THE TK FOUNDRY PAPERS

SECOND PAPER · NO. 02

The State of Agent *Economics in Canada*

A public framework for task-level cost and residual overhead in agentic AI

STEPHEN THOMPSON

Founder · Thompson Kinkade Foundry

Toronto · 2026

This is the second of The TK Foundry Papers. It proposes a shared, testable way to measure what agentic AI actually costs — task by task, and after the best available mitigations — and invites Canadian operators to help build the evidence the field still lacks.

SECTION 01

A DISCIPLINE TAKING SHAPE

Why this report exists

> *The cost of agentic work has to be connected to the work it actually completes.*

Canada's AI for All strategy places sovereign compute and wider adoption at the centre of national policy. It continues more than $2 billion in existing AI-compute investments and proposes further access for Canadian firms.[1] At the same time, more AI work is performed through inference and multi-step agents that plan, call tools, retry, and branch toward an outcome. The cost of those operations must be connected to the work they actually complete.

Agent economics examines the full cost and value of a completed agentic task, where avoidable cost arises, and what share can be recovered without degrading quality or service.[14] FinOps, token economics, billing-data, and telemetry initiatives already supply important pieces of this discipline. This report proposes a complementary task-level taxonomy and a testable method for residual overhead. A comparable published Canadian, cross-estate baseline under this definition remains a research objective.[5,6,7,8]

Thompson Kinkade Foundry (TKF), a federally incorporated Canadian company based in Toronto, describes the measurement taxonomy, waste vocabulary, and residual-overhead method it uses on its own AI estate. The framework is offered for examination by Canadian operators, program officers, and researchers. The method is the subject of a United States provisional patent application; the final publication and reuse terms are set out in the notice at the end of this report.

THE STANDARD THIS REPORT HOLDS ITSELF TO

Public facts are cited; the methodology is disclosed as a proposal; TKF's own observation is identified as internal and qualitative. Evidence classes distinguish measured, modelled, and indicative results. No synthetic dollar figure or unverified saving is presented. Where a Canadian benchmark has not been established under this method, the report says so.

SECTION 02

WHY NOW

The economic ground in 2026

> *When the meter never stops, the test that matters is useful work per dollar.*

THE INFERENCE INFLECTION

At GTC 2026, NVIDIA described an "inference inflection" and said it had line of sight to roughly US$1 trillion in AI-infrastructure demand through 2027.[4] That is NVIDIA's own outlook, not a measurement of all compute demand or the agentic share of it. The direction nevertheless makes recurring inference efficiency commercially important; the economic test is useful work per dollar, subject to quality and service requirements.

AGENTS MULTIPLY THE METER

An agent may call models and tools repeatedly while it plans, retries, and checks a result. The additional calls can raise total spend even when unit token prices fall, but price and workload effects must be measured separately. For operating decisions, cost per successfully completed task should be reported together with completion rate, quality, latency, and the treatment of failed, abandoned, and human-escalated tasks.[5,6]

CANADA'S POSITION

AI for All targets 60% business AI adoption by 2034. The latest observed Statistics Canada measure is 19.2% of businesses using AI to produce goods or deliver services in the 12 months preceding the second-quarter 2026 survey, up from 12.2% a year earlier; a separate third-quarter 2026 survey found 25.2% planned such use in the next 12 months.[2,3] These broad AI measures do not establish agent adoption or agentic spend. The strategy offers SMEs an additional $700 million in affordable sovereign compute. Partnerships still being finalized have proposed 850 MW of capacity by 2030; that is proposed capacity, not capacity already committed or in service.[1] Measurement of useful outcomes would help assess value as these programs develop.

THE TORONTO LESSON

Two Toronto cases illustrate the market for inference efficiency. NVIDIA acquired CentML, whose software improved model performance on existing GPUs; AMD acquired the engineering team of Untether AI, which built inference hardware.[11] These transactions show strategic interest in Canadian optimization talent, but do not establish where all future value will accrue. TKF's proposed position is to measure the workload and the completed outcome across model and infrastructure choices.

A FIELD IN ITS FIRST MONTHS

This is a discipline measuring itself in real time. In October 2026 I attended AGNTCon + MCPCon North America in San Jose, where I heard from the Linux Foundation's newly formed Tokenomics Foundation — stood up only in August 2026 — as it set out an early State of Tokenomics drawing on responses from

close to five hundred companies.[12,13,5] That a body convened to define the economics of AI is barely three months old is not a weakness in the work; it is the clearest signal of how early this field is. The vocabulary is being proposed, debated, and standardized now, in its first months — which is precisely why a Canadian contribution, offered openly and held to evidence, can help shape it.

Selected public sources are numbered in the text and listed after the publication notice. Figures and program status are stated as of 27 October 2026.

SECTION 03

PROPOSED FOR PUBLIC USE

The measurement framework

> *No billing line or token counter reads cost at the level of the completed task.*

An agentic system can be read at four levels, from infrastructure consumption to the outcome served. The levels complement existing billing and telemetry specifications: FOCUS helps normalize billing data, while OpenTelemetry records model and agent operations, including token use. TKF's proposed contribution is the linkage of these inputs to task boundaries, outcomes, and controlled comparisons of overhead. A source's evidence class is retained at every level.[7,8]

LAYER	WHAT IT READS	EVIDENCE BASIS
Canopy — Outcome	Task cost, successful completion, quality, and service level, with failures and escalations counted in the denominator.	Measured only when task boundaries, outcomes, timestamps, and charges are observed; otherwise qualified
Branch — Workflow	Steps, tool calls, round trips, retries, and branch paths per task.	Measured where workflow traces are available
Trunk — Routing	Model choice, input and output tokens, cache use, provider charge, and routing by call.	Measured usage; billed cost reconciled where rates and invoices permit
Root — Silicon	Hardware, capacity, and energy attribution below the API line.	Measured on controlled hardware; otherwise modelled and labelled

THE SEVEN WASTE ELEMENTS

Seven recurring mechanisms can create avoidable cost. Each is a hypothesis to test against the operator's workload, pricing contract, quality threshold, and service level; none is automatically waste in every environment.

ELEMENT	IN PLAIN LANGUAGE
Model overkill	A higher-cost model used where a less costly model meets the same task quality and service level.
Context bloat	Repeated input context beyond what the task requires, demonstrated by a quality-equivalent shorter version.
Retry storms	Paid retries that do not improve completion, including repeated failures without suitable backoff or correction.

ELEMENT	IN PLAIN LANGUAGE
Round-trip excess	Additional billed steps beyond a quality-equivalent route to the outcome, with latency included.
Cache misses	Repeated context that fails to benefit from an available provider cache under the applicable contract.
Verbosity	Output longer than the task requires, after checking that compression does not reduce usefulness or quality.
Batch drag	Under-used reserved or hourly capacity; its economics differ from a pay-per-token API workload.

TWO PRICING RULES

Each candidate saving is compared with an achievable, quality-equivalent baseline and verified rates. One-at-a-time estimates are diagnostic: effects interact and must not be summed. A recoverable total requires a jointly tested configuration with comparable task quality, completion, and latency. API charges, reserved capacity, tool costs, and human review are reported separately when material.

SECTION 04

FROM RAW TO RESIDUAL

The residual-overhead method

> *Residual overhead is what remains after the best available mitigations — the number that cannot be waved away.*

FROM RAW TO RESIDUAL

Infrastructure already employs mitigations such as cache-aware scheduling, paged attention, and tiered storage. The useful question is the overhead that remains after the specified mitigations on the tested stack. "Residual" is therefore relative to a declared workload, qualified reference, service level, and configuration, rather than a universal constant.

THE DUAL-CONFIGURATION MEASUREMENT

Run the same representative workload with identical instrumentation in a current, post-mitigation configuration and a branch-aware comparison configuration. Hold hardware, model, task mix, prices, and quality criteria constant where possible; randomize or pair repeated runs to account for variability. The observed cost difference estimates incremental recovery under those conditions. To report residual overhead, separately define a qualified reference and compare the remaining cost with it. Without that reference, the experiment establishes an incremental difference, not the total residual. Record completion, errors, quality, and latency alongside cost.

THE EPISTEMIC DISCIPLINE

Each reported figure carries its evidence class: measured (directly supported by reported telemetry), modelled (derived using explicit assumptions), or indicative (a weaker signal). A recorded token count does not itself measure outcome quality or recoverable cost. Missing fields remain "not reported," never zero. Dollar estimates require verified applicable rates and a stated treatment of cache, tool, reservation, and other charges; hardware attribution beneath an API bill remains modelled unless directly observed.

OBSERVED IN PRACTICE

TKF operates this framework on its own estate. Three in-house agents — competitive monitoring, sales enrichment, and document review — feed model-call records to a shared telemetry schema. A settled week of operation exposed an instance of context growth in TKF's own workload. This is a qualitative internal observation, not a Canadian benchmark or a quantified saving. Before any numeric result is released, TKF will state the observation window, task and success definitions, sample size, applicable rates, and reproducibility limits.

BENCHMARK AGGREGATION RULES

Cross-organization benchmarks will include only records whose evidence basis and comparable task definitions are documented. No cohort figure will be released from fewer than five estates; small or concentrated cells will also be suppressed where an operator could be inferred. Missing values remain "not re-

ported.” TKF will disclose coverage, sample mix, service levels, and the limits on extrapolation. A minimum cohort size alone is not a guarantee of anonymity.

SECTION 05

THE RESEARCH PROGRAMME

What Canada has not yet measured

> *A discipline begins by naming, honestly, what it has not yet measured.*

The following questions define a Canadian research programme. TKF has not identified a published Canadian, cross-estate measurement using the residual-overhead definition in this report as of October 2026. Broader AI cost and value research already exists.[5,6]

- **The national residual.** What post-mitigation overhead remains across comparable Canadian agentic workloads, against a declared qualified reference?
- **The archetype distribution.** Which task shapes — triage, enrichment, review, and others — dominate deployments, and how do their cost, success, quality, and latency differ?
- **The efficiency of supported compute.** How can public support be linked to completed, quality-acceptable outcomes while accounting for hardware utilization, electricity, water, and local impacts?[1,10]
- **The recoverability split.** Which gains arise from workflow, routing, and caching changes on existing systems, and which require infrastructure or architectural changes?

What I heard in San Jose only sharpened the point: across the room, everyone is early. The vocabulary is still forming, and the measurements with it. Naming these gaps plainly is how the work begins.

SECTION 06

AN INVITATION, NOT A CLAIM

The measurement programme, and a notice of publication

> *No saving is promised before measurement.*

AN INVITATION TO CANADIAN OPERATORS

TKF invites Canadian operators of agentic AI — public sector, regulated industry, and enterprise — to participate as founding design partners in no-charge measurement engagements. TKF proposes to use existing telemetry, with task identifiers, timing, model and token counts, and outcome metadata as agreed with each operator. Collection should exclude prompt and response content by default, limit and protect identifiers, and specify retention, access, and subcontractors. Canadian residency alone does not establish sovereign control: metadata, monitoring, management interfaces, and access pathways must be governed through the full data lifecycle. Only consented, aggregated, measured results meeting the disclosure rules above would be published. No saving is promised before measurement.[9]

ABOUT THOMPSON KINKADE FOUNDRY

Thompson Kinkade Foundry Inc. is a federally incorporated Canadian company based in Toronto, developing measurement and optimization methods for agentic AI. ArborEx is its single-estate instrument; SilvaEx is a cross-estate benchmark in development, subject to the aggregation rules above. Contact: info@tkfoundry.ai · tkfoundry.ai

NOTICE OF PUBLICATION

This report describes a measurement framework and methodology that are the subject of a United States provisional patent application filed in September 2026 (patent pending). It is published to share the framework for examination and use by Canadian operators, program officers, and researchers; it is not legal, financial, or engineering advice, and it grants no licence under any patent or other intellectual-property right except as TKF may separately agree in writing. Public figures and program status are cited as of the dates shown and should be re-checked against the sources before being relied upon. © 2026 Thompson Kinkade Foundry Inc. All rights reserved.

REFERENCES

Sources & Notes

1 Innovation, Science and Economic Development Canada, “Canada’s National Artificial Intelligence Strategy: AI for All” (June 2026). https://ised-isde.canada.ca/site/ised/en/canadas-national-artificial-intelligence-strategy-ai-all

2 Statistics Canada, “Analysis on artificial intelligence use by businesses in Canada, second quarter of 2026” (11-621-M2026010). https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2026010-eng.htm

3 Statistics Canada, “Canadian Survey on Business Conditions, third quarter 2026” (31 August 2026). https://www150.statcan.gc.ca/n1/daily-quotidien/260831/dq260831a-eng.htm

4 NVIDIA, GTC 2026 keynote coverage (March 2026). https://blogs.nvidia.com/blog/gtc-2026-news/

5 Tokenomics Foundation, “The State of Tokenomics,” Release Candidate 1.0 (23 September 2026). https://www.tokenecon omics.com/state-of-tokenomics/

6 FinOps Foundation, “FinOps for AI” framework (accessed October 2026). https://www.finops.org/framework/technology-categories/ai/

7 FinOps Open Cost and Usage Specification (FOCUS), version 1.4 (June 2026). https://focus.finops.org/focus-specification/

8 OpenTelemetry, “Inside the LLM Call: GenAI Observability with OpenTelemetry” (14 May 2026). https://opentelemetry.io/blog/2026/genai-observability/

9 Innovation, Science and Economic Development Canada, “AI Sovereign Compute Infrastructure Program” guide (2026). https://ised-isde.canada.ca/site/ised/en/program-guide-artificial-intelligence-sovereign-compute-infrastructure-program-scip

10 Government of Canada, “Responsible Data Centre Development Principles” (September 2026). https://www.canada.ca/en/innovation-science-economic-development/news/2026/09/government-of-canada-launches-canadas-responsible-data-centre-development-principles.html

11 BetaKit, “NVIDIA acquires Canadian AI efficiency startup CentML” (30 June 2025), with contemporaneous reporting on Untether AI. https://betakit.com/nvidia-acquires-canadian-ai-efficiency-startup-centml/

12 Agentic AI Foundation (Linux Foundation), AGNTCon + MCPCon North America, San Jose, 22–23 October 2026. https://events.linuxfoundation.org/agntcon-mcpcon-north-america/

13 Linux Foundation, “Linux Foundation Launches the Tokenomics Foundation to Define the Economics and ROI of AI Value” (4 August 2026). https://www.linuxfoundation.org/press/linux-foundation-launches-the-tokenomics-foundation-to-define-the-economics-and-roi-of-ai-value

14 Ke Yang and ChengXiang Zhai, “Ten Principles of AI Agent Economics” (University of Illinois Urbana-Champaign, arXiv:2505.20273, 2025). The term “agent economics” is used here in the practitioner sense of this report; TKF claims no coinage of the phrase. https://arxiv.org/abs/2505.20273

THE TK FOUNDRY PAPERS

Second Paper No. 02 · The State of Agent Economics in Canada · Thompson Kinkade Foundry Inc., Toronto · 2026. Published to share a measurement framework for examination and use; the framework and method are the subject of a U.S. provisional patent application (patent pending).

© 2026 Thompson Kinkade Foundry Inc. · tkfoundry.ai