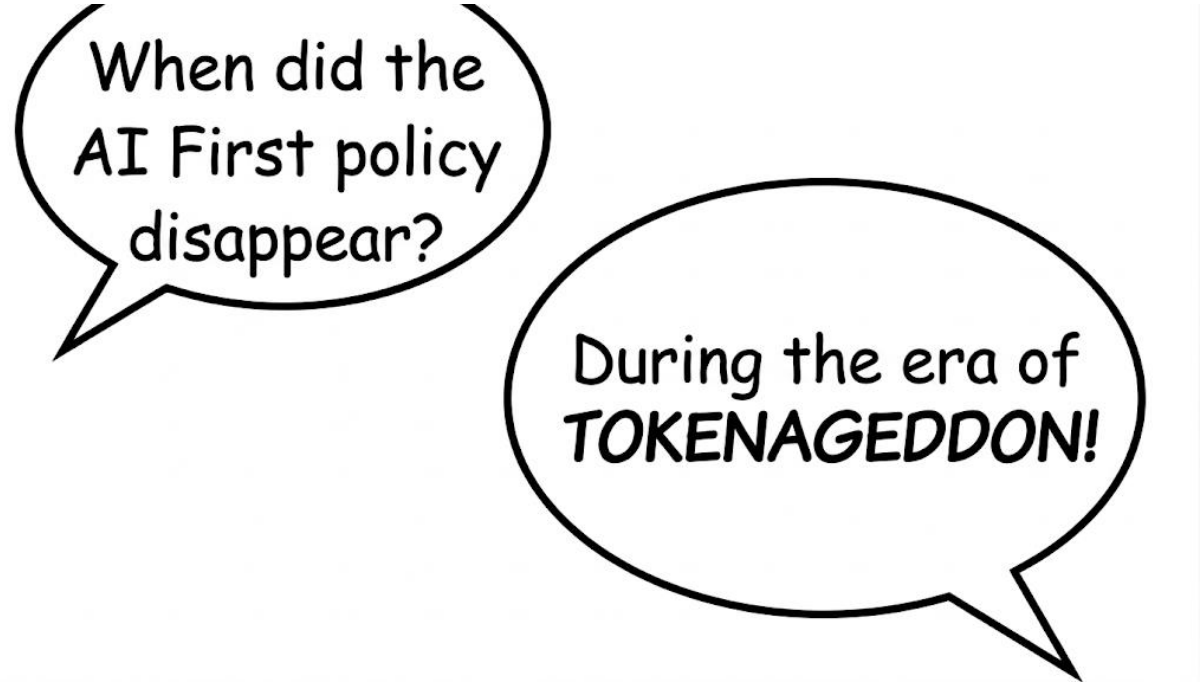


[Jean-Pascal Martin, Ph.D.]
Senior Consultant in AI Responsibility and Sustainable Digital
Digital4Better

Will Your Job Become AI First? And What About the Environmental Impact?

Until now, when you had a new task to perform, you had two options: do it yourself or ask someone else to do it for you. With **AI First**, you add a third option to consider before the other two: **ask AI to do it for you**.

This radical change in work methods could explode our AI consumption and its environmental footprint. Here's the logic:

- Priority will be given to AI agents for performing a large part of tertiary sector work, whether simple tasks or entire jobs.
- Our AI consumption will mainly be through these AI agents, which can themselves subcontract to other AI agents, potentially multiplying our token consumption by over 1,000.
- Companies will need to integrate the carbon footprint of AI services into their carbon balance, which could increase significantly.

- As a direct consequence, the footprint of generative AI could easily add between 550 and 1,650 kgCO2 per employee per year (I'll give you the details of the calculations later in the article)!

Have We Entered the Era of Tokenageddon?

In this article, I share my perspective, drawing on various sources without applying a scientific method. It is a forward-looking analysis aimed at highlighting upcoming challenges.

While everyone agrees that the use of generative AI is growing rapidly, it seems to me that most decision-makers and business leaders have not understood **the extent** and **the speed** of this growth.

Normally, I use AI for information synthesis and to aid reflection, but not for content creation. Specifically, I make about a dozen short queries per day, which corresponds to exchanging fewer than 35,000 words per day, or fewer than 50,000 tokens (1 token = 0.75 words on average).

Of course, if I had intensively used GPT, Mistral, or Gemini to write this article or other content, I could multiply this figure by 10 or 20. In that case, my consumption could reach **one million tokens per day**.

This trend of increasing token consumption is global: Between late 2023 and late 2025, *the average size of a prompt (input) quadrupled*, increasing from 1,500 to over 6,000 tokens [2]. The completion (output) has also tripled [3].

But this is still far below what developers practicing *vibe coding* consume. Their average consumption exceeds **100 million tokens per day**! Ask them, and they'll confirm it. 100 million tokens is about 200,000 pages of text. If we consider that 100 pages are 1 cm thick, that's a stack 20 meters high. **A 7-story building… every day.** This is a change of scale by several orders of magnitude.

To verify this staggering figure, you can visit the **tokscale** website, where you can see that top AI developers consume several hundred million tokens per day, and **sometimes exceed several billion tokens/day** consumed (higher than the Eiffel Tower if we continue with the paper stack metaphor).

The difference between them and me? I am not AI First; they are. Switching from one to the other can multiply our token consumption by 1,000.

Which Jobs Can Be AI-Transformed?

Are we sure that only a few developers are affected by these changes in practices? *Vibe coding* is expanding rapidly and could well be on the path to generalization, especially as it impacts a new audience of non-developers or *occasional developers* through a ripple effect.

I recently tested using AI development environments *for tasks other than software development*. Can we use Claude Code or Google's Antigravity to organize ideas, vacations, notes, write a book, create a board game, compose music, or create PowerPoints on any subject, for all personal or professional projects? I confirm that it works very well thanks to the incredible versatility of large language models and the effectiveness of the agentic approach combined with a dedicated workspace.

I was wondering if it would be relevant to invent an environment for consultants wanting to do ***vibe consulting*** when I saw that all the giants are organizing to shift all professions to this mode of operation. Two examples reveal that this is indeed the strategy of AI giants:

- OpenAI (ChatGPT) is launching **Frontier**, its platform of autonomous business agents that can specialize in payroll, HR, marketing, finance, customer support, purchasing… in short, everything. Currently in the testing phase, these agents will be offered to all companies worldwide.
- Anthropic is launching **Claude Cowork** to equip "white-collar" workers in a similar way to what they offer developers, with strong integration into Microsoft Office 365, Teams, and the workstation. The idea is to deploy specialized agents for all roles in the company. These ready-to-use agents are directly activatable: salesperson, generic manager, product manager, marketer, compliance & regulation assistant, finance, designer, HR assistant…

Does Cowork Consume More Tokens Than Normal Chat?

Yes. Working on tasks with Cowork consumes more of your usage allocation because complex, multi-step tasks are calculation-intensive and require more tokens to execute. Consider grouping related work into single sessions and using the standard chat for simpler tasks that do not require prolonged execution.

In summary, a transformation of all computer-using workers is underway and should make us ALL shift from occasional consumption of *chat-mode AI* to daily and intensive consumption of *agent-mode AI*. The idea that in the future we will all be **AI managers** is becoming a reality in 2026.

Matt Shumer thinks this transformation could happen at COVID speed. He adds that *« The experience tech professionals have had over the past year, seeing AI go from 'useful tool' to 'does my job better than me,' is what everyone is about to experience. Law, finance, medicine, accounting, consulting, writing, design, analysis, customer service. »* For Shumer, this future is not ten years away. It's already here.

Closer to home, **Frederick Marchand** (my boss) recently testified to the ability of AI to autonomously accomplish in one night what would have taken 10 days of full-time work.

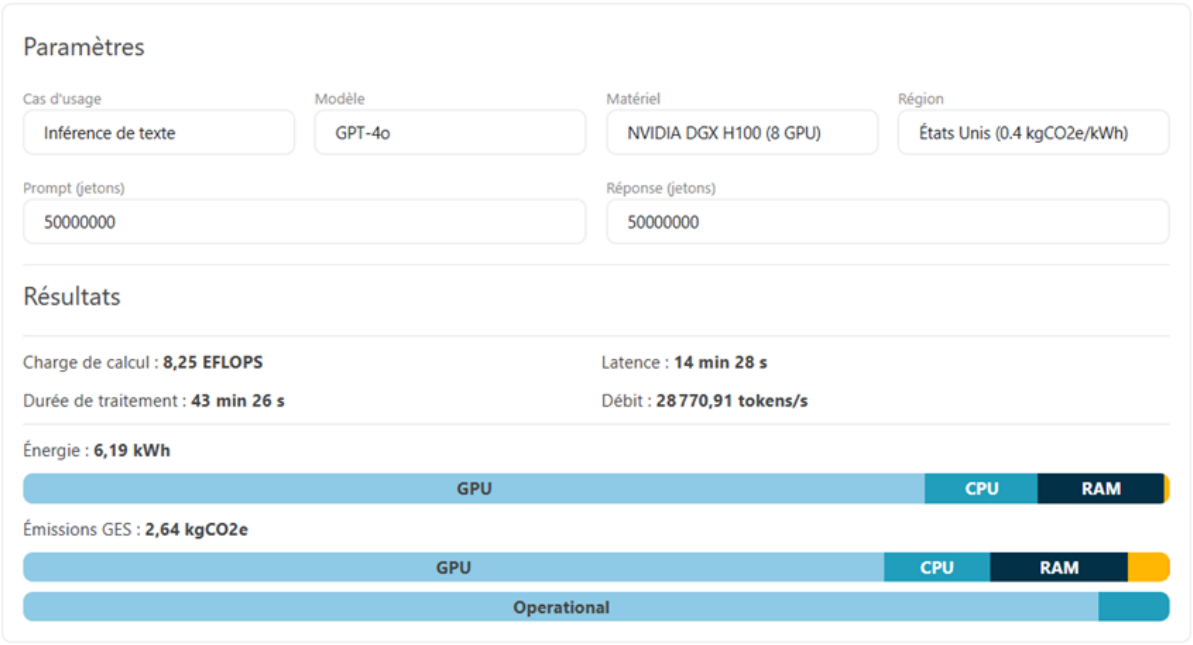
CO2 Emissions from AI First Could Be Measured in Kilograms Per Day

Concretely, what does consuming 100 million tokens per day, year-round, mean for all tertiary sector workers?

This answer is particularly difficult to formulate precisely because AI model providers are not very transparent about the resources consumed by their models. Ideally, we would like to know the amount of CO2e emitted per token. To get an idea, I propose comparing different estimation methods and presenting the annual impact for someone consuming 100 million tokens per day in the right column:

Source of Estimates	**Generative AI Model**	**CO2eq Emissions per Million Tokens**	**Annual Emissions of an AI First Worker (100 million tokens for 220 days)**
Google	Gemini (varied models)	0.075 kgCO2eq	1,650 kgCO2e
Digital4Better	OpenAI GPT 4o (with 50% cache)	0.026 kgCO2eq	572 kgCO2e
Digital4Better	Llama 3.1 405b	0.194 kgCO2eq	4,268 kgCO2e
Ecologits	Anthropic Claude Sonnet 4.0	1 kgCO2eq	22,000 kgCO2e
Ecologits	OpenAI GPT 4o	0.926 kgCO2eq	20,372 kgCO2e
Mistral	Mistral X	2.85 kgCO2eq	62,700 kgCO2e

These results are so staggering that I propose setting aside the most alarming ones based on Mistral or Ecologits methods, as they do not seem to account for the production optimizations that major LLM providers like Google are capable of. I will rely in particular on **our Digital4Better method** (published in OpenSource), which allows us to account for the ability to cache prompts and positions tokens at 50% input and 50% output. A **simulator is available** if you want to test your own hypotheses.



Estimation by the Digital4Better simulator of emissions from 100 million GPT 4o tokens with a 50% split between input and output tokens.

In summary, the annual environmental footprint of AI First for a worker consuming 100 million tokens per day for 220 days would range between 572 kgCO2eq and several tons. It seems clear to me that processing such volumes of tokens cannot be done on poorly optimized LLMs, which is why I focus on a *Mixture-of-Experts* (MoE) model like GPT 4o, which limits the number of active parameters to 88 billion (still a lot). Nevertheless, even with the lowest estimates, adding between 572 and 1,650 kgCO2eq is absolutely problematic. **A study by WenR** estimates that the average annual digital footprint for a European professional is 265 kgCO2eq. **AI First multiplies the digital environmental footprint of the worker by between 3 and more than 6!**

CO2 Emissions That Will Be Hard to Hide

First, I am surprised that the impacts of this work transformation are not more documented. Hiding hundreds of kgCO2e per employee per year should not be so easy unless they are simply not declared! But isn't the current mindset *« let's go full steam ahead, whatever the cost, we'll think about it later »*?

However, several limits could slow down this transformation:

- The cost of AI-First. While Claude Cowork licenses start at €20, you need a €100 license to use it effectively. And often, developers using this license complain about reaching the limits too quickly. The ChatGPT Pro subscription is €230/month. Companies will no longer be able to rely on Shadow AI.

- It is likely that license prices will increase with usage to amortize the considerable investment costs.
- The time required to deploy the new data centers necessary for AI-First. For example, for known investments, the construction of *hyperscalers* for France will extend until 2035. There are land, political, economic, and existing energy access constraints.
- The time required to transform professional practices. This transition seems to me at least as complex as moving from the role of operator to that of manager. It's no longer the same job.

I agree with the conclusion of the latest **Shift Project report**, which, regarding data center construction projects in France, warns that *« The connections validated today will reach full capacity around 2035 and risk inducing tensions in the electrical system and conflicts of use. (…) Maintaining the current dynamic would make the achievement of the sector's decarbonization targets for 2030 unfeasible, both in national inventory and in footprint. »*

In conclusion, in our troubled times, should we put all our energy into AI First when climate change is attacking us, the social impact is unpredictable, and we are not sovereign over this technology?

This article was not written by AI, but AI was used for reflection, information research and translation

[1] OpenRouter API Insights - Evolution of prompt sizes (2023-2025): **openrouter.ai**

[2] Data Engineer Things - Guide to tokens and costs: **dataengineerthings.org**