When Does Agent Collaboration Pay in the Caribbean?

A 100,000-scenario study of workflow and economy-wide thresholds

Adrian Dunkley

17 February 2027

Abstract

This study estimates the minimum coordination exponent at which a team of AI agents can match incumbent workflow throughput and cut the cost per verified case by at least 10%. It then translates eligible cost savings into a Jamaica-weighted value-added equivalent. The experiment draws 100,000 joint scenarios for seven industries and evaluates teams of 2, 4, 8, 16 and 32 agents. In a fixed business-services example, eight agents reach cost parity while matching incumbent throughput at **0.7122** and clear the 10% savings gate at **0.7629**. Four agents require **0.815** and 32 fail the quality gate. Across reference scenarios that reach the economy-wide target within the search range, the median exponent for a **0.25% gross value added (GVA) equivalent** is **0.793**. The 5th and 95th percentiles are 0.632 and 0.943. At a common exponent of 1.00, the median level effect is **0.565% of GVA** under reference assumptions and **0.060%** under concurrent costs, reliability, adoption and domestic-conversion shocks. The national accounting weights are observed. Workflow performance and economy-wide conversion remain scenario assumptions.

Keywords: agent collaboration, coordination, workflow productivity, scenario analysis, Jamaica, small economies.

1. Introduction

A service provider considering more AI agents needs to know what happens to a complete case. Throughput at intake matters little when the case still needs correction at review. The operating question is whether the whole team can maintain quality, finish cases at least as fast as the available alternatives, and charge less per verified resolution.

Imas, Koh and Legg separate competence from coordination in *Organizing Intelligence* [1]. Competence sets the productivity of a team at a given scale. Coordination changes the return to enlarging it. Their theoretical model links sustained increasing returns to firm scale and market concentration under specified assumptions. It builds on work about the cost of organising inside firms and the value of routing hard problems to specialists [2], [3]. The present study asks a finite-scale question and adds explicit cost, backlog and review conditions. It also asks how much of a workflow gain must spread before it matters to a small economy. An exponent of 1 means doubling the agents doubles output before added coordination losses. At an exponent of 0.763 the same doubling gives about 1.70 times the output before those losses.

This paper follows the Caribbean AI Association's Jamaica study of adoption, employment and export demand [16]. That study follows labour-market scenarios through 2030. The new estimand is a coordination threshold at a fixed level of deployment. The present model contains no time-to-adoption function or employment projection. Jamaica provides measured industry weights. Other Caribbean economies can reuse the tests with their own sector weights, operating costs and constraints.

2. Evidence and research design

Task studies motivate the choice of comparators. Brynjolfsson, Li and Raymond report a 15% average increase in issues resolved per hour in a customer-support deployment [4]. Noy and Zhang find faster completion and improved output quality on selected professional writing tasks [5]. In a product-development experiment, people working with AI reached the performance of two-person teams working without it [6]. Neither the task nor the organisational setting in those studies provides a Caribbean estimate of agent-to-agent coordination.

Benchmark evidence motivates the reliability and workload assumptions. Kim and colleagues find that the effect of adding agents varies across sequential and parallel work and with how agents communicate [7]. Cemri and colleagues document coordination and verification failures in multi-agent systems [8]. Kwa and colleagues use software-task duration to characterise changing reliability [9]. These studies motivate measuring verified outcomes across team sizes. Their benchmark scores do not identify the parameters used here.

Caribbean evidence establishes where deployment constraints may matter. The Economic Commission for Latin America and the Caribbean documents variation in national AI readiness [13]. The OECD and Inter-American Development Bank describe differences in digital infrastructure [14]. World Bank and International Labour Organization work assesses occupational exposure and the digital divide across Latin America and the Caribbean [15]. I use those studies as context, with zero imported wage, adoption or task-exposure coefficients. The experience of Jamaica's global digital services sector motivates the business workflow example [17]. The scenario places the worked example in the broad real estate and business activities category. The national accounts table gives no separate outsourcing weight.

2.1 Economic inputs

Jamaica's Statistical Institute publishes quarterly gross value added by industry at constant 2015 prices [12]. I sum the four seasonally unadjusted 2025 quarters for each category and divide by the sum of the four national totals at basic prices. The seven modelled categories account for **65.69%** of the national value-added denominator. The remaining 34.31% contributes zero direct gain in the calculation. No weights are renormalised after exclusion. This is a 2025 weighting base, with no forecast date attached to the results.

Industry category	2025 GVA (J$ million, 2015 prices)	Share of total
Manufacturing	162,495	9.46%
Wholesale and retail trade, including repairs and installation	288,414	16.78%
Accommodation and food service activities	95,127	5.54%
Information and communication	71,442	4.16%
Financial and insurance activities	186,536	10.85%
Real estate and business activities	214,625	12.49%
Public administration and defence	110,239	6.41%
National total at basic prices	1,718,583	100.00%

2.2 Workflow, quality and alternatives

Set the incumbent human workflow to one verified case and one unit of fully loaded cost per hour. A verified case meets the agreed quality standard. In a field test it must also stay closed through the review window.

Each simulated workflow receives a defined set of assumptions for individual AI assistance, standalone AI capacity and the degree to which cases can be processed independently.

Let q_1 be the verified hourly output of one agent in units of the human workflow, a_1 the single-agent acceptance probability, and p the effective share of capacity that additional independent agents can use without sharing case state. An independent pool of m agents produces $q_1[1 + p(m - 1)]$. Its cost combines fixed supervision, per-agent cost and a routing term proportional to $\log_2 m$. The two human workflows meet the acceptance floor by assumption. An independent configuration enters the cost comparison if a_1 meets the same acceptance floor and its throughput reaches one case per human-workflow hour.

For collaboration, γ is a scale exponent before added coordination losses. A reliability decay parameter λ captures rising verification failure as the team grows. Predicted accepted output Q_m, acceptance A_m and fully loaded hourly cost C_m are

$$Q_m(\gamma) = q_1 m^{\gamma} e^{-\lambda(m-1)}, \qquad A_m = a_1 e^{-\lambda(m-1)}, \qquad C_m = f + cm + v \log_2 m.$$

The baseline q_1 already embeds single-agent acceptance. Applying the exponential term incurs only the extra loss as more agents join. The cost terms f, c and v cover fixed integration and human oversight, marginal agent execution, and extra coordination and review, respectively. Their normalisation is relative to the human workflow cost. The acceptance floor varies by sector. It screens cases in the model and does not represent compliance or a serious-error guarantee.

For team size m, L_m is the fastest verified throughput of the human workflow, the human workflow with individual AI assistance, and an acceptable independent pool of the same size. Let B be the lowest unit cost across the two human options and all capacity- and quality-eligible independent pools at sizes 1, 2, 4, 8, 16 and 32. The cost comparison thus allows an incumbent to choose its most efficient tested scale. The speed comparison keeps the team size matched to the pooled agent comparison.

2.3 Exact workflow threshold

For collaboration to qualify, the team must meet three conditions. Its acceptance must clear the sector floor $a_{\min}$. It must match or exceed L_m. Its unit cost must also fall by at least $g = 0.10$ against B. If the quality gate passes, the exact minimum exponent is

$$\gamma_m^* = \frac{\log\left(\frac{\max\{L_m,\, C_m/[B(1-g)]\}}{q_1 e^{-\lambda(m-1)}}\right)}{\log m}.$$

If $A_m < a_{\min}$, no exponent can qualify that size under fixed reliability. The sector threshold is the minimum of the qualifying size thresholds. I also compute a break-even threshold with $g = 0$ and retain the same throughput and quality gates.

DeepMind's $\gamma = 1$ boundary concerns asymptotic returns in its own model [1]. Here $\gamma = 1$ is linear output scaling before losses. Over a doubling from m to $2m$, verified output more than doubles only if $\gamma > 1 + \lambda m / \log 2$. With positive λ, effective returns can fall as more agents join. A sublinear team can still qualify through lower cost and sufficient absolute throughput. The workflow crossover, superlinear verified output and aggregate materiality are three different measurements.

2.4 Economy-wide calculation

For each sector, the simulation chooses the eligible collaborative size with the largest cost saving $s_i = 1 - C_m/(BQ_m)$. A sector with no qualifying size contributes zero. Let w_i be the observed industry GVA weight and e_i assumed addressable spending at the benchmark unit cost and the same workload volume,

expressed as a share of industry GVA. Let d_i be the share of that cost base that adopts the workflow and r_i the share of savings converted into domestic value-added equivalent. The last factor accounts in reduced form for incomplete redeployment, imported services outside measured workflow costs, and demand constraints. It is bounded between zero and one. No cost item should be deducted both in the workflow cost and in the conversion factor.

$$G(\gamma) = \sum_i w_i e_i d_i r_i s_i(\gamma).$$

This fixed-weight cost-to-value calculation follows the discipline of distinguishing task savings from economy-wide gains [10]. It contains no input-output network, price clearing, wage response, trade elasticities or public-spending multipliers. Money saved is a candidate resource release. Only the assumed converted share enters G. Labour-market findings from other countries provide a further reason to avoid mapping workflow gains one-for-one into wages or hours [11].

2.5 Monte Carlo protocol and checks

The reference experiment draws 25,000 joint scenarios from each of four independently seeded streams. Each draw samples one value per parameter from fixed triangular ranges for seven industries. Shared multipliers represent common movements in agent costs, adoption and reliability, while the underlying sector parameters vary independently conditional on these multipliers. The source package records every lower bound, mode and upper bound. There is no likelihood fit, survey weighting or time-series calibration. Each reported hit rate is the share of these draws that passes a specified rule.

The three materiality targets are 0.10%, 0.25% and 0.50% of baseline GVA. A 0.25% level effect corresponds to J$2.50 of modelled gain per J$1,000 of baseline domestic value added. For each target, I locate the earliest common exponent where G reaches the target on a bounded search ending at 1.45. The value calculation is monotone under fixed inputs and may jump when a workflow first qualifies. Bisection finds the lower boundary of attainment to within 1.1×10^{-7} in the exponent. Values outside the upper bound remain censored. Percentiles of minimum exponents describe draws that cross within that bound. The paper reports the uncensored share alongside the percentiles.

The stressed calculations keep the identical scenario draws and independently halve adoption, double λ, raise fixed integration and coordination cost by 50%, or halve domestic conversion. A combined case applies all four. The solver checks the final bracket, monotonicity over the grid and the direction of paired stress effects. The seeds are 261105417 through 261105420. For the 0.25% target at $\gamma = 1$, attainment varied from 98.47% to 98.64% across the four streams. Agreement across seeds addresses Monte Carlo noise under the same assumptions.

3. Results

3.1 The eight-agent crossover

The example uses the most likely assumed values for business activities, with a standalone agent at 0.71 of the human workflow's verified output and $\lambda = 0.004$. At eight agents, the team costs 0.858 per hour and the cheapest eligible incumbent is the 32-agent independent pool, at 0.2826 per verified case. The throughput to match is 2.499 cases per normalised hour, set by the eight-agent independent pool. Cost parity at matched throughput requires a coordination exponent of **0.7122**. An exponent above that point gives a strictly lower unit cost under the same assumptions. The cost rule is binding. The coordinated team must reach 3.373 verified cases per hour to save 10%, which occurs at $\gamma_m^* = 0.7629$. Its acceptance is 95.59%, above the 92% floor.

Agents	Cost parity with matched speed	Exponent for 10% savings
2	0.950	1.102
4	0.739	0.815
8	**0.7122**	**0.7629**
16	0.733	0.771
32	Ineligible	Ineligible

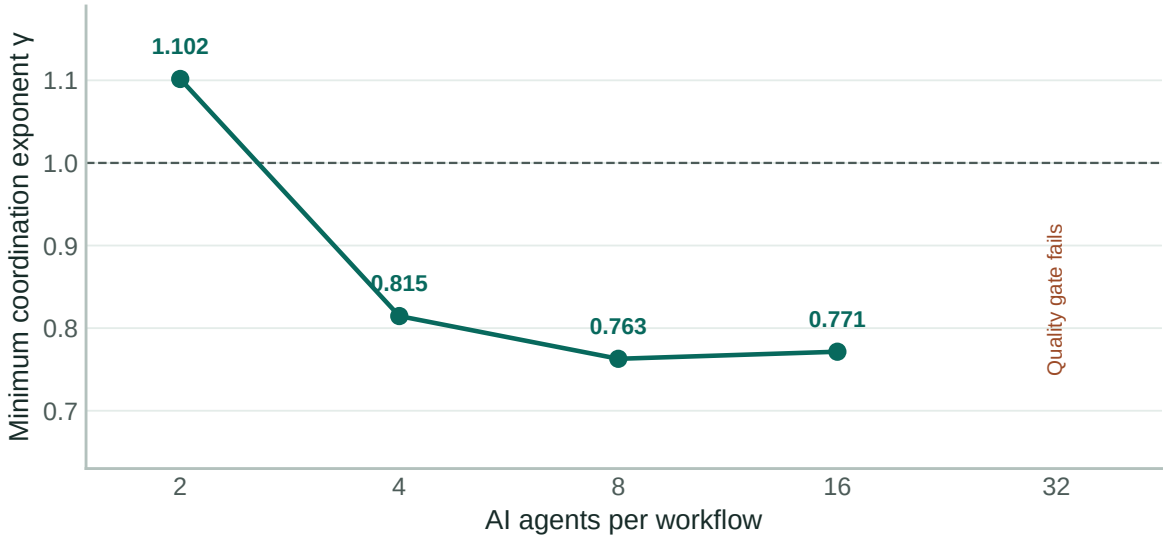


Figure 1: Minimum qualifying exponent for each team size in the modal business-activities example. The 32-agent team fails the quality gate. The horizontal line at 1 marks linear scaling before coordination losses.

The eight-agent result is an operating point for one transparent set of inputs. At $\gamma = 1$, four or eight agents maximise the eligible savings in the majority of draws in each sector. The reference results favour four agents most often in manufacturing, finance and public administration. Eight are most often selected in trade, accommodation, information and communication, and real estate and business activities. The selected size is based on savings at that exponent, which is a different objective from finding the lowest possible crossover.

3.2 Workflow thresholds by sector

Industry	Median break-even γ	Median γ with 10% saving	Conditional 5th–95th percentile	Reaches 10% by $\gamma = 1$
Manufacturing	0.592	0.629	0.425–0.945	94.87%
Wholesale and retail trade	0.646	0.707	0.453–0.936	97.91%
Accommodation and food services	0.603	0.660	0.437–0.892	98.82%
Information and communication	0.768	0.817	0.574–1.018	93.43%
Finance and insurance	0.668	0.715	0.453–1.236	80.80%
Real estate and business activities	0.751	0.803	0.540–1.030	92.84%

Industry	Median break-even γ	Median γ with 10% saving	Conditional 5th–95th percentile	Reaches 10% by $\gamma = 1$
Public administration and defence	0.657	0.689	0.439–1.270	78.13%

The distribution for public administration and finance includes long upper tails. Their tighter acceptance floors screen out high-capacity designs more often under the protocol. Information and communication has the highest median threshold. That sector has stronger independent processing in the scenario inputs, so it sets a tougher comparator. The table illustrates how a strong baseline can raise the hurdle even when the individual agent is productive. It does not rank actual sector readiness.

3.3 Economy-wide thresholds

Target as a share of baseline GVA	Median critical γ	Conditional 5th–95th percentile	Share of all draws by $\gamma = 1.00$	Share of all draws by $\gamma = 1.45$
0.10%	0.662	0.491–0.808	99.99%	100.00%
0.25%	**0.793**	0.632–0.943	98.57%	99.999%
0.50%	0.954	0.790–1.204	64.47%	99.20%

At $\gamma = 1$, the reference median is **0.565% of baseline GVA**, with a 5th–95th percentile range from 0.313% to 0.916%. At 1.15 the reference median is 0.744%. The percentages are changes relative to a fixed level of GVA, so they should not be added to an annual growth rate.

3.4 Sensitivity to costs, reliability and uptake

At $\gamma = 1$, doubling added reliability losses reduces the median to **0.370%** of GVA. Raising integration and routing costs by 50% reduces it to **0.442%**. Halving adoption or domestic conversion reduces it to **0.282%**. Those two shocks have the same effect by construction. They multiply the same cost-saving term and the reference adoption range never binds at one. The combined effect drops to **0.060%** of GVA. Only 8.65% of the combined-stress draws clear the 0.25% target even at the upper bound of 1.45. A coordination improvement can clear a workflow's threshold and leave the aggregate gain small if deployment stalls or released capacity goes unused.

4. Implications for a Caribbean field test

A Jamaican outsourced service provider could start with a queue where a case moves from intake to billing or document review before it closes. Existing work supplies the baseline. Additional test arms use individual AI assistance, the same model in an independent pool, and the same model in a coordinated team. Set the same case acceptance standard, keep serious-error limits independent of the score, randomise matched batches when feasible, and blind the case audit. Sample at 1, 2, 4 and 8 agents before paying for a larger team. Measure loaded costs per verified case, 90th-percentile closure time, reopens, oversight time and escalation failures. Report the same-total-model-budget experiment separately from a test that holds spending per agent fixed.

The field team can estimate the effective elasticity directly from the slope of log verified throughput against log team size. To infer a pre-loss γ for this model, it must also measure acceptance at those sizes. With just two throughput levels, a throughput slope combines raw scaling and reliability and cannot separate

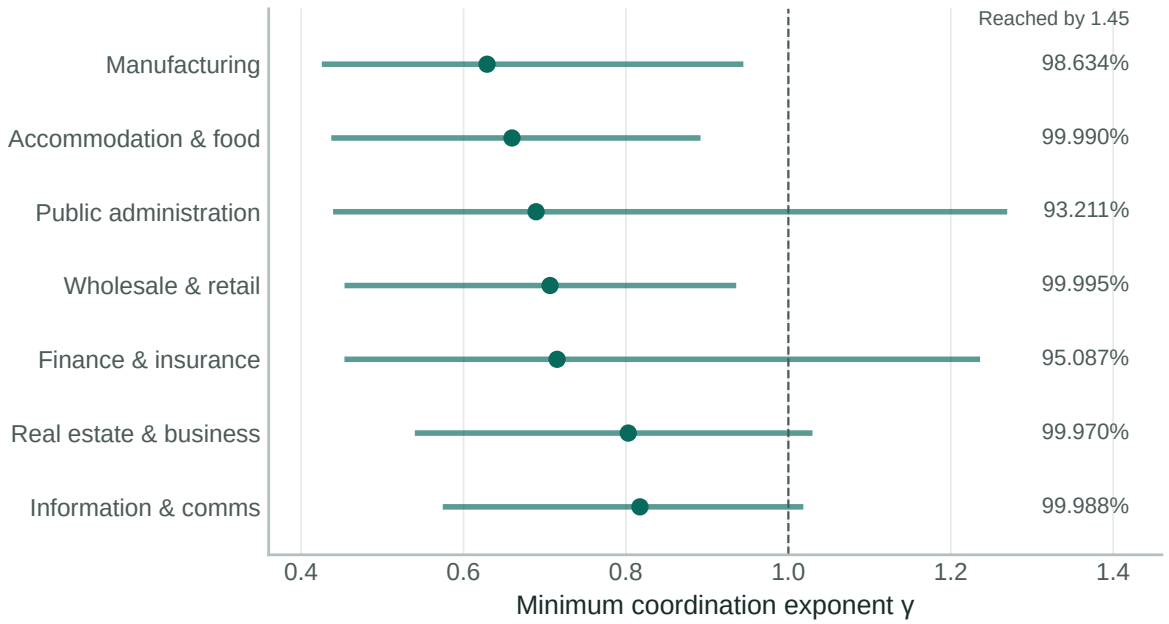


Figure 2: Median and 5th–95th percentile of qualifying workflow thresholds. The right column shows the share of all scenarios that reach the 10% saving gate by 1.45.

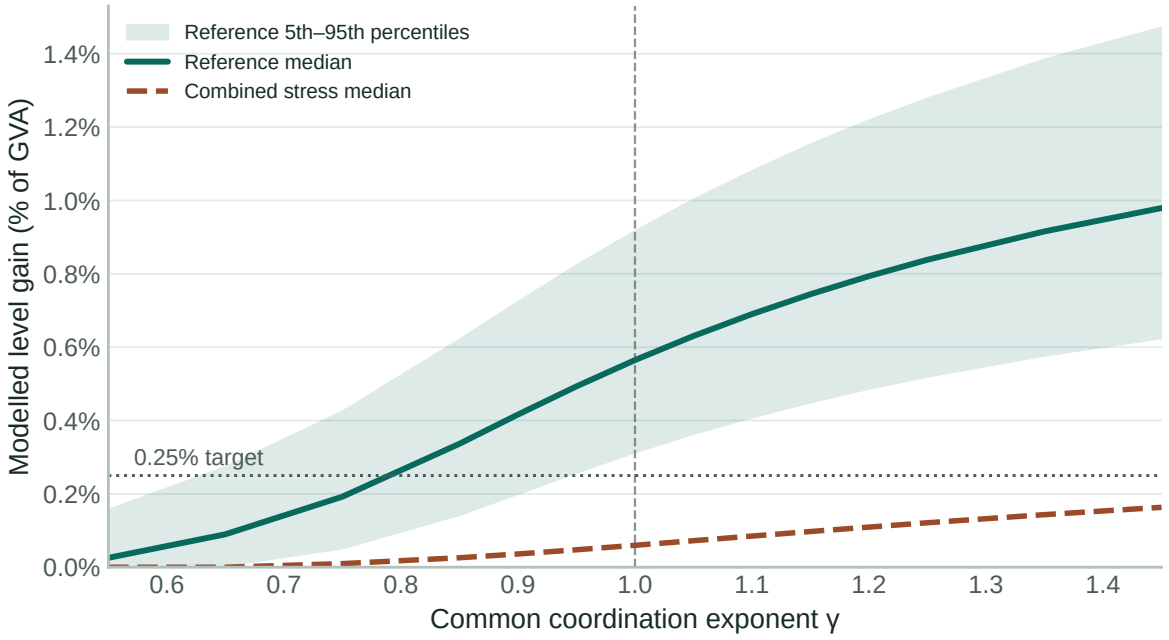


Figure 3: Reference median and 5th–95th percentile for the modelled value-added equivalent, with the combined stress median shown separately. The horizontal line marks the 0.25% target.

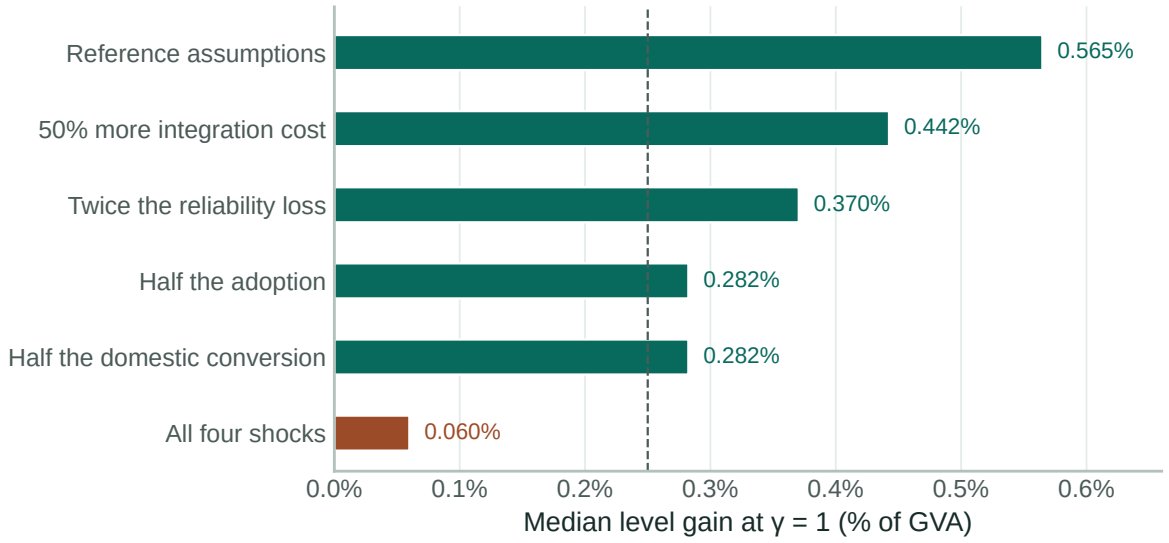


Figure 4: Median level effect at gamma 1.00 under each matched stress. The median falls from 0.565% of GVA under reference assumptions to 0.060% with all four shocks.

them. A model update or a change in the mix of cases should trigger renewed measurement. Report results by workload complexity and language where the sample supports the split.

The same comparison can be used for permit-intake completeness or a simulated regional disaster-information handoff. In each case, replace the accounting and workflow inputs, test queue performance under a surge, and assign the actual authority to approve the work. Audit gains received by providers, government offices and workers separately. For a regional economic estimate, use country-specific national accounts and count cross-border sales once. Otherwise a cost saving in one jurisdiction can also appear as revenue in another.

5. Limitations and interpretation

The figures are conditional scenario outputs. Only the Jamaican industry weights come from observed data. The cost, loss, single-agent capacity, overlap, adoption and domestic conversion distributions are assumptions. A 100,000-draw experiment estimates their numerical implications with small sampling error. It cannot tell whether the scenario inputs describe operations in Jamaica. The selected industries omit 34.31% of Jamaican GVA from direct effects. The 2025 real GVA shares are held fixed and do not adjust to new prices or sector growth.

The reliability model imposes an exponential loss with team size and a hard quality floor. Workflows with strongly correlated failures, architectural improvements that lower errors as agents join, and more complex review queues may behave differently. A common exponent is a convenient way to locate an economy-wide boundary. Firms, workflows and jurisdictions can have different exponents in practice. The smallest tested threshold is conditional on the five permitted team sizes. It cannot determine whether three, six or twelve agents would perform better. The aggregate conversion factor is uncertain and public-sector savings may be realised as service quality rather than additional measured output.

The model compares operating configurations under static backlog and their assumed costs. It does not estimate firm entry, market shares, profits, employment displacement, welfare or a date for anyone to expect the result. A longitudinal study would need verified deployment and worker records plus evidence on how buyers change contracts and demand as costs fall. These limitations keep the simulated crossover distinct from the concentration mechanism in DeepMind's paper [1].

6. Conclusion

Eight agents clear the 10% unit-cost gate at $\gamma = 0.763$ in the worked business case. The analogous national result depends on whether that gain travels beyond one queue. The reference experiment places the median 0.25% GVA-equivalent crossing at $\gamma = 0.793$. The combined stress leaves most runs short of that target even at 1.45. A first field study should measure collaboration against a same-model independent pool and exclude reopened cases from verified throughput and count their rework in cost. That supplies the inputs which currently determine the answer by assumption.

Reproducibility and disclosures

This working paper reports an executed simulation. It uses published aggregate national accounts, without identifiable customer or employee records. It has no external peer-review decision or arXiv identifier. The simulation protocol, four random-number seeds, input ranges, code, derived tables, chart scripts and checksums are provided in the accompanying packages. AI tools were used in preparation of analytical code, graphics and the manuscript draft. No claim of institutional endorsement is made.

References

1. Imas, A., Koh, A. and Legg, S. (2026). *Organizing Intelligence*. arXiv:2611.05417.
2. Coase, R. H. (1937). The nature of the firm. *Economica*, 4, 386–405.
3. Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108, 874–904.
4. Brynjolfsson, E., Li, D. and Raymond, L. R. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140, 889–942.
5. Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381, 187–192.
6. Dell'Acqua, F. et al. (2026). The cybernetic teammate: A field experiment on generative AI and teamwork. *Organization Science*, 37, 1217–1242.
7. Kim, Y. et al. (2025). *Towards a Science of Scaling Agent Systems*. arXiv:2512.08296.
8. Cemri, M. et al. (2025). Why do multi-agent LLM systems fail? *NeurIPS Datasets and Benchmarks*.
9. Kwa, T. et al. (2025). *Measuring AI Ability to Complete Long Software Tasks*. arXiv:2503.14499.
10. Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, 40, 13–58.
11. Humlum, A. and Vestergaard, E. (2025; revised 2026). *Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI*. NBER Working Paper 33777.
12. Statistical Institute of Jamaica. *Quarterly GDP Data Explorer*. Seasonally unadjusted industry GVA at constant 2015 prices, calendar quarters of 2025.
13. Alexander, D., Døhl Diouf, L. and Wooding, C. (2025). *Artificial intelligence readiness in the Caribbean: An exploratory review*. ECLAC.
14. OECD and IDB (2026). *Caribbean Development Dynamics 2026*. OECD Publishing.
15. Gmyrek, P., Winkler, H. and Garganta, S. (2024). *Buffer or Bottleneck? Employment Exposure to Generative AI and the Digital Divide in Latin America*. World Bank Policy Research Working Paper 10863.
16. Caribbean AI Association (2026). *Jamaica: jobs, AI and export demand*. Research report.
17. Jamaica Promotions Corporation. *Global Digital Services*. Sector profile.