Fine-Tuning Language Models for PHI Detection: An Energy Benchmark

Roberto Vergallo
Dept. of Information Science and Technology
Pegaso University
Naples, Italy
roberto.vergallo@unipegaso.it

Matteo Aprile
Dept. of Innovation Engineering
Salento University
Lecce, Italy
matteo.aprile@unisalento.it

Cosimo D'Oria
Dept. of Innovation Engineering
Salento University
Lecce, Italy
cosimo.doria@studenti.unisalento.it

Francesco Grassi
Dept. of Innovation Engineering
Salento University
Lecce, Italy
francesco.grassi@studenti.unisalento.it

Luca Mainetti
Dept. of Innovation Engineering
Salento University
Lecce, Italy
luca.mainetti@unisalento.it

***Abstract*—Smart cities increasingly rely on digital health infrastructures – shared electronic health records, telemedicine platforms, and territorial health data exchanges – whose sustainable operation depends on ICT choices made well before deployment. Protected Health Information (PHI) detection is a critical task in the clinical de-identification pipelines that enable such infrastructures, yet the energy cost of fine-tuning Language Models (LMs) for this purpose has received little attention, despite growing pressure on public health systems to adopt greener AI practices. We present an energy-aware benchmark in which five transformer-based models (110M–770M parameters) are fine-tuned on a synthetic PHI detection dataset under three validation strategies (Hold-out, K-Fold, and Stratified K-Fold), with energy consumption estimated via CodeCarbon. Results show that competitive accuracy (up to 99.47%) can be achieved with substantially lower energy expenditure: CodeT5-base with Hold-out reaches near-optimal performance at only 0.285 kWh, four to five times less than cross-validation configurations that yield no measurable accuracy gain. These findings offer practical guidance for municipalities and healthcare providers designing energy-aware AI pipelines for e-health services within smart city and community ecosystems.**

***Index Terms*—PHI detection, fine-tuning, language models, energy efficiency, Green AI, de-identification, healthcare NLP**

I. Introduction

Training and adapting language models carries a growing environmental cost. Training GPT-3 alone emitted approximately 502 metric tons of CO_2-equivalent[1], and data centers now account for 1–2% of global electricity consumption, with AI workloads representing around 15% of that figure[2]. Much of the Green AI literature has focused on inference-time optimization (quantization, pruning, distillation) while the fine-tuning phase, an essential step connecting general-purpose pre-trained models to domain-specific tasks, remains comparatively understudied from an energy perspective.

This gap is particularly relevant in privacy-sensitive healthcare applications embedded in smart cities digital infrastructures, where fine-tuning constitutes a structural requirement. In clinical systems, the automated detection of Protected Health Information (PHI) (i.e., any data that can identify a patient and is related to their health condition, as defined by HIPAA[3] and the GDPR[4]) is a prerequisite for regulatory compliance and privacy-preserving data sharing. Performing PHI detection at scale requires adapting general-purpose models to clinical text, and the validation strategy chosen during that adaptation determines how many full training cycles the pipeline executes, affecting energy consumption.

We address this gap with an empirical benchmark that asks: **(RQ1)** How does validation strategy choice influence the energy-accuracy trade-off? **(RQ2)** Do the accuracy gains from fine-tuning justify the additional energy required relative to the pre-trained baseline?

II. Background

A. Fine-Tuning and Validation Strategies

Fine-tuning adapts a pre-trained model with weights θ_0 to a target task by minimizing a supervised loss $\mathcal{L}(D_{\text{target}};\, \theta)$ initialized at θ_0. The three validation strategies compared here differ in the number of training cycles they imply. *Hold-out* splits the data once and trains the model a single time, making it the least energy-intensive option. *K-Fold* partitions the data into K folds and repeats training K times; following Kohavi [1], we use $K{=}5$. *Stratified K-Fold* applies the same logic while preserving class proportions in each fold, recommended for sensitive classification tasks [2].

B. Energy Monitoring with CodeCarbon

CodeCarbon [3] estimates energy consumption and CO_2-equivalent emissions by tracking CPU, GPU, and RAM power draw during code execution. Software-based estimation differs from hardware wattmeter measurements (Fischer et al. [4] report errors up to 40%) but consumption trends are consistent with hardware instruments, preserving the validity of relative comparisons across configurations [5].

III. Dataset and Experimental Setup

A. PHI Dataset

The dataset was built for the SHIELD project following the synthetic generation methodology of Savkin et al. [6], using LLaMA-8B. The pipeline produced approximately 50,000 English-language medical consultation request records, each paired with a binary label indicating PHI presence (Label = 1) or absence (Label = 0). PHI entities are derived from GDPR Article 9 and Italian Data Protection Authority Decision No. 55/2019[5], covering six categories: clinical conditions, medical reports, genetic

This work was partially supported by project SHIELD (CUP:B53C22003990006) within the research program PE "Security and Rights in the CyberSpace – SERICS" (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU.

[1]https://www.statista.com/statistics/1465353/total-co2-emission-of-ai-models/

[2]https://www.climateimpact.com/news-insights/insights/carbon-footprint-of-ai/

[3]https://www.hhs.gov/hipaa/

[4]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[5]https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9091942

TABLE I
BENCHMARK RESULTS. FT E = FINE-TUNING ENERGY (KWH); FT ACC = POST-FINE-TUNING ACCURACY; ΔACC = ACCURACY GAIN OVER BASELINE. K-FOLD AND STRATIFIED K-FOLD FT ACC SHOWN AS MEAN $\pm$ STD OVER 5 FOLDS.

Model	Strategy	FT E	FT Acc	ΔAcc
CodeT5-base	Hold-out	0.285	0.9947	+48.37%
	K-Fold	1.221	0.9942±.0004	+48.31%
	Strat-KF	1.145	0.9944±.0004	+48.32%
CodeT5-large	Hold-out	1.317	0.9935	+47.63%
	K-Fold	4.907	0.9935±.0009	+47.63%
	Strat-KF	4.903	0.9942±.0004	+47.70%
ELECTRA-base	Hold-out	0.172	0.9361	+44.54%
	K-Fold	0.691	0.8891±.1899	+39.84%
	Strat-KF	0.693	0.7958±.2340	+30.51%
LongCoder-base	Hold-out	0.909	0.9868	+49.61%
	K-Fold	3.692	0.9817±.0053	+49.10%
	Strat-KF	3.670	0.9852±.0057	+49.45%
BioClinicalBERT	Hold-out	0.247	0.9636	+47.20%
	K-Fold	0.985	0.9868±.0028	+49.52%
	Strat-KF	0.474	0.9882±.0028	+49.66%

data, fertility data, disability data, and addiction data. The class distribution is balanced (≈50/50).

B. Models

Five transformer-based models were selected from Hugging Face:

- **CodeT5-base** (220M, GP, Salesforce, 2022)
- **CodeT5-large** (770M, GP, Salesforce, 2023)
- **ELECTRA-base** (110M, GP, Google, 2020)
- **LongCoder-base** (150M, GP, Guo et al., 2023)
- **BioClinicalBERT** (110M, DS, Alsentzer et al., 2019)

We chose widely used (sort by "Most downloads") and easily accessible (i.e. open weights) transformer models (tasks: "Fill-Mask", "Text Classification", "Text Generation", "Feature Extraction", or direct lookup of known models). All models have a moderate number of parameters (up to $1B$), making them suitable for reproduction by both the industry and research community. The set includes two distinct categories: general-purpose models (75%) and domain-specific models (25%).

C. Experimental Pipeline

Each model was first evaluated on the pre-trained baseline (no fine-tuning) to establish reference accuracy and energy values. Fine-tuning was then performed under each of the three validation strategies, keeping hyperparameters constant across all 15 configurations. Energy consumption was tracked with CodeCarbon for both the fine-tuning phase (including post-training inference on the held-out test set) and the baseline.

IV. RESULTS AND DISCUSSION

Table I summarizes all 15 experimental conditions.

A. RQ1 – Validation Strategy and Energy-Accuracy Trade-off

For CodeT5-base, CodeT5-large, LongCoder-base, and BioClinicalBERT, the choice of validation strategy produces no appreciable accuracy variation (max difference $<0.51\%$). Cross-validation, however, requires three to five times more energy than Hold-out due to the repetition of training across five folds. The energy-accuracy relationship is therefore clearly unbalanced for these models: higher energy expenditure does not translate into measurable performance gains.

ELECTRA exhibits a markedly different behavior. Hold-out yields 93.61% accuracy, but K-Fold and Stratified K-Fold expose deep instability (0.8891 ± 0.1899 and 0.7958 ± 0.2340, respectively). This behavior is not caused by cross-validation but *revealed* by it: the Hold-out result is partly attributable to a favorable data partition, while repeated splits expose the model's actual sensitivity to training set composition.

B. RQ2 – Does Fine-Tuning Justify Its Energy Cost?

Across all 15 configurations, pre-trained models score near 50%, consistent with random classification on a balanced dataset. After fine-tuning, all models reach 79.58–99.47%, with a positive ΔAcc in every configuration (+30.51% to +49.66%). Without task-specific adaptation, none of the models is usable for PHI detection.

The most energy-efficient configuration (CodeT5-base with Hold-out at 0.285 kWh) achieves 99.47% accuracy. The most expensive (CodeT5-large with K-Fold at 4.907 kWh) achieves 99.35%: essentially the same accuracy at $17\times$ the energy cost. Furthermore, CodeT5-large never outperforms CodeT5-base despite having $3.5\times$ more parameters and consuming 4–$5\times$ more energy, confirming that larger is not always better for specialized tasks [7].

V. CONCLUSIONS AND FUTURE WORK

This paper presented an energy-aware benchmark for fine-tuning language models on PHI detection, evaluated across five architectures and three validation strategies. Three key findings emerge: (1) fine-tuning is indispensable; (2) for stable models, Hold-out is the most energy-efficient strategy; (3) model size does not determine performance in specialized tasks.

Future work will integrate parameter-efficient approaches such as LoRA, extend the benchmark to real clinical datasets, and analyze post-training quantization.

REFERENCES

[1] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*. Morgan Kaufmann, 1995, pp. 1137–1143.

[2] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, "Fine-Tuning large language models for specialized use cases," *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 1, p. 100184, 2025.

[3] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, and S. Luccioni, "CodeCarbon: Estimate and track carbon emissions from machine learning computing," *Journal of Machine Learning Research*, vol. 24, no. 130, pp. 1–24, 2023.

[4] R. Fischer, "Ground-truthing AI energy consumption: Validating CodeCarbon against external measurements," *arXiv preprint arXiv:2509.22092*, 2025.

[5] S. Aquino-Brítez, P. García-Sánchez, A. Ortiz, and D. Aquino-Brítez, "Towards an energy consumption index for deep learning models: A comparative analysis of architectures, GPUs, and measurement tools," *Sensors*, vol. 25, no. 3, p. 846, 2025.

[6] M. Savkin, T. Ionov, and V. Konovalov, "SPY: Enhancing privacy with synthetic PII detection dataset," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*. Albuquerque, USA: Association for Computational Linguistics, April 2025, pp. 236–246.

[7] T. Yarally, L. Cruz, D. Feitosa, J. Sallou, and A. van Deursen, "Uncovering energy-efficient practices in deep learning training: Preliminary steps towards Green AI," in *Proceedings of the 2023 IEEE/ACM 2nd International Conference on AI Engineering — Software Engineering for AI (CAIN)*. Melbourne, Australia: IEEE, 2023, pp. 25–36.