models on your desk will not show you the watts

| What                                      | Wh per query           | Boundary · label                                                      |
|-------------------------------------------|------------------------|-----------------------------------------------------------------------|
| Google Search, average query              | 0.3                    | Company blog, 2009. No newer official figure.                         |
| Gemini Apps, median text prompt           | 0.24 (0.10 chips only) | Measured, full serving stack incl. idle and cooling. Google, 2025.    |
| ChatGPT, average query                    | 0.34                   | CEO statement, Jun 2025. Method not published.                        |
| Frontier chatbot, median → long reasoning | 0.31 → 3.91            | Modelled, peer-reviewed. Microsoft Research, Joule 2026. 13×          |
| Same open model, reasoning switched on    | 30× avg; 150–700×      | Measured GPU energy per 1,000 prompts. AI Energy Score v2, 2025.      |
| Claude Sonnet 5.5                         | Not disclosed          | Price public: \$2 / \$10 per MTok. Energy is not.                     |
| Claude Fable 5.1                          | Not disclosed          | Price public: \$10 / \$50 per MTok. 5× the price is not 5× the watts. |
| ChatGPT Astra                             | Not disclosed          | No per-model figure published.                                        |

## REBOUND INDEX

Assumed, not measured

**40 calls × 3.91 Wh = 156 Wh**

per agent research loop

**100 loops a night = 15.6 kWh**

≈ S\$4.40 of electricity; ≈ 6.3 kg CO<sub>2</sub> on Singapore's grid

1.0

baseline

0.1

90% less per run

10

90% less × 100 runs

At desk scale the kilowatt-hours are cents. The exposure is the blank rows: a production input with no receipt. Not disclosed is a position — it is short information.

· Mental model: measurement boundary

# AI-washing is greenwashing's twin

38% ← 15%

Firms using AI in portfolio construction or trading, 2025 survey. Adoption, not a washing rate.

\$400,000

SEC's first AI-washing penalties: Delphia and Global Predictions, Mar 2024. Small on purpose. The next ones will not be.

5 inputs

Deflated Sharpe needs trial count, variance across trials, sample length, skew and kurtosis. It does not cure look-ahead.

1

**What is the model?**

Architecture, provider, version, region.

2

**What data did it see?**

Point-in-time? Survivorship-free? Restatement-aware? Training cutoff before the test window?

3

**What survived out of sample?**

Deflated Sharpe with all trials counted; untouched OOS window; live paper track.

**When an agent can produce a plausible backtest in twenty seconds, the deflated Sharpe is the only number in the industry that became more important, not less. • Mental model: base-rate neglect**

# Three questions, and what is actually binding

| Sub-column                | Ask the manager                                                                                                                            | Status of the rule, honestly labelled                                                                                                                                                                                            | The number                                                                       |
|---------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| <b>Validation</b>         | Deflated Sharpe with all five inputs, and the untouched out-of-sample window? How many variants did the agents try before this one?        | Binding: CFA Standard V(A), reasonable basis. Guidance, out of scope for agents: Fed SR 26-2 (Apr 2026) supersedes SR 11-7 and places generative and agentic AI outside it. Your own policy must cover what the letter declines. | Every trial counted. Paper is not live.                                          |
| <b>Authority and stop</b> | Who can halt the agent at 2 a.m.? Does "stop" halt new actions or flatten the book? Show me the last action your controls actually denied. | Proposal: FSB Sound Practices for Responsible Adoption of AI, Jun 2026 consultation. 12 practices, 0 statutes.                                                                                                                   | 1 named owner. 0 live trading credentials in the research sandbox.               |
| <b>Concentration</b>      | Which model, provider, region? When did you last run the fallback, and does a local open-weight model work?                                | Binding on providers, not on you: AI Act Annex XI (documentation), Article 15 (robustness of high-risk systems, not general explainability). Research, not rule: CFA Institute AMH.                                              | 67% of enterprise LLM spend in two providers. us-east-1 down >15 h, 20 Oct 2025. |

**Two-thirds of enterprise model spend in two providers is not diversification. It is a single point of failure with two logos — and the supervisory letter that used to cover it has just told you it does not.**

Mental model: systemic vs idiosyncratic risk

# Social is not charity. It is bench, title and client.

| Exposure      | The number                                                                                                                                      | What a PM hears                                                                           | Evidence to ask for                                                              |
|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| <b>Bench</b>  | 19%: employment of 22–25-year-olds in AI-exposed jobs sits 19% below less-exposed peers. ADP payroll data to Jun 2026; descriptive, not causal. | Key-person risk, inverted: the agent skips the check the junior never learns.             | Human minutes per accepted decision; override log; apprenticeship tasks retained |
| <b>Title</b>  | \$1.5bn: Anthropic settlement with authors over pirated training books, preliminarily approved Sep 2025.                                        | Licensing is a contingent liability on your vendor's balance sheet, until it is on yours. | Licence check on every ingested feed; no client MNPI in unapproved services      |
| <b>Client</b> | \$400,000: SEC's first AI-washing penalties, Mar 2024.                                                                                          | Mis-selling with an AI label. Standard V(A) already covers it.                            | Truthful disclosure; suitability; remediation route; named owner                 |
| <b>Market</b> | 67%: two providers' share of enterprise LLM spend, Dec 2025.                                                                                    | Liquidity is a social good you consume. When every agent sells, nobody is on the bid.     | Model diversity; tested fallback; outage stress test on the book                 |

ILO 2025: 1 in 4 workers globally is in an occupation with some GenAI exposure. Exposure, not a displacement forecast, and not a census of your desk.

**The agent saved four days. Whose four days, and who is still competent when it is wrong? The desk number does not exist until you create it: human minutes per accepted decision. • Mental model: second-order effects**

# Every price-data hazard applies to ESG data, verbatim

| Failure mode               | In price data                                  | In ESG data                                                      | One fix for both                                                                                                         |
|----------------------------|------------------------------------------------|------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| <b>Look-ahead</b>          | Using tomorrow's close today                   | Rating revisions back-filled into history                        | Point-in-time snapshot store                                                                                             |
| <b>Survivorship</b>        | Delisted tickers dropped                       | Delisted issuers dropped from ESG universes                      | Dead-issuer archive                                                                                                      |
| <b>Restatement</b>         | Earnings restated                              | Emissions restated years later                                   | As-was vs as-restated fields                                                                                             |
| <b>Model as look-ahead</b> | A 2024-trained model scored on a 2018 backtest | A model that has read the 2020 restatement rates the 2018 issuer | Point-in-time model: training cutoff before the test window. Masking the date in the prompt does not remove the weights. |

**A quant treatment of ESG data is stricter than the industry median. Would your ESG data vendor pass a point-in-time audit — and would your language model? • Mental model: information set**

# Sustainability is not a conscience. It is P&L you have not measured yet.

## RECEIPT · E

Not small. Unmeasured at the unit that matters.

- kWh per accepted decision, labelled measured / estimated / not disclosed
- Campaign total beside the ratio
- Region, hour and grid factor
- RAM per run; silicon mix

## PERSON · S

Whose four days. Who is still competent when it is wrong.

- Human minutes per accepted decision
- Override log; apprenticeship tasks retained
- Licence status of every ingested feed
- Client remediation route and named owner

## STOP · G

Your correlation risk, and your stop.

- Trial count and the five deflated-Sharpe inputs
- One named owner; the last denied action
- Kill switch vs orderly wind-down, rehearsed
- Provider share of runs; fallback test date

**EVIDENCE, UNDER ALL THREE:** model + version · data known at decision time · every trial counted · point-in-time for price, ESG and the model itself

Binding: CFA Standard V(A) · SG carbon tax | Guidance: Fed SR 26-2 (agents out of scope) | Proposal: FSB 12 practices · GHG hourly Scope 2 | Your own policy: what the letters decline

**A manager either can or cannot fill Receipt, Person and Stop. Binary, not rhetorical. · Mental model: falsifiability**

# Six questions. A blank is an answer.

| # | Question                                                                      | Number to ask for                                                           | Label you will accept                                           | Owner                |
|---|-------------------------------------------------------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------|----------------------|
| 1 | What did your last accepted idea cost, including every discarded trial?       | \$ and kWh per accepted decision; total trials; campaign total              | Measured or estimated. "Not disclosed" is an answer; record it. | Operational DD       |
| 2 | Which silicon runs the research loop, and why?                                | Silicon mix; ≈20% hourly saving on ARM where compatible                     | Vendor quote is fine                                            | Operational DD       |
| 3 | What is RAM per run, and what breaks if memory prices double again?           | RAM per job; stress case at +90% DRAM                                       | Measured                                                        | Risk                 |
| 4 | Which region, which hour, which grid factor?                                  | kgCO2/kWh by region (SG 0.402; Peninsular Malaysia 0.740); hourly placement | Regulator print                                                 | Risk / ESG           |
| 5 | Who can stop the agent at 2 a.m., and does "stop" flatten the book?           | One name; the last denied action; kill switch vs wind-down                  | Demonstrated, not described                                     | Investment committee |
| 6 | Which provider carries your research, and when did you last run the fallback? | Share of runs on the top provider; fallback test date and result            | Tested, with a date                                             | IC / Risk            |

**Two of these on the next quarterly call. A manager who cannot fill the cell in 90 days has told you something about operational maturity, not about virtue. • Mental model: revealed preference**

# Sustainability is not a conscience. It is P&L you have not measured yet.

- 1 When memory is up 90%, capex is \$800 billion, on-demand GPUs are sold out and Singapore quotes three times Johor — the efficient choice, the resilient choice and the sustainable choice have converged. What is stopping your managers from making it the default before the regulator does?
- 2 If two providers carry two-thirds of the market's model spend, is that in the risk section of your report — or hiding in the AI section?

**No receipt. No person. No stop.** The three blanks the book teaches you to fill.

Jiri Pik, CFA · [jiri@jiripik.com](mailto:jiri@jiripik.com) · [RocketEdge.com](https://RocketEdge.com)

[Hands-On AI Trading with Python, QuantConnect, and AWS](#) · Pik, Chan, Broad, Sun · Wiley



The book and the code

[amzn.to/4fI0hkU](https://amzn.to/4fI0hkU)

Pik · Chan · Broad · Sun · Wiley