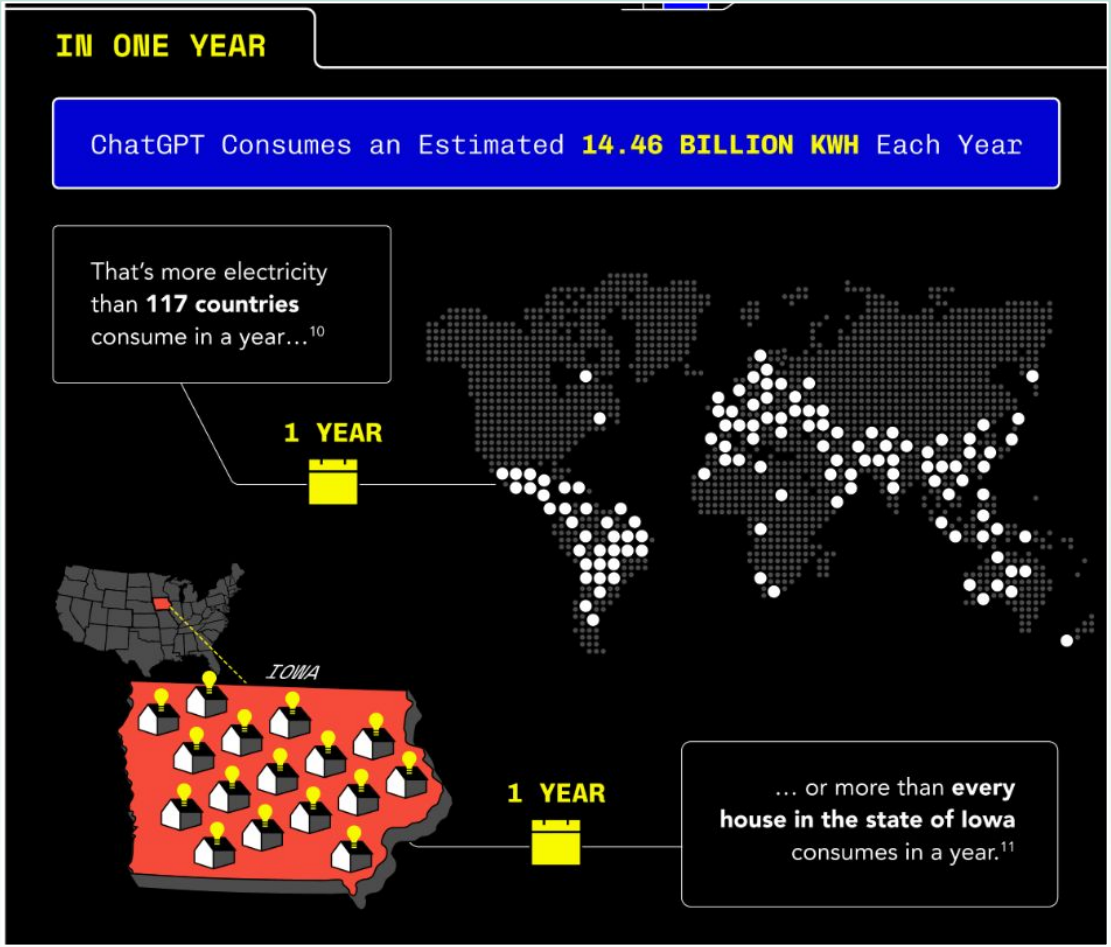
Don’t Look Up: Evaluating the Tradeoff between Accuracy and Sustainability of LLMs for Text Analysis.

Sean Palicki[1], Isaac Bravo[1], Clint Claessen[2]

1. PhD Student, Technical University of Munich
2. PhD Student, University of Basel

Background and Overview

- Large Language Models (LLM), a type of Generative AI (GenAI), perform well across diverse tasks (Grattafiori et al., 2024).
- Benchmarked/trained for data analysis, mathematics and coding (Livebench, 2025; Jiang et al., 2024).
- Rapidly becoming standard tools in social science text analysis (Wuttke et al., 2025; Rodman, 2024; Pipal et al., 2024; Wang, 2023).
- Critical perspectives: availability, validity, privacy, bias & discrimination (Bisbee et al., 2024; Laurer et al., 2024; Häffner et al., 2023; Baden et al., 2022; Lauscher et al., 2021; Bender et al., 2021; Benjamin, 2019; Noble, 2018).



(Business Energy UK, 2025; Washington Post, 2024)

Dr. Sasha Luccioni • 1st
AI & Climate Lead @ Hugging Face, AI & Social Justice Chair (Abeo…
5d •

The "3 Wh per ChatGPT query" estimate is back and haunting my social media feeds! 🤦‍♀️

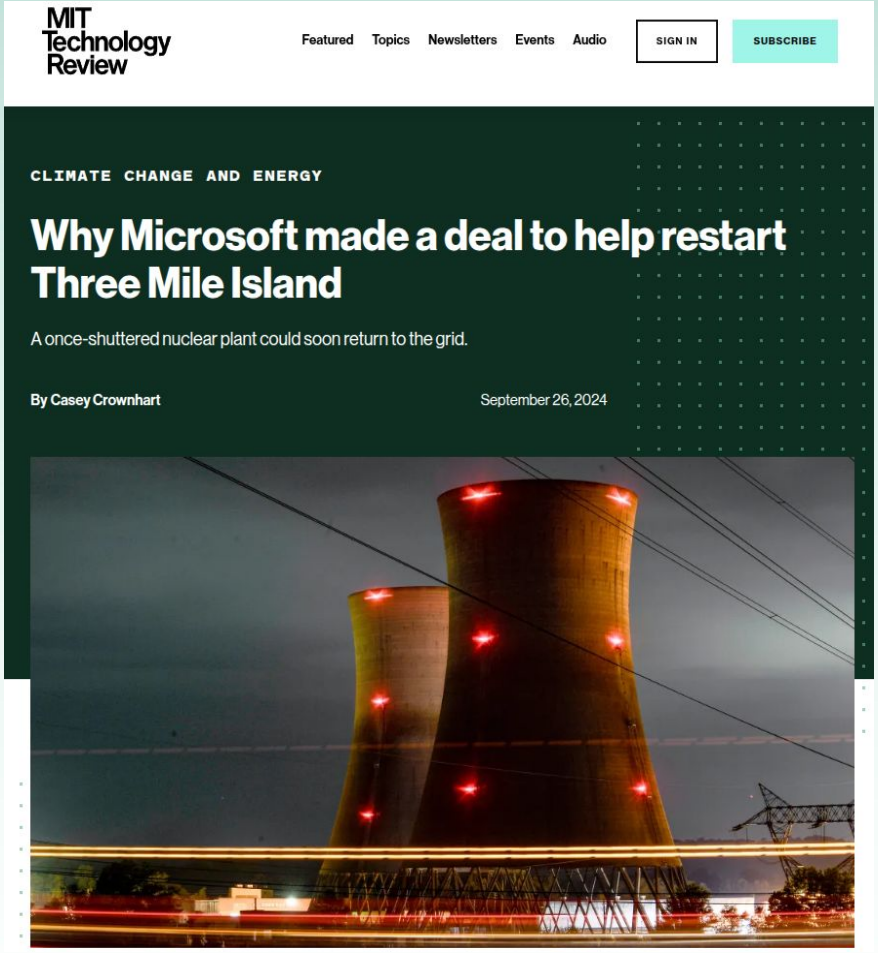
First of all -- that number is a back of the envelope calculation **based on little more than vibes and guesswork**. We don't know much about the model underpinning ChatGPT, how big it is, what hardware it's running on, nor how much energy it's using. 💡

Second -- even if it is true, or close to being true, the **cumulative amount of generative AI we are using is putting strain on energy grids worldwide**. We are switching out technologies that worked well for decades to new, shiny ones that use orders of magnitude more energy, and that comes with a cost. It's not just a matter of asking ChatGPT a question or 2, but also using genAI for searching the web, drafting emails, and generating images and video.🔌

So please don't listen to people who are loudly proclaiming that this "isn't an issue". We don't know the answer to that question -- and actually, there is no single answer to that question, since every gram of CO2 and Wh of energy matters in the fight against climate change. 🌎

The Environmental Costs of GenAI

- Watt Hour (Wh) = Power consumption (W) over time (h).
- ChatGPT query est. 6-10 times more energy than Web search (0.3 Wh vs. 2.9 Wh) (Luccioni, Trevelin & Mitchell, 2024).
- Meta Llama 3.1 405B Training = Manufacture 10,958 BMWs (21,588 MWh) (Li et al., 2024; Grattafiori et al., 2024).
- Water use, est. 0.18 to 1.1L per kWh of energy (Li et al., 2023; Microsoft, 2022).
- LLM Training < Inference (Mehta, 2024; Jian et al., 2024).
- Big Tech not meeting net-zero goals due to AI race (Bloomberg, 2024).



Are LLMs Worth It?

- Power generation is the primary source of CO_2 emissions (IEA, 2024).
- AI lifecycle: training, deployment & components (Luccioni et al., 2022; Strubell et al., 2020).
- Data centers up to 10% of US energy use by 2030, ~40% renewables (Washington Post, 2024).
- AI, climate justice & migration (Boas et al., 2022; Cripps, 2022; Crawford, 2021).

Research Questions

RQ1: How much energy is consumed when using LLMs for text analysis tasks compared to using lightweight alternative methods?

RQ2: What is the optimum balance between LLM size and task performance when completing text analysis tasks?

Methodology: Comparing LLMs to Non-LLM Methods

Tasks | Models | Metrics

Text Analysis:

1. Sentiment Analysis
2. Multi-Class Classification
3. Named-Entity-Recognition

*All tasks run 3 times on the same machine (24 GB GPU) using Human Labelled data.

Tasks

Text Analysis:

1. Sentiment Analysis
2. Multi-Class Classification
3. Named-Entity-Recognition

*All tasks run 3 times on the same machine (24 GB GPU) using Human Labelled data.

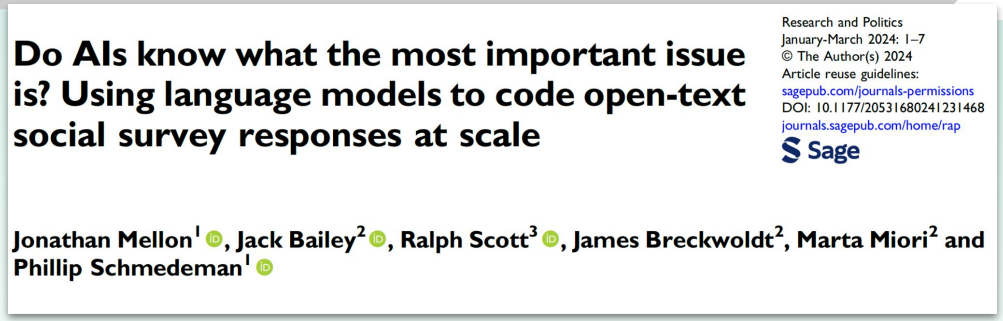
COMMUNICATION METHODS AND MEASURES
https://doi.org/10.1080/19312458.2024.2383453

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

JST and rJST: joint estimation of sentiment and topics in textual data using a semi-supervised approach

Christian Pipal[a], Martijn Schoonvelde[b], Gijs Schumacher[c], and Max Boiten[d]



Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale

Research and Politics
January-March 2024: 1–7
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20531680241231468
journals.sagepub.com/home/rap
Sage

Jonathan Mellon[1], Jack Bailey[2], Ralph Scott[3], James Breckwoldt[2], Marta Miori[2] and Phillip Schmedeman[1]

COMMUNICATION METHODS AND MEASURES
2024, VOL. 18, NO. 4, 371–389
https://doi.org/10.1080/19312458.2024.2324789

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Automatically Finding Actors in Texts: A Performance Review of Multilingual Named Entity Recognition Tools

Paul Balluff, Hajo G. Boomgaarden, and Annie Waldherr

Methodology: Comparing LLMs to Non-LLM Methods

Tasks | Models | Metrics

Text Analysis:

1. Sentiment Analysis
2. Multi-Class Classification
3. Named-Entity-Recognition

*All tasks run 3 times on the same machine (24 GB GPU) using Human Labelled data.

LLMs (Quantized, Local):

1. Mistral Nemo - 12B
2. Gemma 3 - 24B
3. Mistral Small - 27B
4. Deepseek R1- 32B

Non-LLMs:

- Dictionaries
- SVM
- NER Libraries
- BERT (Pretrained/Fine-Tuned)

MISTRAL AI_

Gemma 3

deepseek

Methodology: Comparing LLMs to Non-LLM Methods

Tasks

Text Analysis:

1. Sentiment Analysis
2. Multi-Class Classification
3. Named-Entity-Recognition

*All tasks run 3 times on the same machine (24 GB GPU) using Human Labelled data.

Models

LLMs (Quantized, Local):

1. Mistral Nemo - 12B
2. Gemma 3 - 24B
3. Mistral Small - 27B
4. Deepseek R1- 32B

Non-LLMs:

- Dictionaries
- SVM
- NER Libraries
- BERT (Pretrained/Fine-Tuned)

Metrics

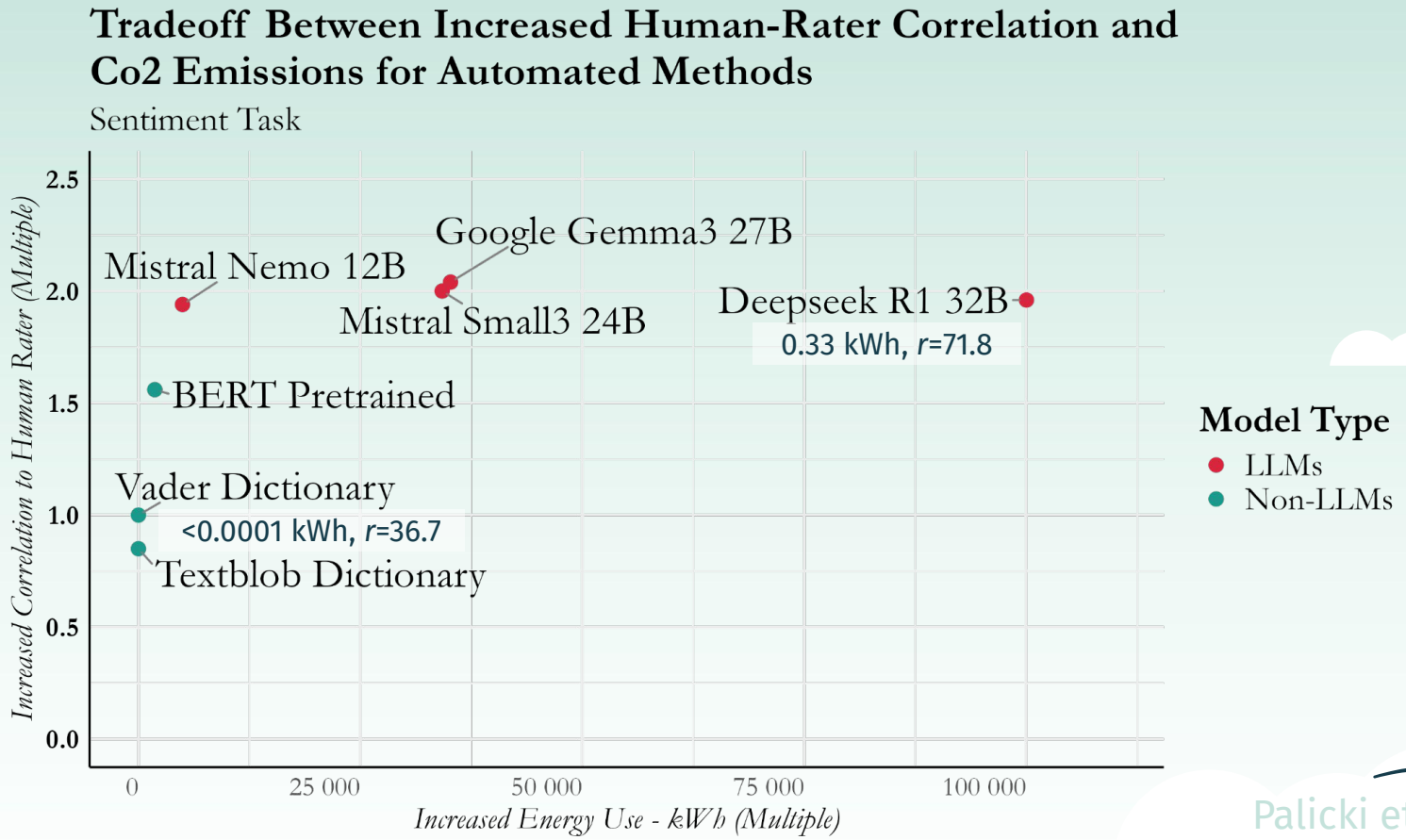
Human-Rater Agreement:
Pearson Correlation, Percent Agreement

Model Effectiveness:
F1, Precision, Recall, Duration

Sustainability:
kWh Energy Use, **CO_2** Emissions

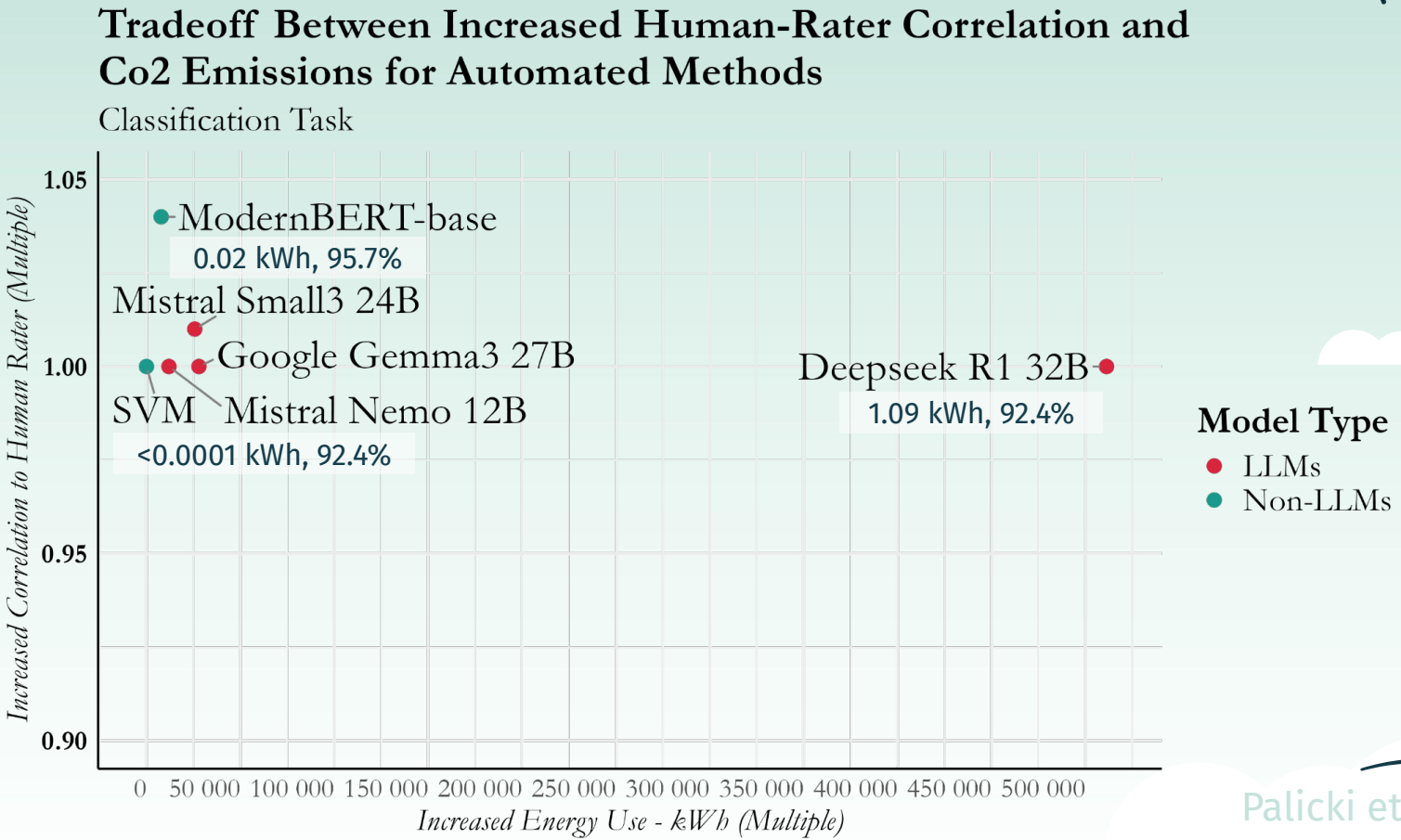
CODE CARBON

Sentiment: LLMs Useful but Costly



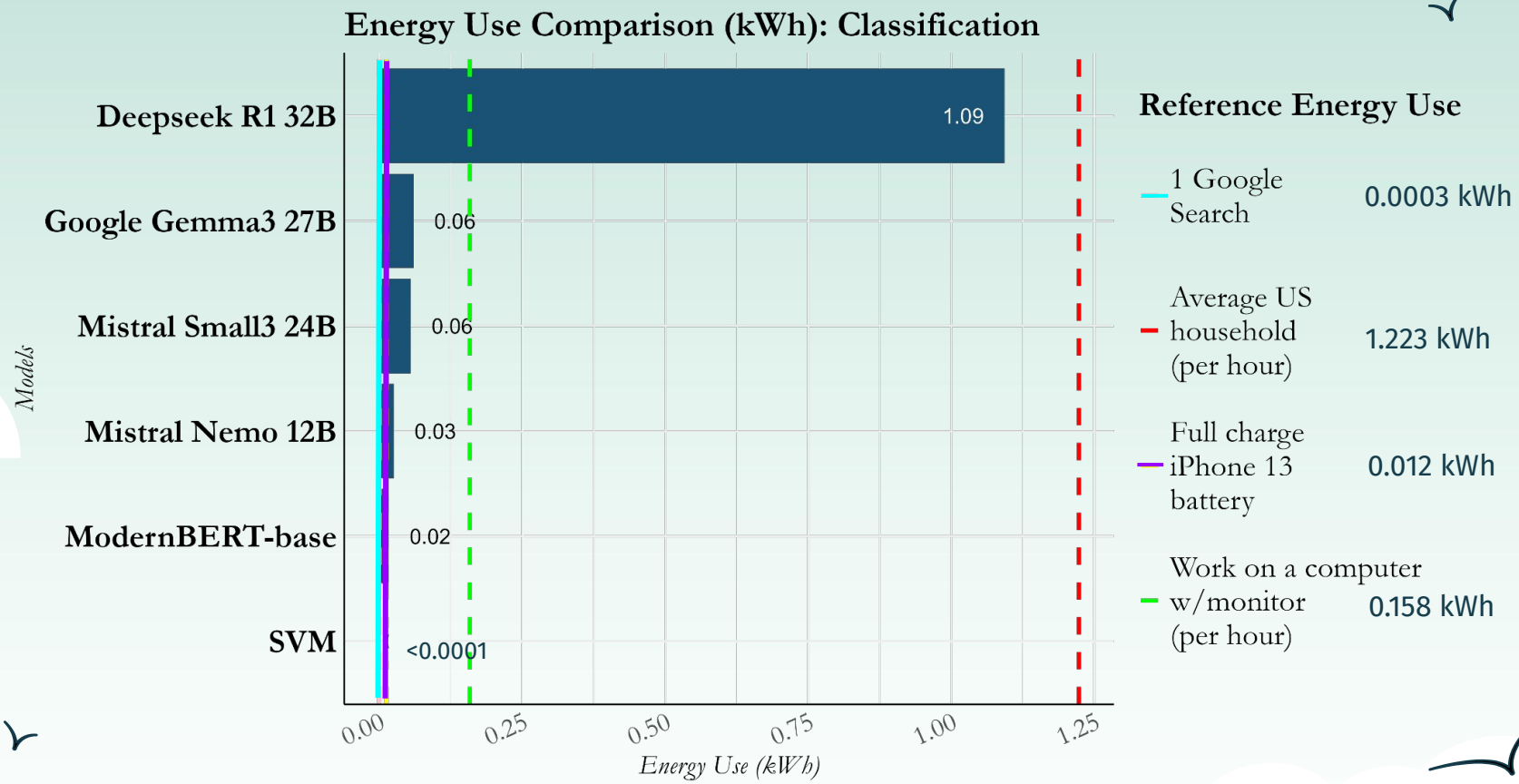
Palicki et al. 2025

Classification: Supervised Models Win



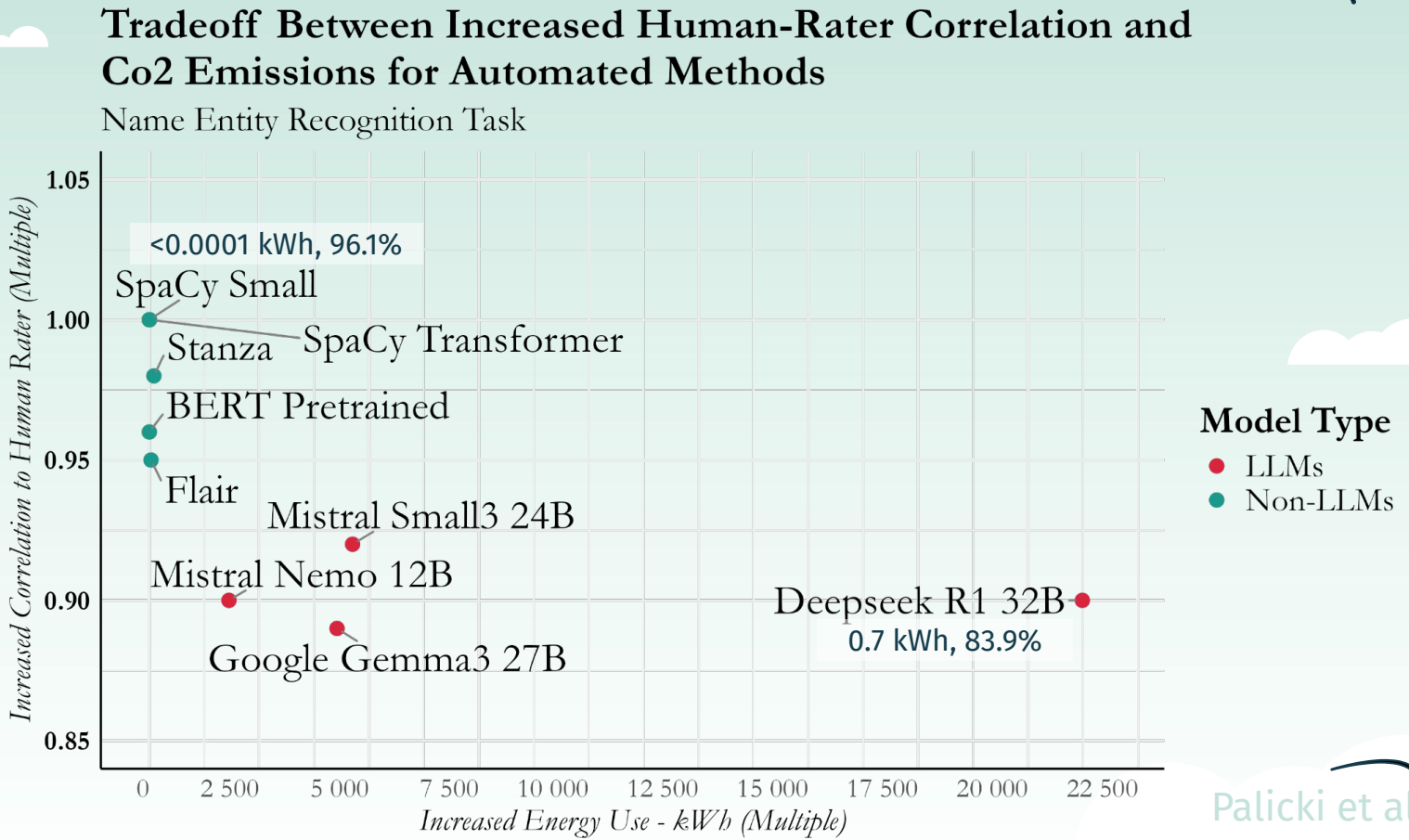
Palicki et al. 2025

Classification: Energy Hog



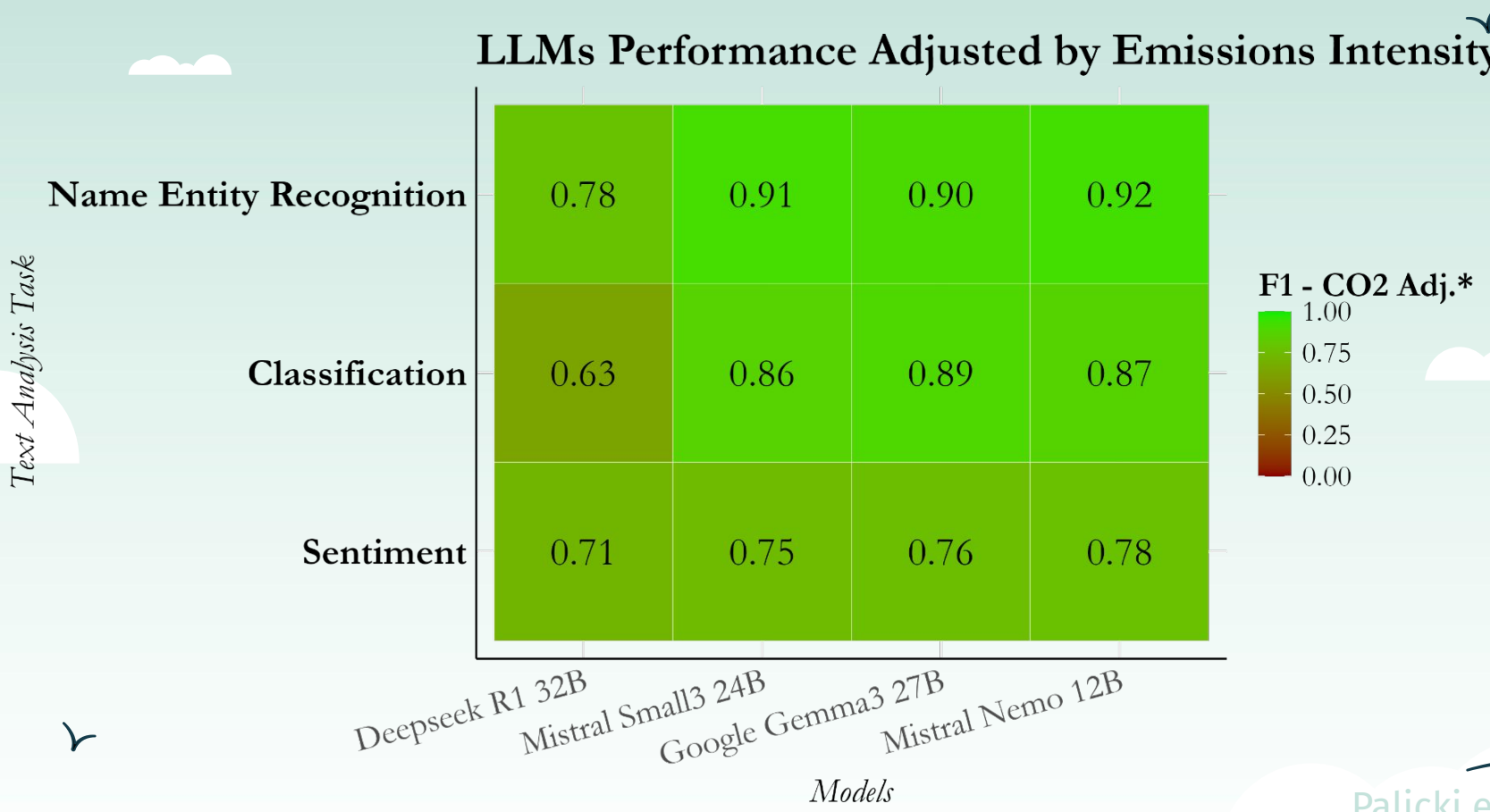
Palicki et al. 2025

NER: Hard for (Local) LLMs



Palicki et al. 2025

Co2 Emissions Adjusted F1



Palicki et al. 2025

Discussion and Implications

- Small LLMs like Mistral-Nemo = Efficient general performers
- LLM size ≠ Task performance
- “One-size-fits-all” LLM use is **not justified**
- Specialized Statistical or ML Model > Pre-Trained Local LLM
- Total Project CO_2 Footprint: 13.42 kWh = ~5.7 lbs (~2.5 kg) of coal burned

Limitations and Future Research

Limitation:

- Limited to models fitting 24 GB VRAM (consumer hardware)

Next:

- Multilingual and Multimodal tasks
- Fine-tune LLMs for specific text-as-data needs
- Environmental reporting as part of evaluation, see CodeCarbon(Py) and emissionsTrackerR(R)

Thank You 🤗

Contact: sean.palicki@tum.de

TO USE OUR RESULTS DASHBOARD AND EMISSIONS TRACKING TOOLS (@IsaacBravo):

https://linktr.ee/comptext25

Appendix

Full Results Tables

Table 1: Comparisons of LLM Size using 4-Bit Quantization

Name	**Provider**	**Model Size Params (Billions)**	**Model Size VRAM (GB)**	**Temperature Ollama**
Mistral-Nemo	Mistral, NVIDIA	12B	7.1 GB	0.3
Gemma 3	Google	27B	17 GB	0.1
Mistral Small 3	Mistral	24B	19 GB	0.15
Deepseek-R1	Deepseek	32B	19 GB	0.6

Table 2: Sentiment Task Results for 200 Parliamentary Speeches.

Model	**Pearson Correlation**	**F1**	**Energy (kWh)**	**Energy (kWh/Query)**	**Time (Minutes)**
Google Gemma3 27B	74.8	63.9	0.11	0.0005	13.9
Mistral Small3 24B	73.2	61.6	0.11	0.0005	13.5
Deepseek R1 32B	71.8	61.3	0.33	0.001	39.3
Mistral Nemo 12B	71.2	63.7	0.02	≤ 0.0001	3.2
BERT Pretrained	57.2	45.2	0.006	≤ 0.0001	2.2
Vader Dictionary	36.7	26.5	≤ 0.0001	≤ 0.0001	0.001
Textblob Dictionary	31.1	26.0	≤ 0.0001	≤ 0.0001	0.001

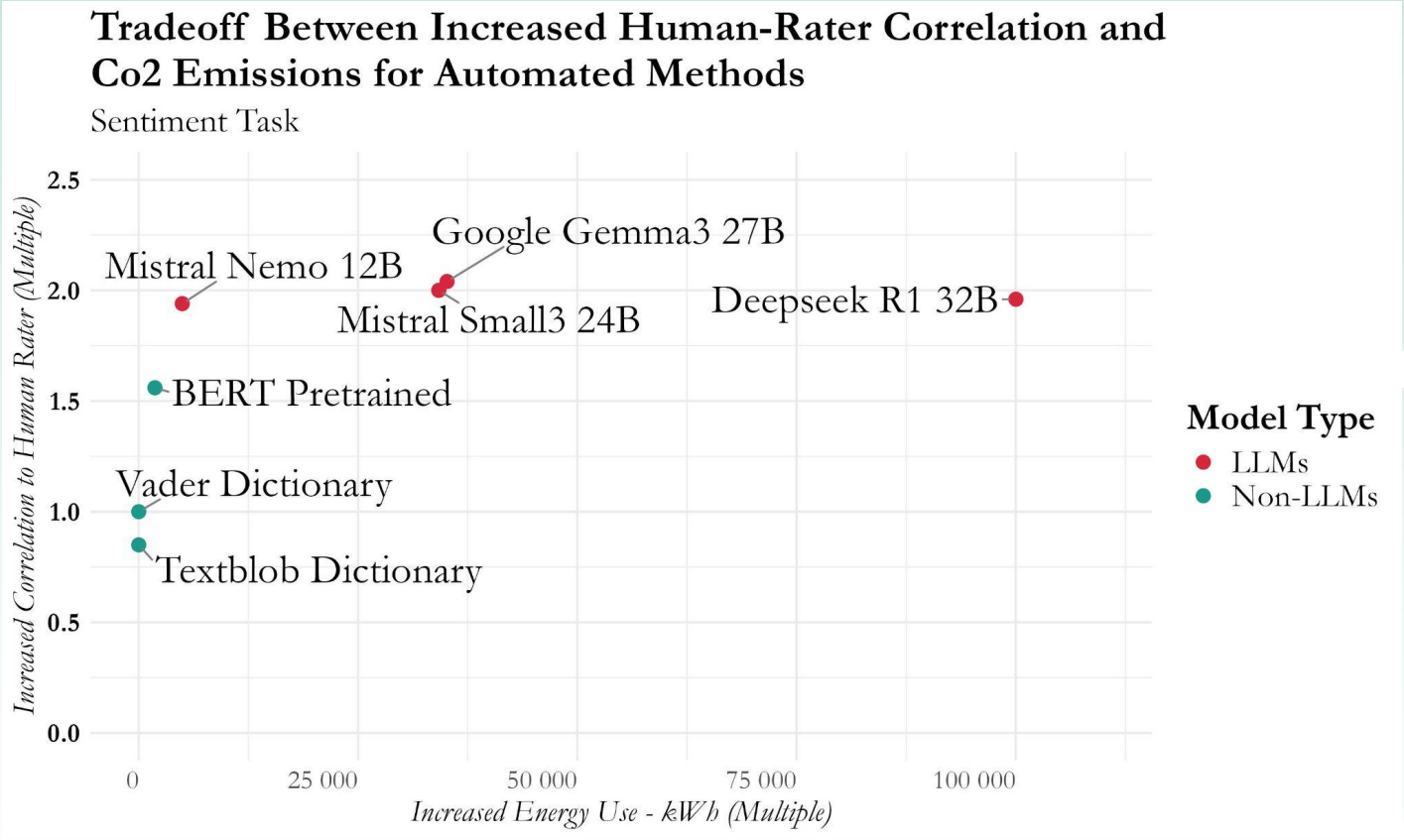
Table 3: Classification Task Results for 1,000 Open-Text BESP Responses.

Model	Agreement (%)	F1	Energy (kWh)	Energy (kWh/Query)	Time (Minutes)
ModernBERT-base	95.7	88.7	0.02	≤ 0.0001	2.5
Mistral Small3 24B	93.4	77.6	0.05	≤ 0.0001	6.6
Google Gemma3 27B	92.8	81.4	0.06	≤ 0.0001	7.6
Deepseek R1 32B	92.4	76.9	1.09	0.00109	128.1
SVM	92.4	89.4	≤ 0.0001	≤ 0.0001	0.002
Mistral Nemo 12B	92.3	77.1	0.03	≤ 0.0001	3.5

Table 4: Named Entity Recognition Task Results for 500 OntoNotes Sentences.

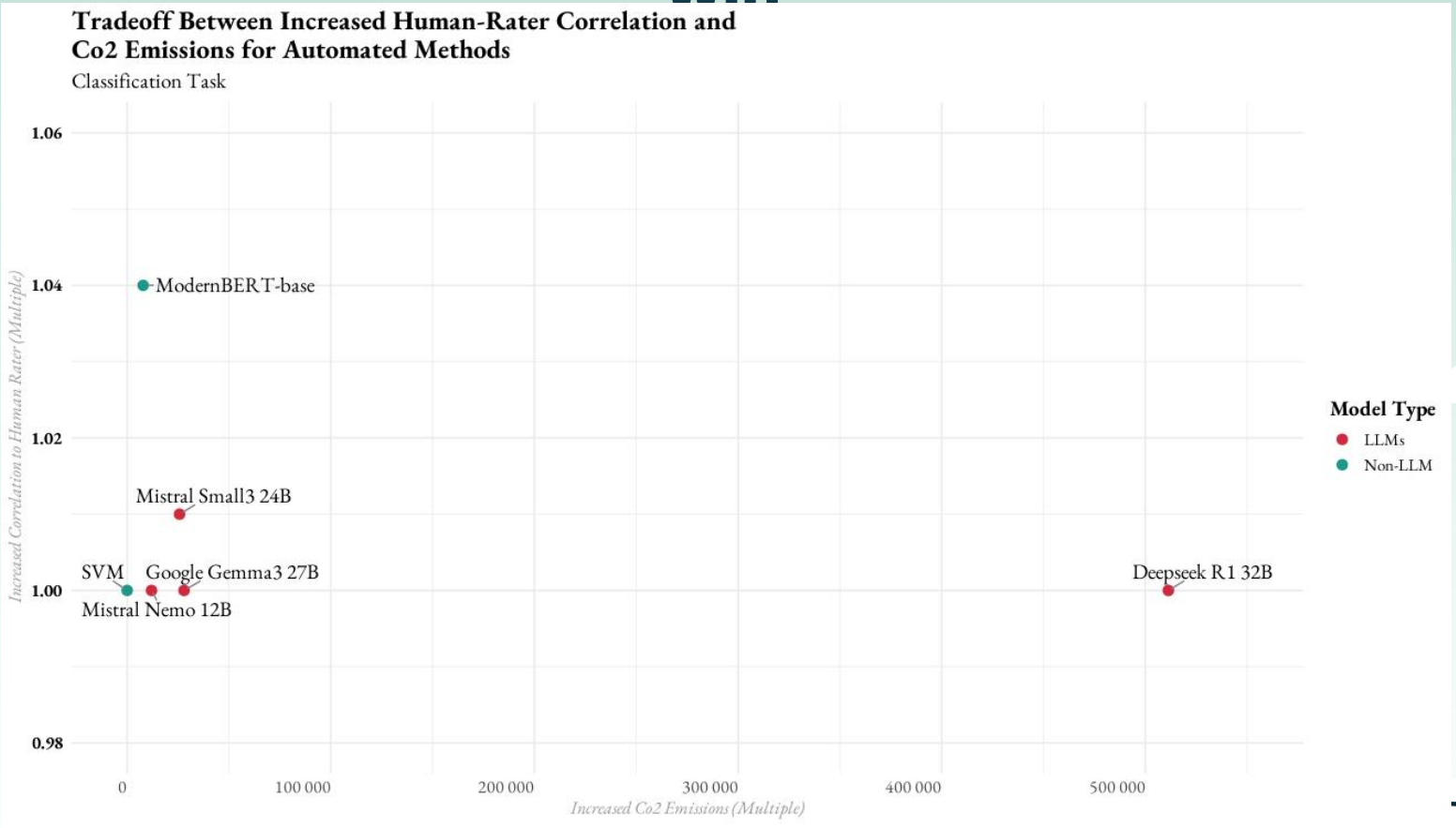
Model	Agreement (%)	F1	Energy (kWh)	Energy (kWh/Query)	Time (Minutes)
Stanza	96.5	93.4	0.003	≤ 0.0001	0.7
SpaCy Small	96.1	91.6	≤ 0.0001	≤ 0.0001	0.02
SpaCy Transformer	95.8	91.8	0.0002	≤ 0.0001	0.2
BERT Pretrained	91.7	85.1	≤ 0.0001	≤ 0.0001	0.02
Flair	89.0	81.0	0.001	≤ 0.0001	0.2
Mistral Small3 24B	87.0	78.6	0.2	0.0003	18.3
Deepseek R1 32B	83.9	68.8	0.7	0.001	82.03
Google Gemma3 27B	83.6	70.7	0.1	0.0003	16.6
Mistral Nemo 12B	83.2	66.5	0.06	≤ 0.0001	7.9
NLTK	53.6	40.2	0.001	≤ 0.0001	0.8

Sentiment Analysis: LLMs Useful, for a Cost



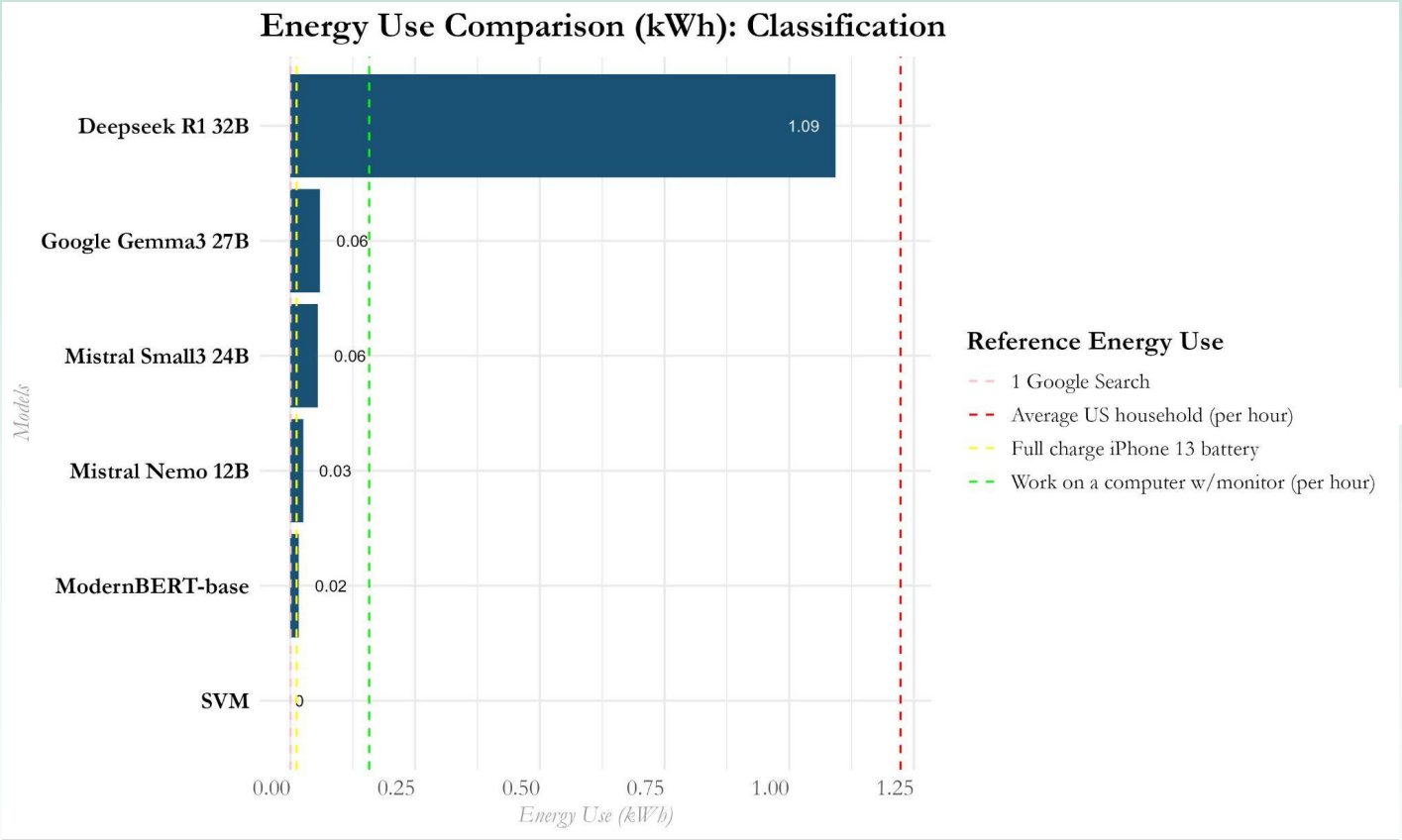
Palicki et al. 2025

Classification: Trained Classification Models Win



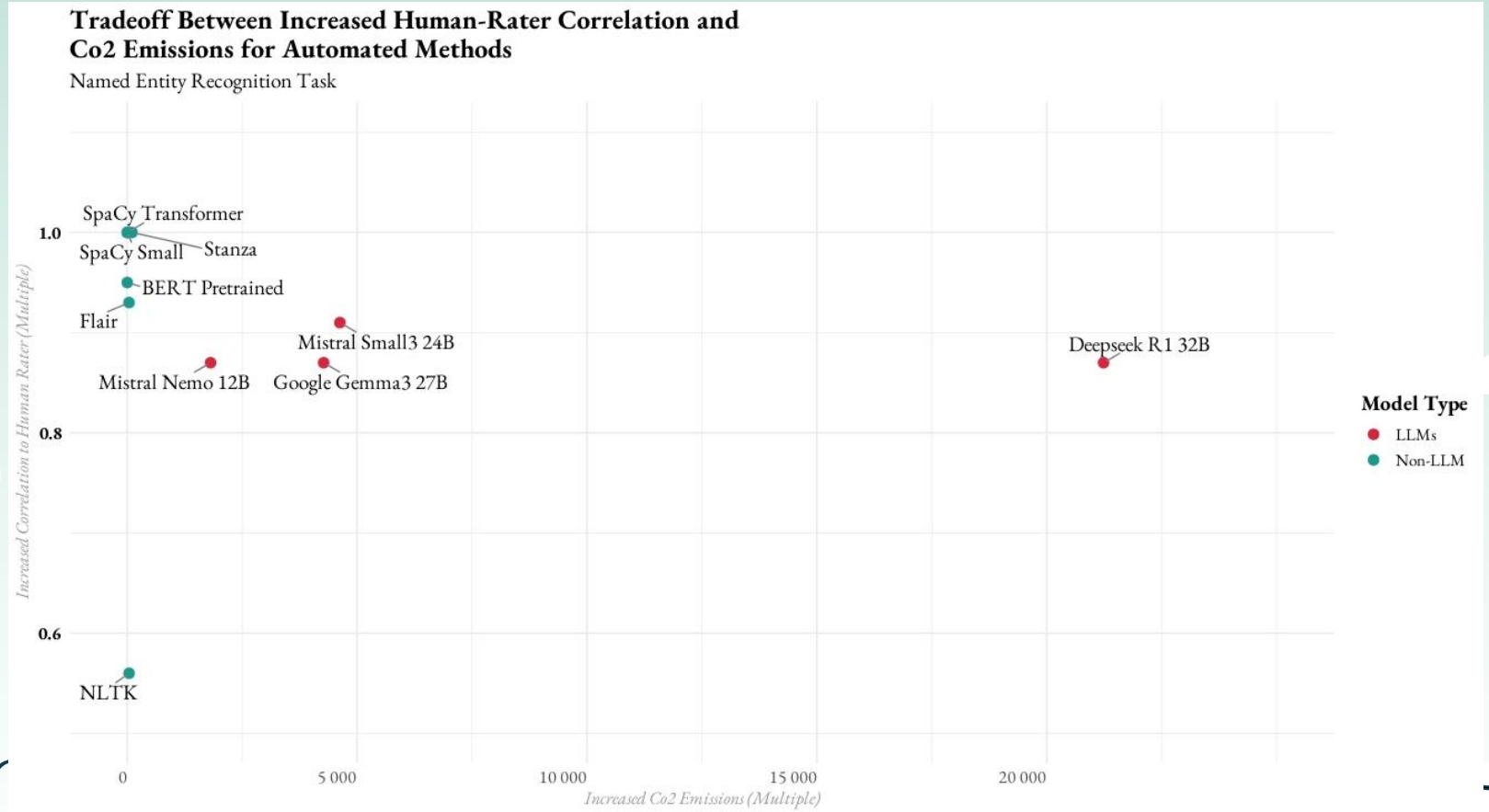
Palicki et al. 2025

Classification: Energy Hog



Palicki et al. 2025

NER: Very Hard for (Local) LLMs



Palicki et al. 2025