You can't consume your way out of AI Cost Opacity

Palantir's government AI momentum is real. So are the contract, budget, and tokenization risks no one should ignore.

Martin Jennings, General Manager, Product Group, Permuta Technologies *16 Jun 2026*

Executive Summary

Artificial intelligence adoption inside the government market is increasingly shifting from experimentation to operational deployment. As that shift occurs, the pricing model behind AI is also changing. Flat subscription assumptions are giving way to hybrid billing structures tied to token consumption, compute utilization, platform access, and variable model selection. The impact of these changes was recently highlighted when Microsoft changed its AI pricing to align with the industry trend of transitioning away from subscription models toward tokenization (https://www.businessinsider.com/github-copilot-token-uage-pricing-change-reaction-2026-6). This pricing transition matters because it changes AI from a predictable software line item into a ungoverned and dynamic operating cost that can vary greatly as usage expands across an organization's user base.

Palantir is one of the most important companies to analyze within this shift because its government business is already large, mission-facing, and deeply embedded in defense, health, and intelligence environments. Palantir's 2024 annual report states that 55% of total revenue came from government customers, and the company continues to position itself for expanded U.S. Federal software spending. Its public contract footprint includes major operational programs such as the Department of Defense Maven Smart System, the Department of Health and Human Services BPA, NHS England's Federated Data Platform, and intelligence-related work in allied governments.

The issue is not simply that Palantir uses AI. The more important issue is that Palantir's own Artificial Intelligence Platform (AIP) documentation shows that model interactions are measured through tokens, translated into compute-seconds to reflect the pricing of the underlying model provider, and exportable by model, application, resource, and currency. That means the market-wide shift toward tokenized and consumption-based AI billing is not theoretical in Palantir's case. It is already present in the architecture that underlies Palantir's AI-enabled offering set.

This paper examines how those tokenization mechanics can affect the economics of Palantir's existing government contracts. It argues that as Palantir programs scale across users, missions, and workflows, AI usage can become a meaningful driver of marginal cost, contract ceiling pressure, and oversight burden. The result is that agencies may face growing

exposure to variable AI costs inside contracts that were initially understood primarily as platform procurements.

The central conclusion is that Palantir's government contract momentum and the industry's tokenization trend are converging in a way that demands closer scrutiny from acquisition leaders, budget officials, and program managers. The key procurement question is no longer limited to whether Palantir's software delivers mission value. It is whether Government buyers can maintain cost transparency and contractual control once unfettered AI usage in the hands of end users becomes operationally indispensable.

Introduction

The first paper in this Permuta Technologies series, *You can't prompt your way out of Enterprise Complexity*, argued that AI enthusiasm often outruns the economic, architectural, and governance realities of enterprise deployment. That argument applies with particular force to Palantir because Palantir sits at the intersection of secure mission software, operational AI, and large public-sector contracts.

Palantir is not a generic developer tool vendor. It operates in environments where customers care about data fusion, decision support, workflow orchestration, auditability, security controls, and operational continuity. Those characteristics make Palantir valuable in government settings, but they also make the economics of AI usage harder to simplify. Once AI functionality is embedded into operational workflows, cost no longer scales only with licenses or labor. It can also scale with prompt size, response size, model selection, concurrency, retrieval, and agent activity.

This creates a practical policy problem. Government buyers are accustomed to evaluating software through fixed licenses, subscriptions, labor rates, and contract ceilings. Tokenization changes that logic by introducing a variable usage layer that may not be obvious from the headline value of the contract itself.

Palantir's Government Position

Palantir's annual report describes the company as having been founded to build software for the U.S. intelligence community, and that history remains central to how the company positions itself today. The same filing states that AIP is bundled with Foundry, Gotham, and Apollo and is designed to enable responsible artificial intelligence across government and enterprise environments.

That positioning matters because Palantir is not selling isolated model access. It is selling an operating environment in which data, workflows, security controls, and LLM-backed functions are connected inside a mission system. This is a materially different economic

structure from a simple software subscription because model-backed interactions can become embedded throughout the workflow stack.

Palantir's contract pattern reinforces this point. Its public-sector work spans defense, health, transportation, and allied intelligence relationships, and several awards show that once adoption begins, enterprise expansion can follow. The most visible example is Maven Smart System, where Palantir's initial five-year, $480 million contract was later expanded by $795 million due to growing demand, bringing the ceiling to nearly $1.3 billion through 2029.

Why Tokenization Matters

Palantir's AIP documentation provides direct evidence that tokenization is built into the company's AI cost model. Palantir states that prompts and responses consume tokens, that those tokens are converted into compute-seconds to reflect the cost of the backing model provider, and that usage data can be exported by application, model, resource, compute-seconds, and currency. It also states that usage can arise across multiple product surfaces, including AIP Assist, AIP Logic, Code Assist, Workshop tools, Quiver tools, Pipeline Builder tools, and direct Language Model Service calls.

This means cost expansion inside Palantir deployments can occur through several mechanisms. Larger prompts consume more tokens. Longer outputs consume more tokens. More users create more interactions. More agentic workflows create more model calls. More demanding mission use cases may require more capable and more expensive models. In short, Palantir's AI economics are usage-sensitive by design.

This is where the broader market shift from VC-subsidized AI pricing to cost-realistic billing becomes relevant. During the early adoption period, many AI vendors could underprice access to accelerate growth and normalize consumption. As model costs, infrastructure demands, and investor expectations become harder to subsidize, vendors are under pressure to recover real cost through variable or hybrid billing. Palantir's documentation suggests that it has already built the telemetry and accounting structure necessary for that transition.

The Hidden Cost Problem

The same dynamic described in the first Permuta paper also appears here in a more contract-specific form. AI value is often marketed through productivity outcomes, but the underlying economic model may depend on token usage, model tier, orchestration complexity, and continued infrastructure spend that are not obvious in topline pricing.

For Palantir customers, that hidden-cost problem is amplified by mission conditions. Government programs do not operate in a vacuum. They experience crisis surges, operational spikes, evolving user communities, and expanding workflow demands. The

moments when agencies most rely on AI may also be the moments when usage costs accelerate the fastest.

This creates at least five practical risks for government buyers:

- Budget unpredictability, because usage may scale with mission tempo rather than planned seat counts.
- Reduced cost transparency, because platform value and model-consumption cost can become intertwined.
- Ceiling pressure, because successful operational adoption can increase obligations faster than expected.
- Source-selection difficulty, because competing offers may structure AI pricing through very different mixes of license, subscription, metering, and cloud pass-through.
- Lock-in risk, because workflows, governance, attribution, and operational habit may all converge inside one vendor environment.

Toward a Better Contract Model

The lesson is not that Palantir should be avoided. The lesson is that Palantir contracts involving AI-enabled workflows should be evaluated with more economic precision than many public-sector software buys historically required.

A more defensible acquisition posture would separate fixed platform and governance costs from variable model-consumption costs, require usage reporting at a level that supports agency chargeback and audit, and establish ceiling protections or tiering rules for surge scenarios. It would also force acquisition teams to evaluate mission value against the full operating cost of AI, not merely the initial access price or productivity claim.

In the same way that the first white paper argued that leaders cannot prompt their way out of enterprise complexity, this paper argues that agencies cannot consume their way out of AI cost opacity. If tokenization remains poorly governed, then even successful Palantir deployments may become economically harder to explain, budget, and compete over time.

Permuta's Contrasting Approach

A useful contrast to Palantir's contract and pricing posture is Permuta's own approach to mission software and AI enablement. Permuta's DefenseReady is a Commercial Off-The-Shelf readiness platform built on Microsoft Dynamics 365 and the Microsoft Power Platform, designed to give defense and government customers an integrated view across workforce, training, medical, assets, security, budget, mission, and workflow functions.

That architecture matters because Permuta is not positioning AI as the foundation of the platform. It is positioning AI as an optional extension on top of an already mature mission

system. Permuta's recent Memorial Day 2026 platform update states that DefenseReady is built on Microsoft Power Platform and Dynamics 365, is deployed across IaaS IL5/6, SaaS IL5, and on-premises environments, and ships its AI capabilities disabled by default. The configuration guide similarly describes AIReady as a metadata framework that prepares DefenseReady data for AI workloads while preserving bring-your-own-model flexibility and avoiding costly vector-search-heavy architectures where they are not required.

This is a meaningful difference from the market's more consumption-forward AI story. Palantir's public AIP documentation shows that model interaction is a meterable layer inside the product environment, translated into compute-seconds and attributed by application, model, and currency. Permuta's model, by contrast, is structured to let customers control whether AI is enabled at all, which model endpoint is used, and whether that endpoint lives in Azure OpenAI, OpenAI-compatible on-premises runtimes, Ollama, or other approved Government first AI environments such as Ask Sage or GenAI.mil.

Why Permuta Is Different

Permuta's ReadinessPlatform offerings have five substantive differences that are worth highlighting in contrast to Palantir.

- Mission platform first, AI second. DefenseReady is presented as a mature readiness platform with established workflows and modules, not as an AI-first product searching for operational use cases after deployment.
- Customer-controlled model choice. DefenseReady allows customers to bring their own model and swap providers without rewriting application logic, which can reduce dependency on a single commercial AI stack.
- Security posture by deployment type. Permuta's AI guidance distinguishes between cloud, government cloud, self-hosted, air-gapped, and on-premises patterns, with controls such as private endpoints, managed identities, IAM roles, secret managers, and disabled-by-default feature activation.
- Cost discipline by design. Permuta's AI roadmap and configuration guidance intentionally avoid forcing every customer into the most infrastructure-intensive AI path. Level 1 AI focuses on metadata-driven experts and conversational assistance, while higher-complexity reasoning and agentic patterns are staged later and tied to more deliberate maturity requirements.
- Engineering realism. Permuta's own white papers and internal messaging do not claim that AI replaces disciplined software engineering. They explicitly frame AI as a force multiplier, caution against hype, and state that pure code-generation approaches were not adopted for core product development because of complexity, security, and rework concerns.

This last point is especially important for the contrast. In recent press releases to their customers and partners, Permuta highlighted that while they use AI tools such as GitHub Copilot, Microsoft Copilot, and a handful of other AI capabilities as force multipliers across the development lifecycle, they do not use pure code-generation AI for core product development because the resulting code requires more rework than manual changes at current AI tool maturity levels when applied to large scale enterprise solutions such as their ReadinessPlatform. That development posture supports a different AI story than the one often sold by token-driven AI vendors. Rather than arguing that AI will replace the need for disciplined architecture, Permuta's documents argue that architecture, accreditation history, tested workflows, and training infrastructure are the things that make selective AI adoption viable in the first place. The result is a more incremental, controlled, and operationally legible adoption path, particularly for defense customers that must preserve continuity, accreditation, and auditability. This is a materially more conservative and operationally grounded position than the broader AI market narrative that assumes AI-generated capability automatically translates into lower total cost.

Architectural And Economic Contrast

The strongest contrast between Palantir and Permuta is not simply that one uses more AI than the other. The stronger contrast is where each company places AI in the system architecture and commercial model.

Palantir's public materials indicate that AI consumption can become an operating layer embedded within the mission platform itself, creating a path where increased mission reliance can translate into increased usage-sensitive cost. Permuta's architecture, by contrast, appears designed to keep the mission platform stable while letting the AI layer remain modular, configurable, and in some cases removable or disabled without disrupting the baseline application environment.

That difference has procurement implications. If AI is deeply fused into the commercial and operational core, then usage growth can become hard to isolate from platform value. If AI is modular, disabled by default, and customer-routable to approved endpoints, then agencies may retain more control over cost, hosting model, data handling, and transition strategy. This does not eliminate variable AI cost, but it can improve bargaining power and reduce the degree to which model-consumption economics are hidden inside a larger enterprise platform story.

What This Means for Buyers

For acquisition leaders, the contrast between Palantir and Permuta is not a simple good-versus-bad comparison. Palantir appears optimized for high-scale operational AI and

decision support in environments where rapid AI expansion can justify larger contract ceilings. Permuta appears optimized for mission software continuity, Commercial Off-the-Shelf (COTS)-based readiness workflows, and a more controlled AI-enablement path that preserves customer control over hosting, activation, and model choice.

That means the better fit depends on the problem being solved. If the customer wants an AI-forward operational environment where model-backed workflows are central to mission execution, Palantir's posture may be attractive, but buyers should expect more scrutiny around consumption, scaling, and hidden cost behavior. If the customer wants to add AI to an existing readiness and workforce management operating model without surrendering architectural control or forcing immediate AI dependency, Permuta's approach may offer a more fiscally legible and operationally conservative path.

Third-Party Market Assessment

A useful additional contrast comes from Hawkins Group's December 2025 third-party industry comparison, which evaluated Permuta relative to competitors using perceptual mapping and Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis focused on total cost of ownership and advanced Human Capital Management (HCM)/Work Force Management (WFM) capabilities. That assessment concluded that Permuta sat in the low-cost, high-capability quadrant, while Palantir sat in a high-cost, moderate-capability position for this specific workforce and readiness problem space.

That third-party framing strengthens the distinction already developed in this paper. Hawkins Group characterized Permuta's value around lower total cost of ownership, strong workforce and readiness foundations, Microsoft ecosystem integration, and established relevance in Federal and DoD environments. It characterized Palantir as expensive and potentially able to accelerate HCM and workforce-management delivery, but still as a higher-cost option with more moderate native fit for this mission set.

The appendix-level competitor matrix in the Hawkins Group analysis is even more direct. It describes Palantir as a data analytics platform rather than a system built for Human Resource (HR) or workforce execution, notes that it requires custom modeling and engineering teams, lacks native support for DoD 8140 , clearance workflows, or DoD HR standards, and carries high total cost of ownership because of software plus heavy services. By contrast, the same matrix describes Permuta as mission-driven, faster to deploy through mission templates, lower in Total Cost of Ownership (TCO) under a COTS model, and stronger in native support for DoD-specific functions such as billet tracking, Personnel Tempo (PERSTEMPO), readiness dashboards, congressional reporting, and integrations with systems such as Defense Information System for Security (DISS) and the Defense Civilian Payroll System (DCPS).

For readers who want the underlying third-party document directly, the hosted Hawkins Group analysis is available here: Permuta Industry Comparison, Hawkins Group, December 2025

Conclusion

Palantir is a useful case study because it shows how quickly AI economics can move from technical abstraction to procurement reality. The company's public disclosures show a government-heavy revenue mix and a product portfolio built to operationalize AI in sensitive environments. Its AIP documentation shows that tokenized usage is already measurable and attributable. Its contract history shows that successful programs can expand materially as adoption grows.

Taken together, those facts support a simple conclusion. The next phase of AI oversight in government will not be defined only by model performance or mission utility. It will also be defined by whether agencies can understand and control the cost behavior of AI once usage becomes embedded in mission workflows.

For acquisition leaders and senior decision makers, the strategic issue is no longer whether Palantir can deliver AI-enabled capability. The more important question is whether the contract structure makes the long-run cost of that capability visible before dependence makes renegotiation difficult.

About the Author

General Manager, Product Group, Permuta Technologies
Martin "Marty" Jennings, Colonel (retired), United States Air Force

For the past 3 years, Marty Jennings has been the General Manager for the Product Group at Permuta Technologies, where he leads a team of Microsoft-certified developers responsible for more than 400 applications supporting unique Federal and Military use cases worldwide. Prior to joining Permuta, he served as Chief Cloud Architect and Chief Engineer for Leidos' 10-year, $10B GSM-O II contract with the Defense Information Systems Agency (DISA). A 30-year U.S. Air Force veteran, Marty retired in the rank of Colonel in September 2021. His military and civilian career spans software development, cybersecurity, and large-scale enterprise system program management in support of multiple combatant commands and agencies.

LinkedIn: https://www.linkedin.com/in/jentek