GOVERNED DATA FOR AGENTIC AI SERIES · EPISODE 3 | VIDEO PODCAST BRIEFING | July 2026

Tokenmaxxing and the Rise of Smarter AI Workflows

Chris Willis
Chief Design Officer, Domo

Joe Clark
Senior Software Architect of AI, Domo

EXECUTIVE SUMMARY

Chris Willis and Joe Clark examine the rise and reckoning of "tokenmaxxing," the idea that AI productivity can be measured by the number of tokens used. They explain why that metric can lead teams toward **higher costs, inconsistent outputs**, and **poorly matched AI workflows**. Drawing on Domo's engineering experience and a customer example, they outline a more practical approach: Match the AI method to the task, manage context carefully, choose models by use case, and use governed tools that connect AI work to business outcomes. Their core message: The future of enterprise AI depends less on using more tokens and more on using the right model, the right context, and the right workflow for the job.

1 The Tokenmaxxing Moment and Its Reckoning

"Tokenmaxxing" emerged as a popular way to measure productivity: The more AI prompts someone sent, the more tokens someone used, the more AI-forward they appeared to be. For companies still learning where AI fits, that kind of metric can encourage people to try new tools and test new use cases. Joe noted the thinking like this:

> *"Some people got focused on 'If you're not spending tokens, you're not trying hard enough.'"*
>
> — Joe Clark

But token volume isn't the same as value. Chris noted that Domo intentionally avoided token leaderboards because usage is easy to measure but doesn't necessarily show whether the work produced a useful outcome.

The takeaway: AI experimentation matters only when it leads to something more measurable than consumption.

2 Usage-Based Pricing Exposed the Cost Problem

Flat-rate pricing made it easy to treat AI as, in Chris's words, "an all-you-can-eat token buffet." But that era is ending. He cited three high-profile signals:

- One company reportedly spent $500 million in a single month on AI after failing to set limits on a Claude deployment.
- Uber burned through its entire 2026 AI budget in a single quarter.
- Amazon took down its internal token leaderboard.

The concern isn't just that AI can become expensive but that spending may not connect to results. You could run an agent, forget it's running, and end up with a large bill even if the agent went on to delete your entire production database.

> *"The cost isn't necessarily just in tokens."*
>
> — Joe Clark

Token usage captures only one layer of cost. Companies also have to account for failed tasks, unnecessary retries, poor outputs, governance risk, and the human time required to review or repair what AI produced.

3 Price Per Token Doesn't Tell the Task Cost

Price-per-token comparisons can be misleading. A model with cheaper tokens may cost more if it needs more reasoning steps, more retries, or more tool calls to finish the same job. A more expensive model may complete the work more efficiently.

> *"A more intelligent model might be able to get to the point a little bit more quickly. It may have to think a little less. A smaller model has to experiment and call more tools, maybe get some wrong, and retry."*
>
> — Joe Clark

Real cost should account for accuracy, time to completion, total token usage, retries, failed attempts, and downstream review or correction alongside the per-token rate.

4 The Core Framework: Match the Tool to the Task

One of the first things Joe looks at when working with customers on inefficient or underperforming agents is whether they're solving deterministic problems with a non-deterministic solution. A deterministic task has a fixed, predictable set of steps. A weekly customer summary that pulls the same metrics, populates the same template, and sends the same email doesn't need a full agent deciding what to do each time. Routing it through an agent may generate a different query each run, producing unreliable results from week to week.

> *"There's no reason to make the model generate that query over and over again, spending all those tokens, when you can just do it one time."*
>
> — Joe Clark

The framework moves through three tiers:

TIER 01

Deterministic Workflows

Use structured workflows when the steps are known, repeatable, and rule-based.

TIER 02

Traditional AI Tasks

Use large language models for classification, sentiment analysis, data extraction, and text generation when flexibility adds value but a full agent loop is unnecessary.

TIER 03

Full Agentic Loops

Use agents when the task requires multi-step reasoning, dynamic decisions, tool calling, or adapting to different inputs. Here, a larger and more capable model is often the more cost-efficient choice because it navigates complexity with fewer retries and better tool-call decisions.

5 Context Engineering: More Isn't Better

More context usually means more tokens. A large context window can help, but filling it doesn't guarantee a better answer.

> *"Just filling up an entire token context window doesn't mean you're going to get the best answers."*
>
> — Chris Willis

Too little context and the agent may not have the information it needs. Too much context and the model has to sort through unnecessary information, which can raise cost and reduce accuracy.

> *"You see a dramatic decrease in accuracy just because the model has to wade through that much more information and it's less likely to find the correct answer. Too many tokens, not only is it more expensive, it does reduce accuracy as well."*
>
> — Joe Clark

The goal is relevant context at the right moment. Instead of sending everything to the model upfront, teams need ways to retrieve the right data, document, or business rule for the task at hand.

Domo supports that approach through a governed knowledge layer that queries the right dataset or document at the right moment rather than loading everything upfront, scoped to each individual's data access permissions.

6 Model Selection: Size for the Task, Evaluate for Your Use Case

Organizations are now building infrastructure around three things:

- Visibility: What are the models actually doing?
- Predictability: Are they producing consistent results in the same situation?
- Control: Are they aligned to the tasks and outcomes the organization is working toward?

Joe's guidance is that complex tasks with full agent loops benefit from larger models; simpler tasks can be handled well by smaller, faster alternatives like GPT Mini, Claude Haiku, or open-source models. But he was clear that general guidelines only go so far:

> *"Nothing trumps evaluation for your specific use case. Get some test cases in. Evaluate accuracy alongside time to execution and cost, and find the right balance."*
>
> — Joe Clark

Chris added that fine-tuning is re-emerging as a practical option for specialized workflows. Rather than building a large bespoke model, companies can tune a general-purpose model for a specific business domain or task.

7 How Domo Helps Customers Optimize AI Usage

Joe identified three recently released DataFlow tiles as among the most impactful products Domo has shipped:

- Text generation
- Classification
- Sentiment analysis

Each tile applies LLM capabilities at scale without requiring teams to write code or make individual API calls.

> *"It lets us really apply the capabilities of LLMs at scale. And one of the things that we get by using batch processing is cheaper token costs."*
>
> — Joe Clark

Batch processing can reduce inference costs for use cases that don't require a full agent loop. It also gives teams a simpler path from AI output to dashboards, apps, and business decisions.

Customer Example

One large Domo customer needed to run sentiment analysis across millions of survey responses, analyzing each one against five specific topics. Previously, they built the pipeline in Domo's Jupyter Workspace, writing Python code and making individual API calls, generating billions of tokens at full cost.

With the new DataFlow tiles, the same workflow requires dragging one tile onto the screen, wiring up a dataset, specifying the topics, and writing to an output. That output then feeds dashboards where actual business decisions are made.

Result: For similar use cases, the new tiles can reduce inference costs, shorten setup time, and make the output easier to feed into dashboards and apps.

Domo also supports more complex use cases through its Agent tile. When a full agentic experience is required, the agent can use knowledge from datasets or document collections and query what it needs to complete the task.

Governance remains central. Models should receive only the information a person has permission to access.

> *"Whatever we send to the model really needs to be the specific information that the user has access to. You can't train private data into the model and expect the model to keep that secret for you."*
>
> — Joe Clark

8 The North Star: Outcomes, Not Tokens

Tokenmaxxing encouraged experimentation, but usage isn't the same as value. As AI costs rise and agents become more capable, companies need stronger measures of success.

Better questions include:

- Did the workflow complete the task accurately?
- Did it reduce manual work?
- Did it produce a consistent result?
- Did it use the right model for the job?
- Did it protect governed data access?
- Did the output help someone make a better decision?

Domo's position: The infrastructure for smarter AI already exists. The path forward is discipline, not more tokens.