Information

How the Government Should Buy AI Tools

How to -NOT- buy AI tools, and some tips on how to buy them too...

Sep 25, 2024

Here's the problem:

Let's set a couple of premises up front:

1. Vendors selling AI tools by the user seat is terrible

2. Vendors selling AI tools where you pay token is likewise terrible

Why is this setup terrible? Pretty simple, it deters or prevents you as the government customer from deploying the tools to your users.

Think about it, if the vendor is selling by the seat, then you have to buy them one at a time, or in blocks. If it's one at a time then you have to make a transaction every time you want to add a new user. Often the amount of paperwork to spend $10, $10,000, or $250,000 is the same, so buying them one at. time is onerous to say the least. Likewise, if you buy them in blocks then you will almost certainly have seats go unused for periods of time, thus wasting your resources.

The same goes double for token limits on your users. No one opens Microsoft Word with the latent stress that they're going to get 5 pages into a document and suddenly run out of words. That would be insane. But that is exactly what happens with these token-limited products. You start using the tool and suddenly you're out of tokens, so now you have to buy more, but let’s go back to whole "amount of paperwork to spend $10, $10,000, or $250,000 is the same" issue, can you just buy more tokens? Probably not, so you're stuck going back to the manual way with your task half-done.

Why is it like this?

For what it's worth, this makes a ton of sense from the vendor perspective. Selling by the seat is easy to structure in a MAC/IDIQ contract price list. As for the tokens, those are a serious variable cost in high-demand AI-powered systems. Think about it, the vendor pays for the LLM to operate, and they get paid the same for a heavy user as a light user. If they don't put token limits on the users, then a heavy user could burn through all the profit margin on their seat and then next 20 seats in their office.

If you break this down from the industry perspective there's a few components of each user seat:

1. The overall platform/product: the vendor is almost certainly changing for the general use of their platform, the development that went into it, the continued support that keeps it running, etc. This is a pretty fixed cost.

2. Configuration of the platform for the overall customer organization: this typically come with some heavy up-front work to tailor or provision a product to a new government user organization. This is a modestly variable cost but is typically one-time.

3. Configuration/provisioning of individual user accounts: this typically comes with some role-based controls, probably some organization configuration. This is a very variable cost; you could have hundreds of users join all at once at the beginning and then much smaller numbers later

4. Ongoing customer service/support: think of your help desk tasks; password resets, browser issues, new account setups, old account shutdowns. Again, a very variable cost.

5. Usage costs: this is where tokens come in pretty heavily with AI tools, think of tokens as the fuel that you pump into the AI engine. Again, if the vendor charges you a set price for a bus ticket (seat cost), they typically limit how far you can go (usage) based on the price of gas (tokens).

Ok genius, what's your idea?

Solve it with contracting, honestly, with a strong preference for enterprise deployments and a combination of Fixed Price, T&M, and CPFF CLINs.

Here's how this could work:

CLIN 1 - FFP: fixed annual price for the overall platform/product

CLIN 2 - FFP: fixed up-front cost for configuration of the platform for the overall customer organization

CLIN 3 - FFP or T&M: fixed per-seat cost/hours for configuration/provisioning of individual user accounts

CLIN 4 - T&M: pool of cost/hours for configuration/provisioning of individual user accounts

CLIN 5 - CPFF or Other Direct Costs: pool of funding for tokens/compute

If certain users or groups consume more of CLIN 4 and 5 than they are due, you can manage that at the org level and leave the vendor out of it.

Also, you can merge CLIN 3 and 4 for simplicity.

Is this set up more complicated than a simple "$100/user/mo. **limitations apply**" model?

Sure, but it also aligns incentives and missions.

The vendor makes a profit, the government has no mission disruption, and everyone gets to use AI.

So What's the Catch?

There's always a catch, or shall we say a tradeoff:

This approach inherently requires a company that has an audited accounting system with a company that can execute cost-type contracts.

What does that mean? It means that your ease of mission execution narrows the pool of vendors who can provide directly to you.

Smaller, newer, companies who don't have a DCAA cert will either be locked out, or will have to work through a Value-Added Reseller (VAR) which is probably a decent way to do it anyway since these same smaller companies don't have access to the massive BIC/MAC/IDIQ contracts through which much software is purchased.

Parting Thought

Don't get too hung up on tokens, they are a near-term cost. In 3-5 years when the market normalizes and democratizes tokens will be commoditized, just like long-distance calling minutes, 4G/5G data access, and cloud storage/compute. This means that we probably should not run out and build an enterprise IDIQ for LLM tokens. This is a near-term problem that obviously requires near-term solutions, which can be dealt with similar to the analogous dynamic/usage-based costs that came before.