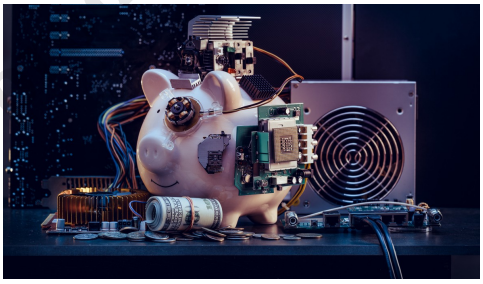
www.pipelinepub.com
Volume 22, Issue 10

Efficient Use of AI

By: Mark Cummings, Ph.D.



The initial exuberance over using GenAI in business and government is running into unexpectedly high online AI services bills. This is leading to a recognition that not all uses of AI are economically justified. This recognition is a sign of maturity. Now, the question becomes: how do you achieve AI efficiency?

Return on Tokens

The question is essentially a question of return. Return on fees paid to service providers and the cost of staff time. In this context, return means how much benefit you get from how much it costs. Online GenAI systems are billing based on tokens. A token is a small word (image, sound, etc.) or part of a larger word (image, sound, etc.). In online commercial systems, tokens are generally counted, and that is how the user is charged. Input to an AI is generally prompt language, data, background information, etc. Input is broken down into tokens. AI processing can be thought of as the AI talking to itself. This conversation creates more tokens. Then, the result is output, creating the final set of tokens.

Charging for retail users is generally through a subscription, which allows users a certain limited number of tokens per unit of time. Business accounts can be done the same way, but are more likely to be charged per number of tokens used. This second way is how businesses are being surprised by their bills. Staffs are using far more tokens than anticipated. are using far more tokens than anticipated.

Thus, the question of efficiency of use of AI boils down to how many tokens plus how much staff time you are using to get what value of result. This is becoming known as Return on Tokens (RoT). There are essentially two sides of RoT: the cost side and the benefit side. Each can be pursued independently. However, they often have to be considered together.

The most important thing when working with AI's is knowing exactly what you want to achieve. Knowing what you want to achieve has two aspects: prompt language and architecture.

Prompt Language

Independent of everything else, the more accurate and complete your prompt language is, the more

efficient you are going to be in using AI.

If you ask a GenAI system to give you something described only generally. The AI will give you something. It will try to figure out what exactly you need. But you are not likely to get what you want. When working in a chat mode, this tends to lead to long, expensive iterations. Iterative interactions that consume more and more tokens and more and more of the user's time. Both expensive.

In an Intelligent Agent (IA) application, prompts that are not completely accurate will cause behavior problems in the IA. If you catch it in development, it will create a string of iterations using more tokens and taking more staff time. A bigger issue is if you don't catch the problem in development and it leads to aberrant behavior in the field.

Struggling with inaccurate prompt language can lead to multiple versions of an AI running. Sometimes, it is easy to forget to turn off one of the previous versions and a Ghost IA result. Ghost IA's consume resources that can be expensive and can pop up in the field and cause hard-to-explain trouble.

In an organization, you are likely to have a wide variety of language capabilities. Some will be able to be naturally exact and complete. Others will be able to learn to become accurate and complete. Some just don't have that skill. Training can, at least, make everybody sensitive to the issue.

One way to improve efficiency in this area is to use an AI to help craft prompt language.

Online vs Local AI

First, there is a trade-off between running locally and online.

Running locally gives you less latency and more control. It also guarantees access, as long as you have battery backup. There is a cost to acquire, power, and maintain local hardware. The resulting cost is not a function of the amount of usage (with the proviso that more usage may mean consuming more utility-provided electricity, and larger, more expensive hardware). You do have to maintain the local system(s) and follow the emergence of new models/hardware.

Running online has the risk of no access, greater latency, less control, and expense based on usage. However, it may give you access to larger, more powerful models. Most existing computers and phones can access it. There is no maintenance expense. As newer, more powerful models appear, they may be made easily available.

Efficient Online AI Usage

If you choose to use an online AI service, you are likely to have a choice of a number of different AI's at a number of different costs. For example, Anthropic today has five different levels of AI's that you can use. In Anthropic's (Frontier model company) documentation, the company recommends that you use the model that is appropriate to your task. Essentially, they are saying, don't use Einstein to assemble a grocery list.

Each one of the five different models that Anthropic offers reasons at different levels. Using the highest-level model consumes the most tokens to execute the same prompt language.

If the intent is to have an AI simply assist with a Web search, then one of the simpler models may be the most efficient, similarly with a simple IA.

On the other hand, trying to execute an extremely complex task or request may need the highest-power model that uses way more tokens. There are also models that have been customized for specific tasks such as software coding. Using a coding model to do a grocery list is also likely to be inefficient.

Less expensive online AI services (not Frontier model companies) are beginning to appear. They tend to fall into one of two approaches: offer open source/open weights models (OS/OW Ms) running in a large data center; or offer older versions of Frontier model companies. OS/OW Ms coming from China are beginning to get close to the performance of Frontier model companies. Some are concerned about using models developed in China, fearing theft of information. There are others who maintain that these models are perfectly safe.

Across all of the alternatives, the trick is to get staff to use the right model for the right task. There are three ways to do this.

First, access to models may be limited depending on function. That is, only give finance staff access to finance-customized models. Only give software developers access to models that have been customized for coding. Etc. Only give staff access to the lower-cost models. Or combinations of these.

Second, staff can be trained to select the right model for the right task.

Finally, an AI may be used to suggest, or actually route, to the most efficient model.

Efficient Local AI Usage

There are three ways that models can be run locally: In an organization data center or a cloud contracted provider; In smaller servers distributed around the organization; and finally on Edge devices such as notebook computers, phones, and small desktop devices like the Mac Mini and Asus version of the DGX Spark. Running locally generally involves running OS/OW Ms. Many come from China. However, Google's Gemma OS/OW Ms are competitive with many of the Chinese models. Also, there are older OS/OW Ms from Meta that may be very practical for small IA applications.

OS/OW Ms have to be downloaded. Hugging Face is a good source for models, infrastructure supporting their use, etc. An organization running models locally can avoid the per-token pricing plans. As pointed out above, there is cost associated with the hardware and maintenance of the OS/OW M software.

Another consideration is longevity. Local models will not be automatically updated. That can be an advantage or disadvantage. On the advantage side are consistency both for users and for Intelligent Agents. On the disadvantage side is the work required to investigate new models and to manage the upgrading process.

All of these considerations apply to both data center and Edge device implementations. However, the cost/performance comparison between Edge and data center will change over time. Data centers are going to get more powerful. However, Edge devices may improve more dramatically over time. For example, what runs today on a $10,000 device will run in a few years on a $1,000 one. Of course, the frontier models are likely to increase in size, at least for the next couple of years.

In this evolving environment, what is most efficient to run in one place today may be more efficient in another place tomorrow.

Architecture Alternatives

Based on the above, a number of different basic architecture approaches emerge:

Online Frontier models; Online proprietary models that are no longer Frontier;
Online OS/OW Ms:
Local OS/OW Ms:
Hybrid part online / part local (Apple seems to be taking this approach):
Hybrid Frontier model / OS/OW Ms.

There is one further architecture alternative. One in which activity is funneled through an AI that figures out which of the above is best suited to efficiently handle the task at hand and helps with crafting complete, accurate, and efficient prompt language.

Organization Decision Making Guidelines

Currently, organizations have to make usage and architectural decisions based on their functional and efficiency needs. Over time, it is likely that vendors will appear wth products that support a variety of the emerging architecture alternatives. Some of those solutions will be useful for some organizations. Some organizations will want to continue to create their own. What seems clear is that there will be an ongoing high degree of volatility.

Implications for organizations of ongoing volatility are: 1.) well-structured staff training will continue to pay dividends even as change occurs; and 2.) whatever architectural, usage, and billing decisions are made today, they should be made with an eye to being easily changed tomorrow.

Although there are indications that the results of AI can not be copyrighted, it appears that patent protection around what is covered here is still in effect. Thus, some consideration for, and search of, patents is prudent in making decisions in this space.