Priced in Dollars: Token Costs, Exchange Rates and Two-Speed AI Diffusion in Emerging Economies

Nov 1, 2026 · @Pawan Maheshwari

Abstract

Cross-country indices link the slow diffusion of generative AI in emerging economies mainly to income, skills and connectivity. We study a firm-level channel these indices cannot observe: currency mismatch. Emerging-market firms buy the enterprise AI stack — cloud infrastructure, model inference and vendor support — at prices set in US dollars or other hard currencies, while their revenues and annual IT budgets are set in local currency. We develop a model of complementary adoption for firms whose AI workloads require cloud migration or equivalent infrastructure modernisation, with spending limited by a budget fixed in local currency. The model gives two adoption thresholds in the exchange rate, one set by profitability and one by the budget; shows that falls in token prices can be offset by modest depreciation because inference is a small share of stack cost; and implies that, once the budget binds, inference volumes become locally more than unit-elastic to the exchange rate. Because local pricing is more established in consumer AI subscriptions than in the enterprise services we examine, the model also predicts two-speed diffusion. Using monthly rand exchange rates for 2015–2026, our historical rolling-window analysis shows that, under an unhedged, fully dollar-denominated baseline, the average local-currency cost exceeded the planning rate by more than 10% in 16% of 12-month rolling planning windows and 27% of 24-month windows. Because the windows overlap, these are descriptive frequencies, not probabilities of future overruns. A field observation from a large South African retailer illustrates the mechanism. The evidence is descriptive; we set out the predictions and data needed for causal tests.

JEL codes: O33, F31, D22, L86, G31

Keywords: artificial intelligence, technology diffusion, exchange rates, dominant-currency pricing, cloud computing, budget constraints, emerging markets

1. Introduction

A large South African retailer recently put on hold a plan to move its point-of-sale (POS) systems onto Google Cloud Platform. The migration was the foundation of its AI roadmap: store-level forecasting and personalisation need transaction data in the cloud. The project

was not stopped for lack of demand, skills or connectivity. It was stopped because implementation and running costs, priced in US dollars, and support from a European partner, priced in pounds sterling, exceeded an IT budget approved in rand.

Cases like this are invisible in current measures of AI diffusion. The Anthropic AI Usage Index finds that a 1% higher GDP per working-age person is associated with 0.7% higher per-capita use of Claude (Appel et al., 2025). Microsoft's telemetry-based AI User Share shows the gap between the Global North and South widening, from 10.6 to 12.6 percentage points between the second half of 2025 and the second quarter of 2026 (Microsoft, 2026b). Both measures are dominated by consumer use, and both are naturally read as reflecting income and capabilities.

We ask how the currency in which the enterprise AI stack is priced shapes adoption by firms in emerging economies. We use **AI stack** to mean the cloud infrastructure, model inference and vendor support a firm buys to run AI in production, and **hard-currency pricing** to mean list prices set in a currency other than the firm's own, whether or not the invoice is converted to local currency. Our argument is that hard-currency pricing creates a currency mismatch between stack costs and local-currency budgets that income per head cannot capture: two countries with the same income can face very different exchange-rate paths. We call the resulting pattern currency-constrained AI diffusion: access to AI need not translate into enterprise adoption when the complementary infrastructure is priced in foreign currency and financed from rigid domestic budgets.

The paper makes three contributions. It applies the dominant-currency paradigm in international prices (Gopinath et al., 2020) to enterprise AI, a market in which, for the hyperscaler and frontier-model services we examine, hard-currency pricing is the norm. It develops a model of sequential, complementary adoption under a local-currency budget constraint, with four propositions on adoption thresholds, the limited reach of token-price cuts, budget rationing and two-speed diffusion. And it measures the size of the currency shock facing South African firms, with sensitivity analysis across planning horizons, contingencies, currency mixes and hedging, and an illustrative field observation.

We do not claim that currency explains most cross-country variation in AI use, and the evidence here does not identify a causal effect on adoption. We show that the channel is large enough to matter for firm budgets and set out the predictions and data that would test it. Section 2 reviews related work, Section 3 describes how the AI stack is priced, Section 4 presents the model, Sections 5 and 6 describe the measurement and evidence, Section 7 discusses limitations and a research agenda, and Section 8 concludes.

2. Related literature

The paper draws on six literatures, each supplying one element of the model.

AI as a general-purpose technology. AI's returns depend on complementary investment (Bresnahan and Trajtenberg, 1995; Agrawal, Gans and Goldfarb, 2019), and such complements delay measured productivity gains (Brynjolfsson, Rock and Syverson, 2021). We treat cloud migration of core data, or equivalent infrastructure modernisation, as the first complement for enterprises that lack the data infrastructure and scalable computing production AI requires. Our addition is that the technology and its complement are priced in the same foreign currency, so their costs move together.

Enterprise technology adoption. Firm-level IT adoption depends on organisational complements and management practices (Bresnahan, Brynjolfsson and Hitt, 2002; Bloom, Sadun and Van Reenen, 2012), and adoption of new technologies is slow and uneven even where benefits are clear (Hall and Khan, 2003). Survey evidence for the United States shows AI adoption concentrated in large, young and high-capability firms (McElheran et al., 2024). This work explains which firms adopt in a given country; it is silent on the price of the technology in local currency, which is our focus.

Cross-country diffusion. Adoption lags explain much of the cross-country income gap (Comin and Hobijn, 2010). The generative AI indices extend this evidence, finding steep income gradients and, in the latest data, growing concentration across countries (Appel et al., 2025; Anthropic, 2026; Microsoft, 2026a, 2026b). Appel et al. (2025) also estimate an elasticity of enterprise API use with respect to task cost of −0.29, which they describe as weak price sensitivity. That estimate is in dollars, for firms already using the API; it does not speak to firms that have not yet migrated.

Dominant-currency pricing. Most trade is invoiced in a few currencies, mainly the dollar, so exchange-rate changes pass through almost fully to importers' local-currency costs in the short run (Gopinath et al., 2020; Burstein and Gopinath, 2014). Imported inputs matter for firm productivity (Halpern, Koren and Szeidl, 2015), so their local-currency cost shapes technology upgrading. To our knowledge, dominant-currency pricing has not been applied to enterprise AI.

Financial constraints and currency mismatch. Investment is sensitive to internal funds when external finance is costly (Fazzari, Hubbard and Petersen, 1988). Firms with costs or liabilities in foreign currency and revenues in local currency cut investment after devaluations (Aguiar, 2005). Our budget constraint is an internal version of this: a fixed local-currency allocation that cannot absorb a rise in hard-currency costs within the year.

Investment under uncertainty. Irreversible investment carries an option value of waiting that rises with uncertainty (Dixit and Pindyck, 1994), and uncertainty shocks delay investment (Bloom, Bond and Van Reenen, 2007; Bloom, 2009). Cloud migration is largely irreversible and its future running costs depend on the exchange rate, so exchange-rate volatility should raise the bar for migrating. We note this extension in Section 4 but keep the core model static.

3. How the AI stack is priced

The distinction that matters is between **local billing** — a hard-currency price converted to local currency on the invoice — and **local pricing** — a price set for the local market. For the hyperscaler and frontier-model services we examine, enterprise prices are set in hard currency and at most locally billed, whereas local pricing is more established in consumer AI subscriptions. Locally priced enterprise contracts may exist through negotiation or resellers, but they appear less established and less transparent.

Cloud and inference. Google Cloud converts US dollar list prices into supported local currencies at rates set at the start of each month; the rand is not a supported billing currency, so South African accounts may be charged in dollars (Google Cloud, 2026). Microsoft shows Azure plan prices in dollars only and publishes exchange rates that partners use to bill in their own currency (Microsoft, 2025). AWS accepts payment in rand among other currencies but invoices in dollars by default (Amazon Web Services, 2026). Model inference sold through these platforms, or directly by model providers, is priced per million tokens in dollars. In every case the customer bears the exchange-rate risk, with the conversion reset monthly or at each payment.

Vendor and integration support. Enterprise software in emerging markets is often implemented and supported by foreign partners whose contracts are priced in their home currency. In the case we study, support was priced in sterling, creating a second hard-currency exposure.

Consumer AI. In August 2025 OpenAI moved ChatGPT subscriptions in India to rupee prices and launched a lower-priced plan, ChatGPT Go, at ₹399 a month (TechCrunch, 2025). Rupee prices were not straight conversions: the Plus plan was set at ₹1,999, above the dollar price at the prevailing rate, while Go was priced far below it. These are prices as reported at launch in August 2025; we have not re-verified them since, and displayed consumer prices and their tax treatment may vary by billing channel and over time. This is local pricing in the full sense. We found no comparable published local pricing for the enterprise cloud and API services examined, and this asymmetry underlies Proposition 4.

4. Model

4.1 Set-up and assumptions

We model firms whose production AI workloads need infrastructure they do not yet have: where enterprises lack the data infrastructure and scalable computing that production AI requires, cloud migration or equivalent infrastructure modernisation becomes a prerequisite investment. Such a firm can adopt cloud infrastructure (C) and then AI services (A). Per-period benefits in local currency are V_C from cloud and V_A from AI on

top of cloud; let V = V_C + V_A. Per-period costs are c_C (cloud) and c_A (inference) in dollars and s (support) in pounds; let c = c_C + c_A. With e the local-currency price of a dollar and g of a pound, the local cost of the AI stack is:

$$K(e,g) = e\,c + g\,s$$

The firm's IT budget is fixed in local currency for one fiscal year, set at planning-date rates (e_0, g_0) with a contingency share β:

$$B = (1+\beta)\,(e_0\,c + g_0\,s)$$

We maintain four assumptions throughout.

- **A1 (hard-currency pricing).** Dollar and sterling list prices are fixed over the planning horizon; vendors do not price to market.
- **A2 (local-currency benefits).** V does not depend on e, as for a firm selling non-tradable goods or services at home. Remark 1 relaxes this.
- **A3 (sticky budget, no hedging).** B cannot be revised within the year, and the firm does not hedge. Remark 2 relaxes hedging.
- **A4 (sequential complementarity).** For the workloads modelled, AI can be used only once the cloud, or equivalent modernised infrastructure, is in place. Firms that already have such infrastructure, or that run AI on existing private-cloud, hybrid or on-premises systems, are outside the model's scope.

For analytical tractability, we model the cloud and AI investments as a combined adoption decision, while preserving their sequential technological dependency: the AI benefit V_A accrues only if the cloud investment is made. A full dynamic treatment of the two investment stages is left for future research.

4.2 Two adoption thresholds

Proposition 1. The firm adopts the stack in a given year if and only if $e \leqslant \min(e^V, e^B)$, where

$$e^{V} = \frac{V - g\,s}{c}, \qquad e^{B} = \frac{B - g\,s}{c}$$

The first threshold comes from profitability ($V \geqslant K$), the second from budget feasibility ($K \leqslant B$). Substituting B, and holding the pound at its planning rate ($g = g_0$), gives

$$\frac{e^{B}}{e_0} = 1 + \beta\,(1+\sigma), \qquad \sigma = \frac{g_0\,s}{e_0\,c}$$

where σ is the ratio of sterling to dollar costs at planning rates. The budget binds once the dollar has appreciated against the local currency by more than β(1+σ) since the planning date.

Two regimes follow. When e^B < e^V, the firm is **budget-constrained**: the project is profitable at the current rate but cannot be funded, and is deferred. When e^V ⩽ e^B, it is **profitability-constrained**: the project no longer pays. Income-based diffusion measures cannot distinguish these regimes, and the field observation in Section 6 falls in the first.

Remark 1 (natural hedges). If a share λ of the firm′s benefits accrues in hard currency — an exporter or a business-process outsourcer, say — e^V rises with e and the profitability threshold is crossed later. The budget threshold still binds unless the budget is also indexed. The model therefore predicts lower exchange-rate sensitivity of adoption among exporters.

4.3 Token prices and the exchange rate

Let θ_A = e c_A / K be inference′s share of stack cost, θ_D = e c / K the dollar share and θ_S = g s / K the sterling share, with θ_D + θ_S = 1. Suppose dollar token prices fall by a fraction x, holding token quantities fixed.

Proposition 2. Under A1 and A3, the stack′s local cost is unchanged by the token price cut if the local currency depreciates by d, where:

(a) if it depreciates equally against the dollar and the pound,

```
d = \frac{\theta_A\,x}{1 - \theta_A\,x}
```

(b) if it depreciates against the dollar only,

```
d = \frac{\theta_A\,x}{\theta_D - \theta_A\,x}
```

With an illustrative inference share of θ_A = 0.2, halving token prices (x = 0.5) is offset by an 11.1% depreciation in case (a). In case (b) with a dollar share of θ_D = 0.7, the offset is 16.7%. These fixed-volume figures isolate the mechanical price effect; they are not the full expected expenditure response, because firms use more tokens when prices fall. If token volume responds to price with elasticity η, inference spending falls by x′ = 1 − (1 − x)^(1+η) rather than by x, and x′ replaces x in the formulas above. Appel et al. (2025) estimate an elasticity of API use with respect to task cost of −0.29. Applied here, halving token prices raises volume by 22.3% and cuts inference spending by 38.9% rather than 50%, and the offsetting depreciation falls to 8.4% in case (a) and 12.5% in case (b). The transfer of that estimate is uncertain: it relates usage frequency to a task-cost index for firms already using the API, and may not carry over to token-price changes or to all enterprise workloads. The offset equalises spending, not welfare, since the extra tokens

have value to the firm. The proposition explains why the rapid fall in dollar token prices — between 9× and 900× a year at constant performance, depending on the task (Epoch AI, 2025) — need not lower the cost of enterprise AI in local currency by much: the falling component is a small share of the stack, and the exchange rate moves the whole stack.

4.4 Rationing once the budget binds

Within a year, suppose the firm has committed dollar cloud costs F and sterling support s, and chooses an inference volume q at dollar price p, with local-currency benefit u(q), increasing and concave. Unconstrained demand q^u(e) solves u′(q) = e p. The budget constraint e(F + p q) + g s ⩽ B implies a maximum volume q^B(e), and actual use is the smaller of the two:

```
q(e) = \min\left\{\, q^{u}(e),\; q^{B}(e) \,\right\}, \qquad q^{B}(e) =
\frac{B - g\,s - e\,F}{e\,p}
```

Proposition 3. (i) There is a threshold ê at which q^u(ê) = q^B(ê). Below it, inference responds to the exchange rate only through the unconstrained elasticity, which existing estimates suggest is small. Above it, the elasticity of use with respect to e is

```
\frac{\partial \ln q^{B}}{\partial \ln e} = -\frac{1}{1-\varphi}, \qquad
\varphi = \frac{e\,F}{B - g\,s}
```

For 0 ⩽ φ < 1 and q^B(e) > 0, this budget-constrained local elasticity is at most −1 and grows in magnitude with φ, the share of the hard-currency budget net of sterling costs that is committed to fixed cloud costs. At φ = 0.8 it equals −5 at that operating point: a small depreciation from there cuts inference volume by about five times its percentage size. It is a local elasticity, not a universal response to every exchange-rate change or operating regime; as e rises, φ rises and q^B(e) falls towards zero.

(ii) A one-off migration project that would breach the budget at the current rate is deferred. If the budget is re-planned each year at the new rate, the delay lasts one budget cycle. If the budget is also capped as a share of local-currency revenue, the delay lasts until the rate falls back below e^B, until dollar prices fall enough, or until revenue grows enough.

The first part explains why firms report "cutting usage" in response to depreciation even though measured price elasticities are low: the measured elasticities come from firms off the budget constraint. The second part explains why migration timing is lumpy and tied to the budget calendar.

4.5 Two-speed diffusion

Let consumer and enterprise adoption rates be a_c = f_c(Y, P_c) and a_e = f_e(Y, K(e, g)), where Y is income per head and P_c is the local-currency price of consumer AI. We maintain three assumptions: (i) P_c is set in local currency and does not move with e over the period considered, so a_c does not depend on e once Y is held fixed; (ii) enterprise adoption falls with stack cost, ∂f_e/∂K < 0; and (iii) income is held fixed.

Proposition 4. Under (i)–(iii),

```
\frac{\partial \ln(a_e/a_c)}{\partial \ln e} = \frac{\partial \ln f_e}
{\partial \ln K}\,\theta_D < 0
```

since ∂ln K/∂ln e = θ_D > 0 and ∂ln a_c/∂ln e = 0. Economies with larger depreciations should show lower enterprise use relative to consumer use, and consumer-weighted indices will overstate how far AI has reached production. Depreciations also move income in practice; the ratio removes income effects only if consumer and enterprise adoption respond to income in the same proportion, which is why the result is stated holding Y fixed.

Remark 2 (hedging and uncertainty). Hedging a share h of the hard-currency exposure scales the budget overshoot by (1 − h), at the cost of the hedge, which for emerging-market currencies includes a forward premium roughly equal to the interest differential; basis risk and incomplete coverage also remain. Exchange-rate volatility also adds an option value to waiting before an irreversible migration (Dixit and Pindyck, 1994), so with uncertainty the effective adoption threshold lies below min(e^V, e^B). Both effects are outside the static model; the second strengthens its predictions.

5. Data and measurement

Because firm-level data linking billing currency to adoption are not public, we measure the size of the currency shock a firm's budget must absorb, and treat adoption data as context.

Exchange rates. Annual averages of daily spot rates for 2015–2025 come from the Federal Reserve H.10 release via FRED (series AEXSFUS, AEXUSUK, AEXINUS, AEXBZUS). Monthly averages for January 2015 to September 2026 (EXSFUS, EXUSUK; 141 months) are used for the budget analysis. Rand per pound is rand per dollar times dollars per pound.

Stack Cost Index (SCI). The local-currency cost of a constant hard-currency bill, relative to a base year b:

```
SCI_{t} = 100 \times \frac{e_{t}}{e_{b}}
```

Holding list prices constant isolates the currency channel (assumption A1). We report it for three base years, since the choice of base changes the measured shock.

Budget overshoot. For a budget planned in month m at the prevailing rate, with a share w of hard-currency costs in dollars and 1 − w in sterling, the overshoot over a horizon of H months is measured two ways:

$$O^{avg}_{m,H} = \frac{1}{H}\sum_{k=1}^{H} r_{m,k} - 1, \qquad O^{max}_{m,H} = \max_{1\le k\le H} r_{m,k} - 1, \qquad r_{m,k} = w\,\frac{e_{m+k}}{e_{m}} + (1-w)\,\frac{g_{m+k}}{g_{m}}$$

The average measure applies to a bill paid monthly over the budget year; the maximum measures the worst single month, relevant where a large one-off payment, such as a migration milestone, falls in that month. In the model's terms, the budget binds when the overshoot exceeds the contingency β. The planning rate is the monthly average spot rate in month m. We compute the share of planning windows in which the overshoot exceeds β of 5%, 10%, 15% and 20%, for horizons of 12 and 24 months and dollar shares w of 1.0, 0.8 and 0.6. The 24-month horizon measures the cumulative exposure of a multi-year commitment, such as a cloud migration, against its original planning rate; it is not necessarily a single, unrevised two-year operating budget. The hedge case assumes a proportional hedge of 50% of the exposure at the planning rate, so each overshoot is halved; hedge costs, including the forward premium, are not deducted (Remark 2). Because adjacent planning windows share most of their monthly observations, they are not independent: the reported shares describe historical rolling-window exposure, not independent event probabilities, and should not be read as forecasts of future budget overruns. Appendix A and the supplementary workbook give the data and formulas needed to reproduce every figure.

Adoption context. Microsoft's AI User Share is the share of people aged 15–64 who used a generative AI product in the period, estimated from device telemetry (Microsoft, 2026b). Anthropic's index compares each country's share of Claude use with its share of the working-age population (Appel et al., 2025; Anthropic, 2026). Both are dominated by consumer use, so neither can test Proposition 4 on its own.

Field observation. The retailer case comes from the author's direct involvement as a delivery leader for enterprise technology programmes in South Africa. It is reported anonymously, without commercially sensitive figures, and is used to illustrate the mechanism, not to estimate it.

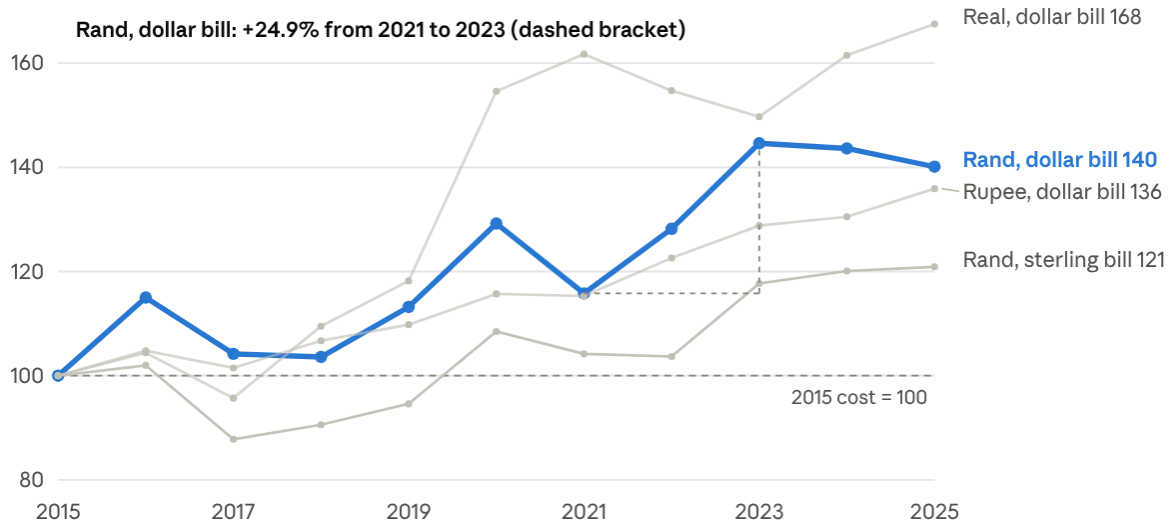
6. Evidence

The currency shock is large relative to plausible budget contingencies, concentrated in short windows, and only partly reversed by the rand's recovery in 2026.

6.1 Stack Cost Index

A fixed dollar AI bill cost South African firms 40% more in rand by 2025 than in 2015

Local-currency cost of a constant foreign-currency bill, 2015 = 100, annual average exchange rates



FRED / Federal Reserve H.10 annual averages (AEXSFUS, AEXUSUK, AEXINUS, AEXBZUS), 2015–2025; rand per pound computed as rand per dollar × dollars per pound

Measured against 2015, a dollar bill cost 40.1% more in rand in 2025; the rand's shocks came mainly in 2015–16, 2020 and 2022–23. A dollar bill also cost 35.9% more in rupees and 67.5% more in reais over the same period, so South Africa is not an outlier. The dollar and sterling exposures do not diversify each other reliably: in 2023 the rand weakened 12.8% against the dollar and 13.4% against the pound.

The measured shock depends on the base year and end point, and the rand strengthened in 2026: its January–September average was 8.5% below the 2025 average.

Base period	Dollar bill, 2025 avg	Dollar bill, Sep 2026	Sterling bill, 2025 avg	Sterling bill, Sep 2026
2015 average	+40.1%	+27.2%	+20.8%	+11.5%
2019 average	+23.8%	+12.3%	+27.8%	+17.9%
2021 average	+20.9%	+9.7%	+15.9%	+7.0%

Rand cost of a constant hard-currency bill relative to the base period, from monthly FRED series (EXSFUS, EXUSUK). The annual-average series gives +24.9% for 2021 to 2023 (dollar bill); the monthly-average series gives +24.7%.

6.2 Budget overshoot

For a budget planned at the prevailing rate, the typical year brings little or no overshoot; the risk sits in the tails. In 47% of 12-month rolling windows the average dollar cost came in below plan. But in 16% of windows it exceeded plan by more than 10%, and for 24-month windows the figure rises to 27%. The 24-month measure is the cumulative exposure of a multi-year commitment against its original planning rate; it does not assume a single two-year budget, since many firms re-approve budgets annually while a migration runs. These are descriptive historical frequencies across overlapping windows, not probabilities of future overruns.

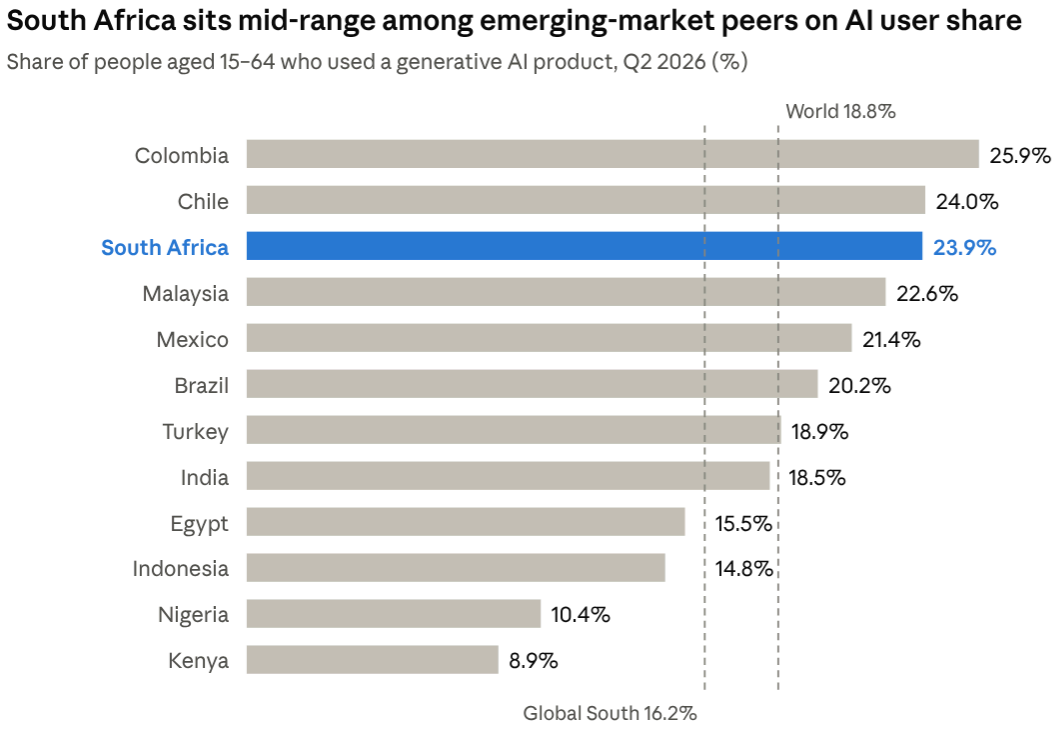
Dollar share w	Horizon	Measure	Median	Below plan	> 5%	> 10%	> 15%	> 20%
1.0	12 months	Average	0.5%	47%	35%	16%	5%	0%
1.0	12 months	Worst month	7.5%	17%	60%	40%	29%	22%
1.0	24 months	Average	4.2%	36%	44%	27%	15%	2%
1.0	24 months	Worst month	15.1%	9%	74%	62%	50%	44%
0.8	12 months	Average	1.0%	46%	34%	16%	4%	0%
0.8	24 months	Average	4.2%	35%	47%	26%	9%	0%
0.6	12 months	Average	1.1%	44%	33%	16%	3%	0%
0.6	24 months	Average	3.8%	37%	49%	22%	8%	0%

Share of planning windows in which the overshoot exceeds each contingency level β; "Below plan" is the share with a negative overshoot. All figures are reproduced in the Summary sheet of the supplementary workbook. Planning months January 2015 to September 2025 (12-month horizon, 129 windows) and to September 2024 (24-month horizon, 117 windows). Windows overlap and are not independent observations, so the shares describe historical rolling-window exposure rather than event probabilities.

Several results stand out. Multi-year commitments such as cloud migration carry far more cumulative exposure, against their original planning rate, than annual run costs. A one-off payment timed badly within the year faced an overshoot above 10% in 40% of 12-month windows. Mixing in sterling costs barely reduces the risk, because the rand tends to weaken against both currencies at once. Under the stated assumptions, a proportional 50% hedge halves the measured exposure: no 12-month window then shows an average overshoot above 10%, and 16% show one above 5%. This reduces measured exchange-

rate exposure; it does not remove the economic cost of currency risk, since forward premiums, basis risk, imperfect hedge coverage and other contractual risks remain (Remark 2).

6.3 Adoption context



Microsoft Global AI Diffusion Report, Q2 2026 (September 2026); selected emerging economies

South Africa's AI User Share is above the Global South average and in the middle of its emerging-market peers. It was 23.9% in Q2 2026, up from 23.1% in Q1 (Microsoft, 2026b). These figures measure consumer-weighted use of generative AI, not enterprise deployment. They cannot show whether firms have moved core data and AI workloads to the cloud, which is the margin on which the currency channel operates. Anthropic's latest report, for February 2026, finds usage across countries becoming slightly more concentrated, with low-adoption countries falling further behind (Anthropic, 2026).

6.4 Field observation

A large South African retailer planned to move its POS estate to Google Cloud Platform as the foundation for store-level AI. Implementation and running costs were quoted in dollars and ongoing support from a European partner in sterling. Together they exceeded the IT budget approved in rand, and the programme was held back.

The observation maps onto the model's budget-constrained regime (Proposition 1): the project was wanted, and the immediate constraint identified was the rand-denominated funding limit. The AI roadmap was delayed with the migration, whatever happened to

token prices (Propositions 2 and 3). The timing matters for interpretation. The deferral came in 2026, when the rand was stronger than in 2025, so it reflects the level of hard-currency costs relative to rand budgets built up over a decade — a dollar bill still cost 27% more in rand in September 2026 than in 2015 — not a recent depreciation. One case cannot separate the currency channel from other reasons for deferral. It is offered as an illustration consistent with the model, not as independent validation of it.

6.5 What the evidence does and does not show

Claim	Evidence	Status
Enterprise AI stack is priced in hard currency; consumer AI is starting to be priced locally	Vendor documentation; OpenAI India pricing	Documented
Rand cost of hard-currency AI bills rose substantially, in short windows	FRED exchange rates, 2015–2026	Descriptive
The shock regularly exceeds typical budget contingencies, especially over multi-year horizons	Monthly overshoot analysis with sensitivity	Descriptive
Budget constraints, not profitability, delay adoption (Proposition 1)	One firm	Illustrative
Token-price cuts are offset by modest depreciation (Proposition 2)	Model, with illustrative cost shares	Theoretical
Usage is more than unit-elastic to the rate once budgets bind (Proposition 3)	Model	Not yet tested
Enterprise use lags consumer use more where currencies weaken (Proposition 4)	Model	Not yet tested

7. Limitations and research agenda

The evidence measures the size of the currency shock to budgets, not its effect on adoption. Three limitations matter most. The overshoot analysis assumes no hedging, no local-currency contracts and no reseller absorption, so it is an upper bound for large firms that hedge and closer to the truth for mid-sized firms that cannot. The field observation is a single firm, and depreciations often coincide with weak domestic demand, which could

delay investment on its own. And the model collapses the two investment stages into one decision and abstracts from the option value of waiting, which would strengthen its predictions but changes how thresholds should be estimated.

Testable predictions.

1. Enterprise cloud and AI spending in local currency should show kinks at budget-cycle boundaries after large depreciations, and inference volumes should be more than unit-elastic to the exchange rate when budgets bind (Proposition 3).
2. Across countries of similar income, larger depreciation should predict a lower ratio of enterprise API use to consumer use (Proposition 4). Anthropic already reports API and consumer traffic separately; country-level API data would allow a direct test.
3. Exporters and firms with hard-currency revenue should adopt with less sensitivity to the exchange rate than domestic-market firms in the same country (Remark 1).
4. Local pricing of enterprise cloud or inference should raise adoption more than an equal-sized dollar price cut, because it reduces the exchange-rate component of budget risk as well as lowering the level (Propositions 2 and 3). It would not remove all budget uncertainty: usage, contractual and other cost risks remain.

Data and identification. The cleanest designs compare firms within the same country and period that differ in the share of their stack priced in hard currency, or compare consumer and enterprise use over the same period, so that domestic demand shocks are held constant. Three data sources would make this possible: hyperscaler and reseller billing by currency and customer country, for event studies around depreciations; IT budget and procurement records from large emerging-market firms, which systems integrators observe, to date approvals, deferrals and re-approvals; and country-level API usage from AI vendors.

8. Conclusion

For an emerging-market firm, the price of enterprise AI is the vendor's hard-currency price multiplied by an exchange rate the firm does not control, applied to cloud, inference and support together, and paid from a budget fixed in local currency a year ahead. The model shows how this creates a budget threshold separate from profitability, why token-price cuts reach such firms only partly, and why usage can become highly sensitive to the exchange rate once budgets bind. For South Africa, the shock regularly exceeds plausible contingencies, especially over the multi-year horizons that cloud migration requires.

For measurement, the implication is that consumer-weighted diffusion indices may understate the divergence in enterprise AI. Local pricing is more established in consumer AI subscriptions than in the enterprise cloud and model services we examine. If the two-

speed prediction holds, the gap that matters for productivity is wider than the indices show, and it moves with exchange rates as well as income.

References

- Agrawal, A., J. Gans and A. Goldfarb, eds. (2019). *The Economics of Artificial Intelligence: An Agenda*. Chicago: University of Chicago Press for NBER. Link
- Aguiar, M. (2005). "Investment, devaluation, and foreign currency exposure: The case of Mexico." *Journal of Development Economics*, 78(1): 95–113.
- Amazon Web Services (2026). "What currencies does AWS currently support?" AWS re:Post Knowledge Center. Accessed 1 November 2026. Link
- Anthropic (2026). "Anthropic Economic Index report: Learning curves." 24 March 2026. Link
- Appel, R., P. McCrory, A. Tamkin, M. McCain, T. Neylon and M. Stern (2025). "The Anthropic Economic Index report: Uneven geographic and enterprise AI adoption." arXiv:2511.15080. Link
- Bloom, N. (2009). "The impact of uncertainty shocks." *Econometrica*, 77(3): 623–685.
- Bloom, N., S. Bond and J. Van Reenen (2007). "Uncertainty and investment dynamics." *Review of Economic Studies*, 74(2): 391–415.
- Bloom, N., R. Sadun and J. Van Reenen (2012). "Americans do IT better: US multinationals and the productivity miracle." *American Economic Review*, 102(1): 167–201.
- Bresnahan, T., E. Brynjolfsson and L. Hitt (2002). "Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence." *Quarterly Journal of Economics*, 117(1): 339–376.
- Bresnahan, T. and M. Trajtenberg (1995). "General purpose technologies: 'Engines of growth'?" *Journal of Econometrics*, 65(1): 83–108.
- Brynjolfsson, E., D. Rock and C. Syverson (2021). "The productivity J-curve: How intangibles complement general purpose technologies." *American Economic Journal: Macroeconomics*, 13(1): 333–372.
- Burstein, A. and G. Gopinath (2014). "International prices and exchange rates." In G. Gopinath, E. Helpman and K. Rogoff, eds., *Handbook of International Economics*, Vol. 4: 391–451. Amsterdam: Elsevier.
- Comin, D. and B. Hobijn (2010). "An exploration of technology diffusion." *American Economic Review*, 100(5): 2031–2059.
- Dixit, A. and R. Pindyck (1994). *Investment under Uncertainty*. Princeton: Princeton University Press.
- Epoch AI (2025). "LLM inference price trends." 12 March 2025. Link

- Fazzari, S., R. G. Hubbard and B. Petersen (1988). "Financing constraints and corporate investment." *Brookings Papers on Economic Activity*, 1988(1): 141–195.
- Federal Reserve Bank of St. Louis (2026). FRED series AEXSFUS, AEXUSUK, AEXINUS, AEXBZUS, EXSFUS and EXUSUK, from Federal Reserve Board H.10 release. Accessed 1 November 2026. Link
- Google Cloud (2026). "Local currency billing overview." Accessed 1 November 2026. Link
- Gopinath, G., E. Boz, C. Casas, F. Díez, P.-O. Gourinchas and M. Plagborg-Møller (2020). "Dominant currency paradigm." *American Economic Review*, 110(3): 677–719. Link
- Hall, B. and B. Khan (2003). "Adoption of new technology." NBER Working Paper 9730.
- Halpern, L., M. Koren and A. Szeidl (2015). "Imported inputs and productivity." *American Economic Review*, 105(12): 3660–3703.
- McElheran, K., J. F. Li, E. Brynjolfsson, Z. Kroff, E. Dinlersoz, L. Foster and N. Zolas (2024). "AI adoption in America: Who, what, and where." *Journal of Economics & Management Strategy*, 33(2): 375–415.
- Microsoft (2025). "Azure plan price list." Partner Center documentation. Accessed 1 November 2026. Link
- Microsoft (2026a). *Global AI Diffusion Q1 2026: Trends and Insights*. May 2026. Link
- Microsoft (2026b). *Global AI Diffusion Q2 2026: Trends and Insights*. September 2026. Link
- TechCrunch (2025). "OpenAI launches a sub-$5 ChatGPT plan in India." 18 August 2025. Link

Appendix A. Reproducibility

Every figure in Sections 4.3 and 6 can be reproduced from the series and settings below. The supplementary workbook *Priced_in_Dollars_supplementary_data.xlsx* contains the monthly data and live formulas for each calculation; changing an assumption cell recalculates the results.

A.1 Data

Series	Description	Frequency	Period used	Retrieved
EXSFUS	Rand per US dollar, average of daily H.10 rates	Monthly	Jan 2015 – Sep 2026	1 Nov 2026
EXUSUK	US dollars per pound, average of daily H.10 rates	Monthly	Jan 2015 – Sep 2026	1 Nov 2026

Series	Description	Frequency	Period used	Retrieved
AEXSFUS	Rand per US dollar	Annual	2015–2025	1 Nov 2026
AEXUSUK	US dollars per pound	Annual	2015–2025	1 Nov 2026
AEXINUS	Rupees per US dollar	Annual	2015–2025	1 Nov 2026
AEXBZUS	Reais per US dollar	Annual	2015–2025	1 Nov 2026

All series are from FRED (Federal Reserve Bank of St. Louis), sourced from the Federal Reserve Board's H.10 release. Rand per pound is computed as rand per dollar × dollars per pound, month by month or year by year. Adoption figures are taken as published in Microsoft (2026b).

A.2 Calculation choices

- **Stack Cost Index:** 100 × e_t / e_b, with list prices in hard currency held constant. The chart in Section 6.1 uses annual averages with b = 2015; the base-period table in Section 6.1 uses averages of monthly data.
- **Planning rate:** the monthly average spot rate in planning month m.
- **Overshoot:** r(m,k) = w·e(m+k)/e(m) + (1 − w)·g(m+k)/g(m); average overshoot = mean of r(m,k) over k = 1…H, minus 1; worst-month overshoot = max of r(m,k), minus 1, computed for w = 1 only.
- **Windows:** every planning month with a full horizon of data — January 2015 to September 2025 for H = 12 (129 windows) and to September 2024 for H = 24 (117 windows). Windows overlap.
- **Contingency test:** a window counts when its overshoot exceeds β; shares are the count over the number of windows.
- **Hedge:** a proportional hedge of share h at the planning rate scales each overshoot by (1 − h); hedge costs are not deducted.
- **Proposition 2, elasticity-adjusted case:** x′ = 1 − (1 − x)^(1+η) with x = 0.5 and η = −0.29, giving x′ = 0.389; then d = θ_A·x′/(1 − θ_A·x′) in case (a) and θ_A·x′/(θ_D − θ_A·x′) in case (b), with θ_A = 0.2 and θ_D = 0.7.

A.3 Settings

Parameter	Values used	Baseline
Dollar share of hard-currency cost (w)	1.0, 0.8, 0.6	1.0
Horizon (H)	12, 24 months	12 months
Contingency (β)	5%, 10%, 15%, 20%	10%
Hedge share (h)	0, 0.5	0
Inference share of stack cost (θ_A)	0.2 (illustrative)	0.2
Token-volume elasticity (η)	0, −0.29	0

A.4 Supplementary workbook

Sheet	Contents
README	Sources, retrieval dates, definitions and the assumption cells (w, h, β)
Monthly	141 monthly observations of EXSFUS and EXUSUK, with rand per pound computed
Overshoot12m	Overshoot by planning month for the 12-month horizon: w = 1.0, 0.8, 0.6, worst month, and hedged
Overshoot24m	The same for the 24-month horizon
Summary	Medians, below-plan shares and shares above each β — reproduces the table in Section 6.2
Annual_SCI	Annual series and the Stack Cost Index plotted in Section 6.1
Prop2	Fixed-volume and elasticity-adjusted offsetting depreciation for Proposition 2, with editable x, η, θ_A and θ_D
Checks	Observation and window counts, rand-per-pound consistency, hedged versus unhedged rows, and yearly means of the monthly data against FRED's annual series (all within 0.2%)