Efficiency and environmental impact of LLMs

Jill-Jênn Vie

March 26, 2025

In-context learning

Traditional ML: fit(train), predict(test)

Transfer learning: pretrain, finetune(train), predict(test) (e.g. word embeddings)

In-context learning: pretrain (LLM), predict(test, train)

Tom Brown et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems*. Vol. 33, pp. 1877–1901

Long Ouyang et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35, pp. 27730–27744

Noah Hollmann et al. (2025). "Accurate predictions on small data with a tabular foundation model". In: *Nature* 637.8045, pp. 319–326

Article

Accurate predictions on small data with a tabular foundation model

https://doi.org/10.1038/s41586-024-08328-6

Received: 17 May 2024

Accepted: 31 October 2024

Published online: 8 January 2025

Open access

Check for updates

Noah Hollmann[1,2,3,7] ✉, Samuel Müller[1,7] ✉, Lennart Purucker[1], Arjun Krishnakumar[1], Max Körfer[1], Shi Bin Hoo[1], Robin Tibor Schirrmeister[4,5] & Frank Hutter[1,3,6] ✉

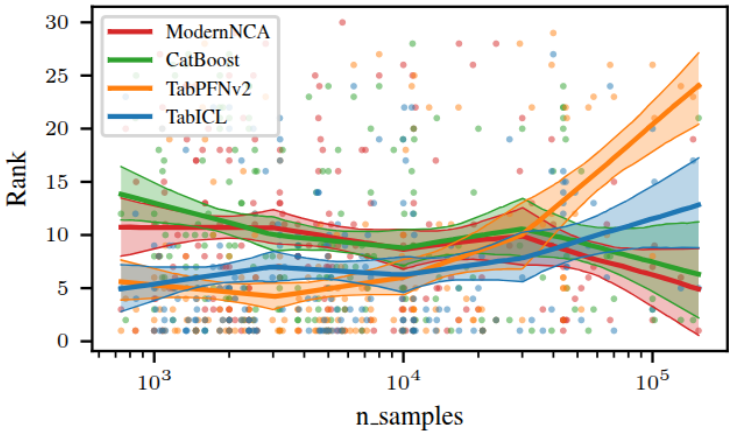
Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science[1,2]. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for various applications as diverse as biomedical risk models, drug discovery and materials science. Although deep learning has revolutionized learning from raw data and led to numerous high-profile success stories[3–5], gradient-boosted decision trees[6–9] have dominated tabular data for the past 20 years. Here we present the Tabular Prior-data Fitted Network (TabPFN), a tabular foundation model that outperforms all previous methods on datasets with up to 10,000 samples by a wide margin, using substantially less training time. In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting. As a generative transformer-based foundation model, this model also allows fine-tuning, data generation, density estimation and learning reusable embeddings. TabPFN is a learning algorithm that is itself learned across millions of synthetic datasets, demonstrating the power of this approach for algorithm development. By improving modelling abilities across diverse fields, TabPFN has the potential to accelerate scientific discovery and enhance important decision-making in various domains.

TabPFN is pretrained on 1 million synthetic datasets

- Pretrain a foundation model to "train and test" (very meta)
- Better than training from scratch (4 hours $\rightarrow$ 2.8 seconds)
- Directly estimate $p(y_{test} \mid x_{test}, D_{train})$ using in-context learning $\rightarrow$ strong baseline
- Limitations: up to 10000 samples, does not handle missing data very well

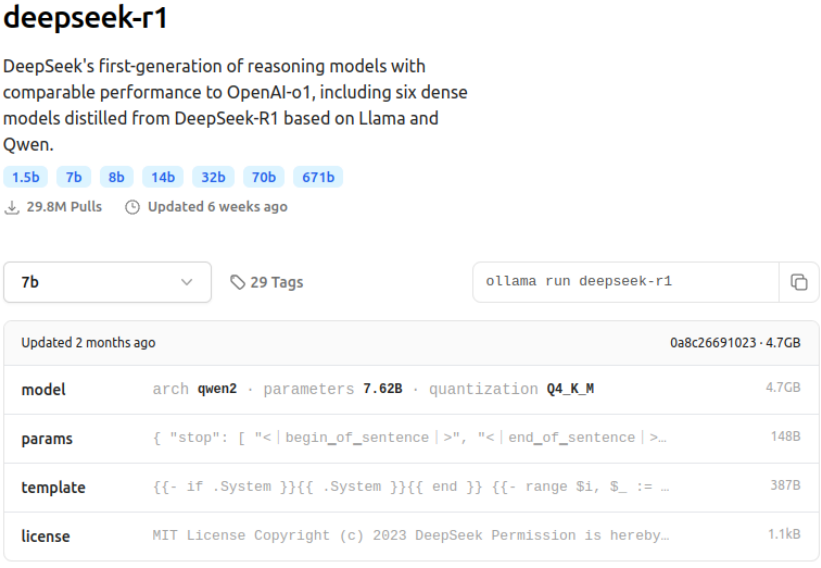
TabPFN is pretrained on 1 million synthetic datasets

- Pretrain a foundation model to "train and test" (very meta)
- Better than training from scratch (4 hours → 2.8 seconds)
- Directly estimate $p(y_{test} \mid x_{test}, D_{train})$ using in-context learning → strong baseline
- Limitations: up to 10000 samples, does not handle missing data very well



Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan (2025). "TabICL: A Tabular Foundation Model for In-Context Learning on Large Data". arXiv preprint arXiv:2502.05564

Open weights models: download & run them on your computer / phone



Speech recognition: Whisper

The smallest model is 75 MB (!) and takes 8 seconds to transcribe 30 seconds of video:

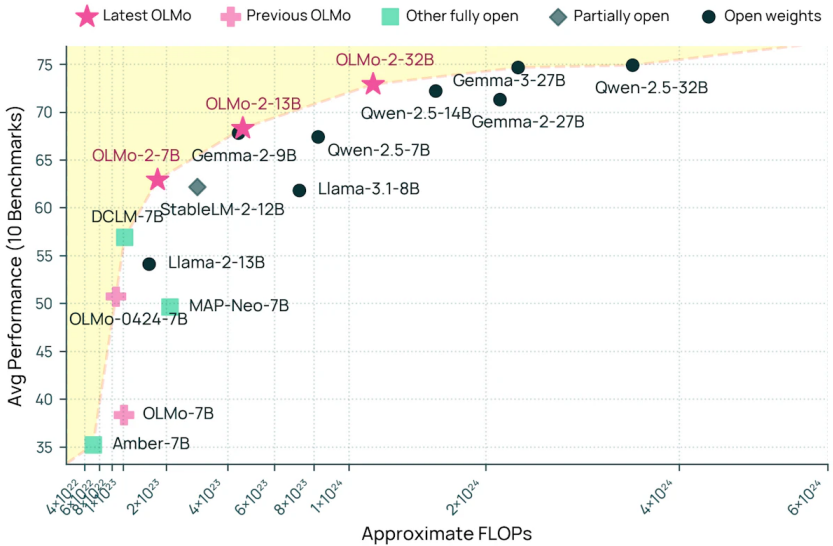
```
time ./main -m models/ggml-tiny.bin -l fr -f sample-show.wav

[00.000 --> 02.500]   - Ça va, merci. - C'est pas trois que je parle.
[02.500 --> 04.800]   Regarde ce que tu as fait, c'est de pourfouter de chose.
[04.800 --> 06.100]   Est-ce qu'il ne respire?
[06.100 --> 08.100]   - Euh, bah je crois. - Oui, je crois.
[08.100 --> 11.000]   - Ne restes pas planté là, tu vois bien qu'il a besoin d'un médecin.
[11.000 --> 13.000]   Il y a un centre Pokémon, non loin d'ici.
[13.000 --> 15.200]   Il faut que tu y ailles le plus vite possible.
[15.200 --> 16.700]   - C'est vrai et un centre.
[16.700 --> 20.600]   - Oui, pour Pokémon. - Est-ce que tu peux m'indiquer la direction?
[20.600 --> 21.700]   - Par là.
[21.700 --> 25.700]   - Oh non, ils arrivent!
[25.700 --> 28.900]   - Humé! Qu'est-ce que tu fais?
[28.900 --> 30.900]   Je comprends pas mis.
```

insanely-fast-whisper is (pretrained on 5M hours of data) optimized for GPU
and transcribes 150 minutes (2.5 hours) of audio in 98 seconds on a Nvidia A100 80 GB.
A compressed version (166M parameters) fits within 1.5 GB RAM.

Trade-offs between compression and performance

OLMo 2 32B: First fully open model to outperform GPT-3.5 and GPT-4o mini



Distillation and other cutting-costs tricks

Before: humans were annotating data $\rightarrow$ Cost: a lot

Later: GPT-4 is annotating data $\rightarrow$ Cost: still a lot :)

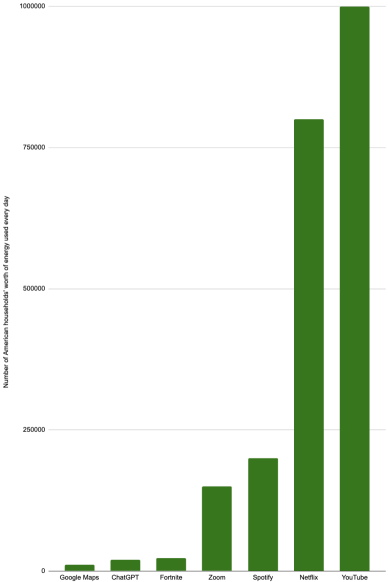
Now: smaller models are trained on GPT-4 answers

3 dimensions for improvement:

- Parallelize the search of an answer: sample many responses and score them
- Giving more time to compute an answer (o1): score partial promising answers ($\rightarrow$ not what DeepSeek is doing)
- Improve the base model (bigger)

DeepSeek had (and shared) many other tricks for reducing costs (45x improvement)

- Caching frequent requests
- 671B parameters but only 37B active in the VRAM
- Multi-head latent attention (MLA) reducing key-value cache
- Multi-token prediction



Maps ChatGPT Fortnite Zoom Spotify Netflix YouTube

Let's now talk about CO_2

"It's okay[a]" $\rightarrow$ "No it's not"

What does it replace?

If getting the answer from an LLM prevents me from watching a YouTube video: cool

Or instead of browsing many pages with trackers ($\sim$ 50–70% of carbon emissions[b]): cool

"500 billion[c] logged events per day on Netflix in 2016"

---

[a] https://andymasley.substack.com/p/individual-ai-use-is-not-bad-for

[b] https://marmelab.com/blog/2021/12/20/mesurons-lempreinte-carbone-des-plus-gros-sites-medias.html

[c] https://theconversation.com/que-sait-on-des-impacts-environnementaux-de-la-video-en-ligne-lexemple-de-netflix-229955

Pretraining

- 552 metric tons eq. CO_2
  (5x the lifetime emissions of an American car, 550 roundtrip flights NYC–SF)

Finetuning

- Gemma-2B-it took 1/7 emissions of GPT-3 to finetune
- QLoRA directly finetunes the 4-bit 65B quantized model
  (24 h on a single 48 GB GPU $\rightarrow$ reached 99.3% performance of GPT-4)

Inference

- Despite lower per-query energy use, inference dominates total emissions due to scale
- Google in 2022 reported 60% of ML energy use stems from inference
- ChatGPT had 1.7B visits in October 2023: within weeks or months, inference would surpass training costs

David Patterson et al. (2022). "The carbon footprint of machine learning training will plateau, then shrink". In: *Computer* 55.7, pp. 18–28
Sasha Luccioni, Yacine Jernite, and Emma Strubell (2024). "Power hungry processing: Watts driving the cost of ai deployment?" In: *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 85–99
Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). "QLoRA: Efficient finetuning of quantized LLMs". In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 10088–10115

Comparison to human labor

Digitalization and OCR (being improved by multimodal LLMs) helped reduce heavy transport environmental costs

Environmental impact for workers in the US:

- 53x less emissions for typical LLM (LLaMa-3-70B)
- 4400x less emissions for lightweight LLM (Gemma-2B-it)

For workers in India:

- 13x less emissions for typical LLM (LLaMa-3-70B)
- 1100x less emissions for lightweight LLM (Gemma-2B-it)

Shaolei Ren, Bill Tomlinson, Rebecca W Black, and Andrew W Torrance (2024). "Reconciling the contrasting narratives on the environmental impact of large language models". In: *Scientific Reports* 14.1, p. 26310

"An algorithm that optimizes flight costs makes people fly more which is 'bad' "

To me:

- It is yet to be proven that such an algorithm actually changes human behavior (so-called rebound effect is hard to estimate)
- It's the pricing that is 'bad', not the algorithm.
  So probably we should: enforce carbon tax everywhere, $cost = price + \lambda CO_2$

Paradox: if a LLM or AI helps reduce carbon emissions of a company, then it's worth using it

Estimating the carbon emissions of a LLM before training it (ICLR 2024)

LLM	T5	GPT3	GShard	Switch	XLM
reference	(Patterson et al., 2021)				(Wu et al., 2022)
developer	Google	OpenAI	Google	Google	Meta
type	dense	dense	MoE	MoE	dense
parameter # (B)	11	175	619	1500	0.55
base model param. # (B)	-	-	2.3	7.41	-
token # (B)	500	300	1K	2K	7K
CO_2eq/KWh	0.545	0.429	0.177	0.33	0.413
PUE	1.12	1.1	1.09	1.1	1.1
computing device	TPUv3	V100	TPUv3	TPUv3	V100
device TPD (W)	450	300	450	450	300
avg. system power (W)	310	330	288	245	342
peak TFLOPs/s	123	125	123	123	125
achieved TFLOPs/s	45.6	24.6	48	34.4	26.5
hardware efficiency	37%	19.7%	39%	28%	21.2%
device #	512	10K	1K	1K	512
total zettaFLOPs	40.5	314	13.3	82.2	23.9
training days	20	14.8	3.1	27	20.4
actual tCO_2eq	46.7	552.1	4.3	59.1	39
mlco2 predicted tCO_2eq	89.4	955.2	8.4	137.3	66.96
mlco2 Δ	+91.3%	+73%	+95.3%	+132%	+69%
LLMCarbon predicted tCO_2eq	45.66	553.87	4.46	63.9	37.6
LLMCarbon Δ	**−2.22%**	**+0.32%**	**+3.8%**	**+8.2%**	**−3.54%**

Ahmad Faiz et al. (n.d.). "LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models". In: *ICLR 2024*. URL: https://arxiv.org/pdf/2309.14393

Troll : “L’IA frugale n’existe pas, l’IA c’est de l’optimisation, donc ça dépend ce qu’on optimise”
En fait :

Les notions de frugalité et d’efficience doivent s’articuler autour des distinctions suivantes :

	Définition	Notions connexes	Raisonnement	Approche	Précisions
Efficience	Aptitude à optimiser les moyens alloués pour atteindre un résultat défini	Efficacité, optimisation	En relatif/par unité d’usage Le besoin prime : optimisation d’une solution jugée celle répondant le mieux au besoin	Recherche d’un optimum local ou d’un compromis sur un niveau de résultat fortement contraint	Prise en compte des effets de premier ordre pour les minimiser Prise en compte des parties prenantes de l’IA
Frugalité	Aptitude à se contenter d’un niveau de résultat jugé suffisant en redéfinissant les usages et les besoins	Sobriété (ou *Sufficiency* [10] en anglais)	En global La contrainte sur les ressources prime : recherche de la solution utilisant le moins de ressources possible et apportant une réponse satisfaisante au besoin	Recherche d’un optimum global ou d’un compromis large sur un niveau de résultat, ce qui nécessite d’élargir ou d’assouplir le besoin	Prise en compte des effets de premier ordre et de second ordre pour minimiser les impacts environnementaux négatifs Prise en compte de tous les acteurs au-delà des seules parties prenantes de l’IA

AFNOR (2024). *Référentiel général pour l’IA frugale - Mesurer et réduire l’impact environnemental de l’IA*. URL: https://www.afnor.org/actualites/referentiel-pour-mesurer-et-reduire-impact-environnemental-de-ia/

Take-home message

It's all about trade-offs computation / performance / memory / time / carbon / energy
Still, datacenters represent 1–2% of global carbon emissions
Need for transparency (cf. Altman vs. Luccioni)

Avoid doing the same thing multiple times
(cache, use pretrained instead of train from scratch, reproducibility, sovereignty paradox)

Avoid: usage peaks (everyone at the same time), mad race for incremental benefits
(everyone wants their own model, cf. many lines built from different companies serving the exact same routes in the 1800s)

Je n'aime pas trop "IA frugale" je suggère plutôt "optimisation optimisée"

Cela est bien dit, répondit Candide, mais il faut cultiver notre jardin
→ continuer à avoir notre impact de chercheur ou ingénieur en informatique

Merci à Benjamin Ninassi, Antoine Pietri, Michel Blockelet, Damien Sileo pour les discussions

Gaël Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker (2024). "Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI". arXiv preprint arXiv:2409.14160

- AFNOR (2024). *Référentiel général pour l'IA frugale - Mesurer et réduire l'impact environnemental de l'IA*. URL: https://www.afnor.org/actualites/referentiel-pour-mesurer-et-reduire-impact-environnemental-de-ia/.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems*. Vol. 33, pp. 1877–1901.
- Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). "QLoRA: Efficient finetuning of quantized LLMs". In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 10088–10115.
- Faiz, Ahmad, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang (n.d.). "LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models". In: *ICLR 2024*. URL: https://arxiv.org/pdf/2309.14393.

Hollmann, Noah, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter (2025). "Accurate predictions on small data with a tabular foundation model". In: *Nature* 637.8045, pp. 319–326.

Luccioni, Sasha, Yacine Jernite, and Emma Strubell (2024). "Power hungry processing: Watts driving the cost of ai deployment?" In: *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 85–99.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35, pp. 27730–27744.

Patterson, David, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean (2022). "The carbon footprint of machine learning training will plateau, then shrink". In: *Computer* 55.7, pp. 18–28.

Qu, Jingang, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan (2025). "TabICL: A Tabular Foundation Model for In-Context Learning on Large Data". arXiv preprint arXiv:2502.05564.


Ren, Shaolei, Bill Tomlinson, Rebecca W Black, and Andrew W Torrance (2024). "Reconciling the contrasting narratives on the environmental impact of large language models". In: *Scientific Reports* 14.1, p. 26310.


Varoquaux, Gaël, Alexandra Sasha Luccioni, and Meredith Whittaker (2024). "Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI". arXiv preprint arXiv:2409.14160.