FROM PARAMETER COUNT TO HUMAN-CENTERED LOCAL-FIRST AI

Hardware Reach, Environmental Externalities, and a Five-Year Human-in-the-Loop Strategy for Middle-Market Organizations

PUBLIC RESEARCH EDITION

Professor Timothy E. Bates
Battle-Scarred CTO Research Series
Pyrinas.co
Version 1.0 | August 2026

Research proposition

For a substantial class of early enterprise workloads, task-aware local inference can let middle-market organizations use existing or incrementally upgraded computing assets during a multi-year AI apprenticeship period. The goal is not to eliminate cloud or data-center AI. It is to right-size routine inference, keep humans in control, and escalate to heavyweight systems only when workload evidence justifies it.

What this edition intentionally excludes

Proprietary prompts, evaluator code, exact routing logic, production policy rules, customer data, internal endpoints, deployment secrets, and TAi implementation mechanisms are not published here.

Executive Summary

The enterprise AI market often treats model size as a proxy for capability. This public research edition tests a different proposition: organizations buy useful business outcomes, not parameter counts. The underlying benchmark evaluated ten enterprise-oriented tasks using local and network-accessible language models ranging from approximately 1.2B to 108.6B nominal parameters, and measured quality, throughput, task latency, accelerator power, energy per task, and cold-versus-warm execution behavior.

Across the tested workload suite, the lowest-energy model that cleared the provisional quality gate was never a 20B-, 30B-, or 100B-class model. The task-optimal routes used compact or specialist models at 8.3B parameters or below. This result does not establish that small models are universally superior. It supports a systems conclusion: model size should be an outcome of workload requirements, not the starting point of enterprise architecture.

Core operating principle

Use the minimum computational system that reliably produces the required business outcome. Reuse. Benchmark. Specialize. Escalate.

Dimension	Public result
Model scale tested	~1.2B to 108.6B nominal parameters
Enterprise task categories	10
Warm local model-task runs	120
Primary measured accelerator	NVIDIA GeForce RTX 3090 Ti, 24 GiB-class VRAM
Provisional quality gate	80 / 100
Task-optimal model ceiling	8.3B or smaller across all 10 tested tasks
Mean measured routed energy	0.1086 Wh/task
Warm-residency energy reduction	17.8% to 88.2% across evaluated models

1. What This Public Edition Shares - and What It Protects

A credible public research paper should reveal enough evidence to be evaluated without publishing the operational machinery that creates competitive advantage or security risk.

Published	Protected

Research questions, aggregate methods, benchmark scope, quality gate, representative hardware facts, aggregate results, limitations, scenario assumptions, environmental caveats, and citations.	Exact prompts, evaluator implementation, scoring weights beyond the public gate, orchestration/routing code, TAi policy logic, customer-specific datasets, private model aliases, internal network details, production credentials, and deployment secrets.

2. Research Questions

- Does increasing total parameter count consistently improve enterprise-task quality?
- Does total parameter count reliably predict local inference throughput?
- Can sub-10B models satisfy a provisional enterprise-quality threshold on tasks commonly delegated to much larger general models?
- How materially does warm model residency change latency and accelerator energy relative to cold execution?
- Can task-aware routing select lower-energy models without violating a defined quality threshold?

Primary hypothesis: enterprise AI efficiency is better predicted by task-model fit, execution architecture, and residency policy than by total parameter count alone.

3. Public Methodology

The study used a three-phase workflow. The public edition reports the experimental structure while withholding prompt text, evaluator code, exact routing implementation, and production orchestration rules.

Phase	Purpose	Public output
1	Inventory and throughput baseline	Model/task performance matrix and host capability summary
2	Deterministic quality screening and local energy pass	Quality-adjusted energy measurements
3	Warm residency, anomaly review, and task-aware selection	Aggregate routing findings and cold-versus-warm comparison

Each recommendation required the model to clear the provisional 80/100 quality gate before energy or latency could be used to prefer it. Missing telemetry was not treated as zero. Remote power was not inferred. Automated rubric results are screening metrics and not substitutes for expert adjudication.

4. Enterprise Workload Classes

The benchmark intentionally used diverse enterprise work rather than a single academic benchmark. Public task descriptions are kept at the category level:

- Cybersecurity incident analysis and evidence-bounded reasoning
- Secure-code review and control identification
- Contract and policy risk analysis
- Retrieval-grounded policy reasoning
- Hallucination resistance
- Tool-selection and routing decisions
- Architecture and least-privilege control design
- Executive tradeoff reasoning
- Enterprise summarization with uncertainty preservation
- Model-escalation decisions

5. Findings: Parameter Count Was Not the Decision Variable

The benchmark found no monotonic relationship between nominal model size and useful enterprise performance. Compact and specialist models repeatedly crossed the quality threshold on defined workloads, and the lowest-energy passing route for every tested task stayed at or below 8.3B parameters.

Maximum model tier allowed	Cumulative task coverage
1.2B	3 of 10 tasks
7.6B	6 of 10 tasks
8.3B	10 of 10 tasks
20B+	No additional task was required to obtain the lowest-energy passing route

This is an observed property of this ten-task benchmark, not a universal statement about enterprise work. Heavyweight models remain appropriate for frontier research, complex multimodal reasoning, difficult long-context work, synthetic-data generation, teacher models, and workloads that fail smaller tiers.

6. Warm Residency Is an Architecture Variable

Keeping a model resident after load materially changed measured latency and accelerator energy. Across evaluated models, warm-residency energy reduction ranged from approximately 17.8% to 88.2% relative to the earlier cold execution pass. This means energy models based only on parameter count or device TDP can miss a major operational effect.

Do not show me watts. Show me watts doing what?

Energy should be interpreted per useful task, and preferably per correct task. A lower-power device can consume more energy per completed outcome if it takes substantially longer or requires repeated retries.

7. Hardware Reach: Use the Computing Capital Already Manufactured

The public hardware-reach analysis separates memory feasibility from interactive performance. Direct performance measurements were made on the RTX 3090 Ti host. Other hardware classes are reach estimates until they are executed under controlled tests.

Representative model class	Observed quantized weight file	Conservative modeled practical memory
1.2B Q4	~0.68 GiB	~2.85 GiB
7.6B Q4 specialist	~4.36 GiB	~7.45 GiB
8.3B Liquid MoE Q4	~4.70 GiB	~7.88 GiB

Under that conservative one-model-at-a-time memory model, a 16 GB legacy PC has enough memory capacity to hold the largest task-optimal model from this benchmark. That does not prove acceptable interactive latency on an old CPU. It means the machine should not be discarded solely because the model will not fit.

Hardware tier	Public interpretation
Legacy 16-32 GB CPU-only PCs	Potential batch/background or low-concurrency pilot candidates; interactive speed remains unmeasured.
Older 6-12 GB discrete GPUs	Strong memory-feasibility candidates for the tested compact winners; runtime compatibility and throughput must be measured.
16-24 GB GPU workstations	Practical local inference tier; 24 GB RTX 3090 Ti was directly measured in this study.
40+ TOPS NPU AI PCs	Useful emerging local tier for supported runtimes and small models; requires device-specific validation.
High-memory unified-memory workstations	Can host much larger local models; still should be justified by workload rather than capacity alone.

8. Five-Year Local-First Scenario

A separate scenario model asks what happens if benchmark-like routine tasks are routed locally while complex exceptions escalate to centralized systems. The baseline is a planning model, not a prediction.

Assumption	Baseline
Mature task demand	20 potential AI tasks per employee per workday
Workdays	250 per year
Five-year adoption ramp	25%, 45%, 65%, 80%, 90%
Local eligibility	90% of benchmark-like routine tasks
Measured local routed mean	0.1085942 Wh/task
External comparison reference	0.34 Wh/query

Under those assumptions, centralized request volume falls by 90% for the modeled workload and total inference electricity falls by approximately 61% relative to sending the same task volume to the centralized comparison reference. The result is sensitive to the share of locally eligible work and to the energy intensity of the centralized workload.

Employees	5-year tasks	Local tasks	Centralized tasks	Modeled net kWh reduction
100	1,525,000	1,372,500	152,500	317.6
500	7,625,000	6,862,500	762,500	1,588.0
1,000	15,250,000	13,725,000	1,525,000	3,176.0
5,000	76,250,000	68,625,000	7,625,000	15,880.2

9. Environmental Interpretation: Marginal Pressure, Not Magical Savings

Local-first AI does not make data centers disappear. It can reduce marginal centralized inference demand for qualifying workloads, which may lower exposure to electricity demand, evaporative

cooling, cooling-tower blowdown, water-treatment chemistry, network traffic, hardware expansion, and premature equipment replacement.

- Cooling and water impacts are site-specific. Grid mix, climate, server efficiency, cooling architecture, utilization, and local HVAC can materially change the result.
- The paper uses a direct cooling-water proxy only to illustrate a calculation pathway. It is not a universal total-water-footprint claim.
- Hardware reuse can avoid some embodied impacts of premature replacement, but a complete lifecycle assessment is still required.
- The correct comparison is useful business work per unit of energy, not device watts in isolation.

10. The Five-Year AI Apprenticeship

The infrastructure argument is inseparable from the human argument. Organizations need time to learn what AI should do, where it should stop, and how human intent should be represented before broad autonomy becomes normal.

Stage	Human-machine relationship
Year 1 - Observe	AI watches structured work and helps document patterns; humans remain fully responsible.
Year 2 - Assist	AI drafts, retrieves, classifies, and summarizes; humans decide.
Year 3 - Recommend	AI proposes actions with evidence and uncertainty; humans approve or reject.
Year 4 - Act with approval	AI executes bounded actions after explicit authorization.
Year 5 - Bounded autonomy	Only demonstrated, reversible, monitored tasks receive constrained autonomy.

Human-centered principle

A model's ability to perform a task is not, by itself, justification for eliminating a job. Autonomy should be earned through measured quality, evidence, workforce participation, explicit approval gates, observability, and reversibility.

11. What an Enterprise Architecture Should Do

The practical conclusion is not “always run small models.” It is “build an intelligence orchestration system.” A mature architecture should classify the task, determine the minimum required capability, select an appropriate execution tier, evaluate the result, and escalate when justified.

Execution tier	Use when
Deterministic software	Rules, calculations, validation, lookups, and transformations do not require generative reasoning.
Compact/specialist local model	Routine enterprise tasks can be performed reliably with low latency, lower energy, and stronger data locality.
Larger local or departmental model	The task needs more capability or context but still benefits from local control.
Centralized/frontier model	Complex reasoning, advanced multimodal work, unusual context, burst capacity, or quality failure justifies escalation.
Human expert	Risk, uncertainty, policy, ethics, or business consequence exceeds the machine's authority.

An enterprise usually needs a better system before it needs a bigger model.

12. Limitations and Publication Boundaries

- The benchmark is not an MLPerf submission and should not be represented as one.
- Direct energy measurement was accelerator-focused on the primary local host rather than whole-facility AC power.
- Remote endpoint power was not inferred.
- The 80/100 quality threshold is an engineering gate, not an industry standard.
- Automated scoring can under- or over-score response formats; expert adjudication remains important.
- Hardware memory fit is necessary but insufficient; throughput, context length, runtime compatibility, thermals, and concurrency must be tested.
- The five-year scenario is assumption-driven and should be recalibrated with real organizational workload distributions.
- Environmental outcomes vary by location and infrastructure; site-specific data should precede regional claims.

13. Conclusion

This study began with a practical question: what business workload actually requires an 80B-, 400B-, or 800B-class model? The benchmark does not answer that universally. It demonstrates a disciplined way to ask the question.

For the ten enterprise-oriented tasks examined here, compact and specialist models repeatedly met the provisional quality threshold while using less measured accelerator energy than larger alternatives. Warm residency materially changed the economics. Task-aware selection did not require a 30B or heavyweight model as the lowest-energy passing route for any tested task.

The strategic implication is to let measured workload demand determine infrastructure. Reuse existing computing capital where it is capable. Benchmark real tasks. Specialize models and software. Escalate only when evidence warrants it. This approach can buy organizations time to learn workflows, governance, security boundaries, and human intent before making irreversible infrastructure and automation commitments.

Public research takeaway

The purpose of efficient local AI is not merely to save money. It may buy society time - time to learn what we want AI to do before we build enormous permanent infrastructure around assumptions we have not validated.

Public Use Note

This document is a public research edition intended for enterprise, policy, educational, investor, and architecture discussion. It is not a deployment specification, security configuration, procurement guarantee, environmental certification, or legal/compliance opinion. Organizations should validate workload quality, hardware performance, security controls, privacy obligations, and environmental assumptions in their own operating context.

Research by Professor Timothy E. Bates. Published by Pyrinas.co. TAi means Transparent AI: you decide what runs, where it runs, and why.

Selected References

[1] MLCommons. MLPerf Client Benchmark, v1.6 documentation, 2026.

[2] Amini, A., Banaszak, A., Benoit, H., et al. “LFM2 Technical Report.” arXiv:2511.23404, 2025.

[3] International Energy Agency. Energy and AI. Paris: IEA, 2025.

[4] Lei, N., Lu, J., Shehabi, A., & Masanet, E. R. “The water use of data center workloads: A review and assessment of key determinants.” Resources, Conservation and Recycling 219 (2025): 108310.

[5] U.S. Department of Energy, Federal Energy Management Program. “Cooling Water Efficiency Opportunities for Federal Data Centers.”

[6] U.S. Environmental Protection Agency. WaterSense at Work: Best Management Practices for Commercial and Institutional Facilities.

[7] Microsoft. “Measuring energy and water efficiency for Microsoft datacenters.” FY25 global WUE reported as 0.27 L/kWh.

[8] Oviedo, F., Kazhamiaka, F., Choukse, E., et al. “Energy Use of AI Inference: Efficiency Pathways and Test-Time Compute.” arXiv:2509.20241, 2025.

[9] Luccioni, A. S., Jernite, Y., & Strubell, E. “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” arXiv:2311.16863, 2023.

[10] NIST. AI Risk Management Framework 1.0, Appendix C: AI Risk Management and Human-AI Interaction.

[11] U.S. Environmental Protection Agency. “Electronics Donation and Recycling.” Updated 2025.