Where do LLMs fit in NLP pipelines?

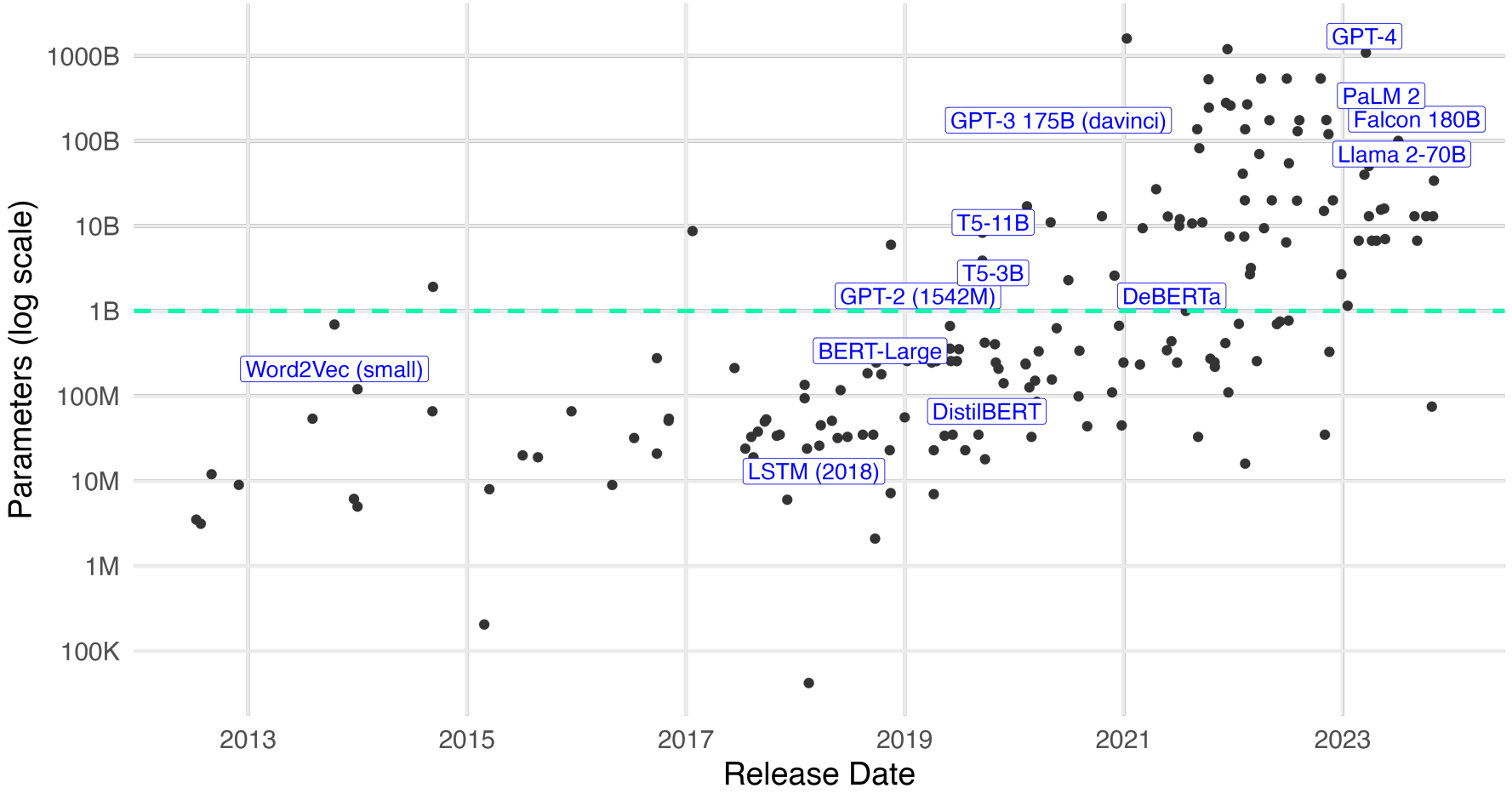
Paul Simmering and Paavo Huoviala

General Online Research 2024

Where do LLMs fit in NLP Pipelines?

1. Extractive NLP with LLMs
2. Benchmarks
3. Case studies
4. Finetuning and distillation
5. Discussion

What counts as a “large” language model?



Data source: Epoch (2023) via Our World in Data

Extractive vs. Generative NLP

Extractive NLP

- Text classification
- Named entity recognition
- Sentiment analysis

Generative NLP

- Question answering
- Creative writing
- Agents

Lists are not exhaustive

Is multi task prompting all you need?

Text

Dr. James Smith from the University of Greene has an exciting invention. The method 'Coffee Cup Reading' is a game-changer. This method forecasts consumer behavior based on patterns left in empty coffee cups at BrewHouse shops.

Prompt

Extract as JSON:
- language
- topic (one term)
- sentiment
- named entities

Response by GPT-4

```
{
  "language": "English",
  "topic": "Market Research",
  "sentiment": "Positive",
  "named_entities": {
    "persons": ["Dr. James Smith"],
    "organizations": [
            "University of Greene”,
            "BrewHouse shops”
    ],
    "other": ["Coffee Cup Reading"]
  }
}
```

text-to-text is inefficient for extractive NLP

Output layer of text classifier

P(Class 1)

P(Class 2)

P(Class 3)

- Directly read probabilities
- Straightforward loss function (cross entropy)

Output layer of GPT model

P(Token 1)

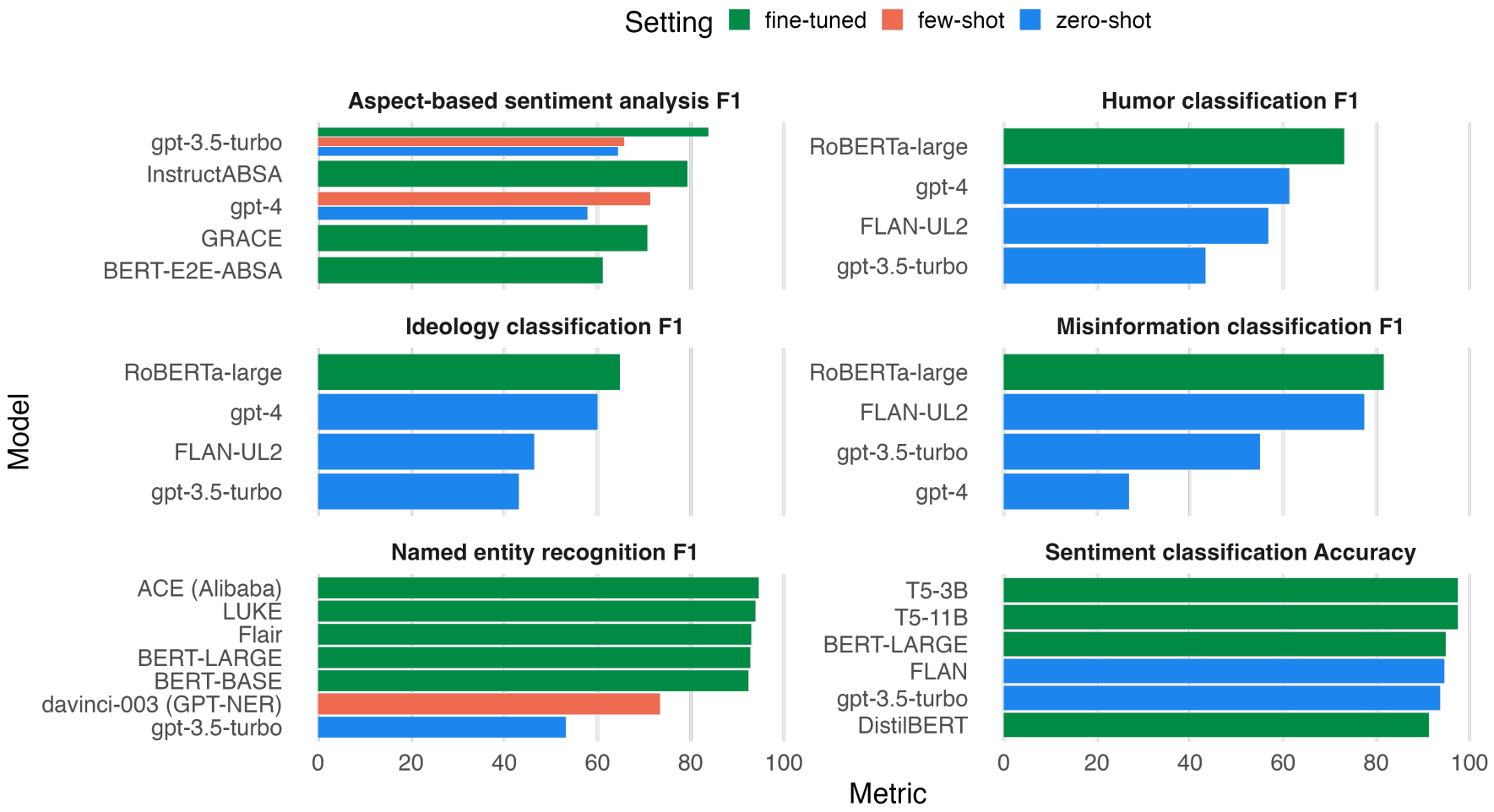
P(Token 2)

⋮

P(Token k)

- Need to teach output format
- Model training loss function is not the classification loss function

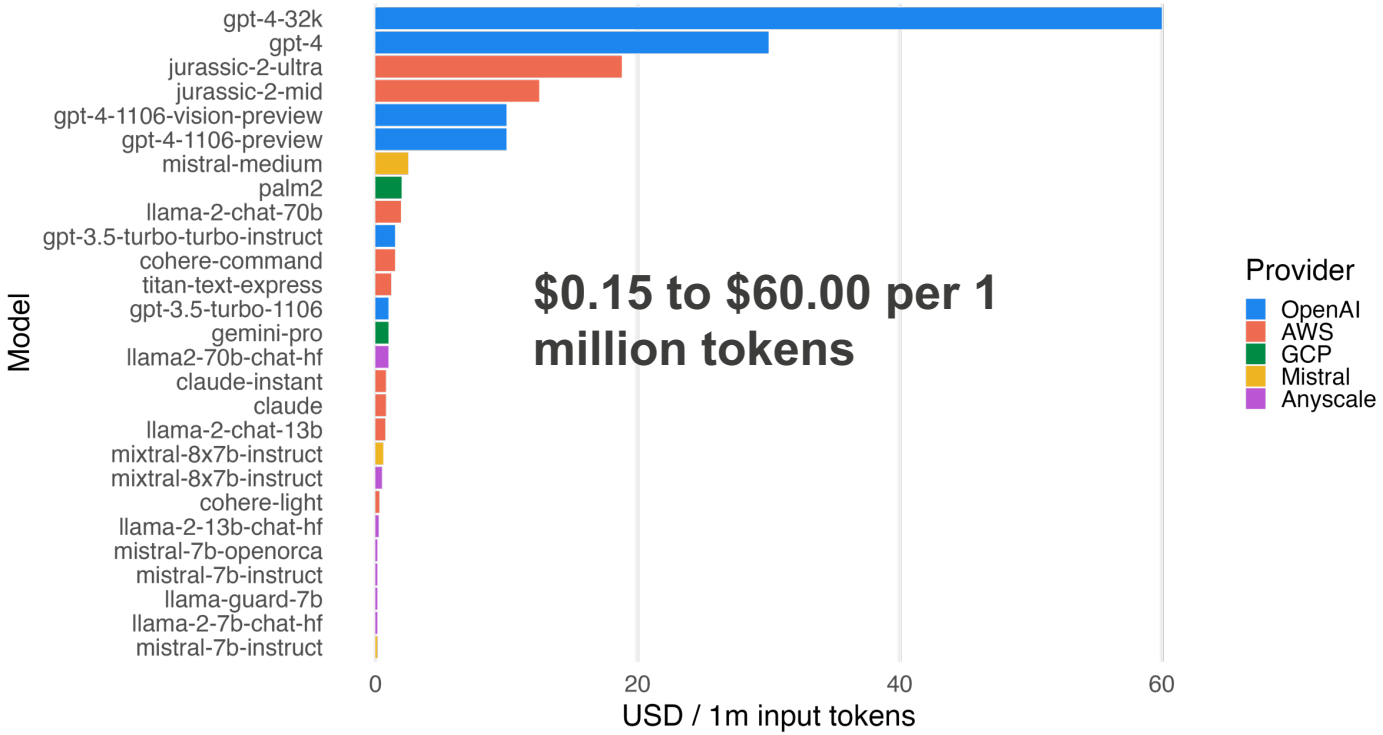
Fine-tuned models are most accurate



Simmering and Huoviala (2023)
Qin et al. (2023)
Ziems et al. (2023)
Wang et al. (2023)

LLM inference is expensive

Token cost by model and provider



Data source: Pricing pages of each provider (2024-01-11)

DistilBERT benchmark

It takes 36s to process 1m tokens using an Nvidia T4 GPU. On demand cost is $0.526/h on AWS. That makes **$0.005/1m tokens**, which is 30x less than the cheapest LLM API.

Tested on Google Colab using SST 2 dataset. See: https://colab.research.google.com/drive/17_Xds6aQAzZlbg0q_dHs836Owm344lfy?usp=sharing

Social media monitoring for oncology
GPT-4 offers schema flexibility

- Cost with OpenAI ~$400
- Changed label scheme multiple times
- No model training, just check of accuracy

→ OpenAI

10k social media posts

gpt-4 via OpenAI API
Multi task text classification

Social media monitoring for cosmetics
CPU model scales to 100m+ texts

- Cost with CPU model < $500
- Trained on 20k human labeled examples
- Label schema is fixed
- Cost with GPT-4 would be > $200k

→

spaCy

aws Lambda

>100m social media posts

2 trained *en_core_web_lg* models: text classification, NER

LLM Finetuning is affordable

OpenAI API finetuning

Fine-tuning GPT-3.5 on 5,759 ABSA examples cost $50.

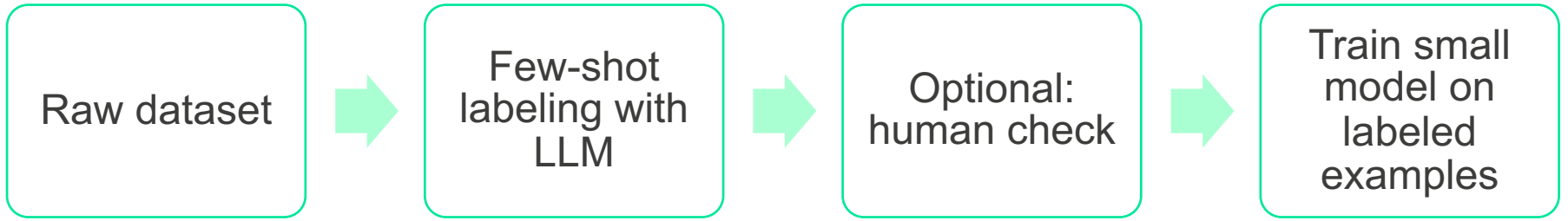
Simmering, Paul F., and Paavo Huoviala. "Large language models for aspect-based sentiment analysis." *arXiv preprint arXiv:2310.18025* (2023).

Finetuning open source LLMs via HuggingFace transformers

With low-rank adaptation (LoRA) and quantization, LLMs can be fine-tuned on consumer GPUs

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Distillation: LLM-based labels for smaller models



Right tool for the job

Aspect	LLM	BERT-like model
Output	Text	Task-specific
Zero/few-shot accuracy	Task-dependent	Doesn’t work
Schema flexibility	High	Low
Prompt engineering	Yes	No
Finetuned accuracy	Very high	High
Fine-tuning cost	Medium (with LoRA)	Low
Multi tasking	Yes, via prompt	Yes, via architecture
Inference cost	High, but falling	Low
Inference throughput	Low	High
Hardware needs	High end GPUs / API	Older GPUs or CPU

Need a partner for NLP projects?

Paul Simmering and Paavo Huoviala

Q Agentur für Forschung

Web: teamq.de

Mail: paul.simmering@teamq.de

X: @Q_InsightAgency

LinkedIn: Q Agentur für Forschung