Think Big:
Entering Large Scale Affective Computing

Björn W. Schuller

ELLIS Summer School on Large-Scale AI for Research and Industry
Modena, 18-22 September 2023

Universität Augsburg University
Imperial College London
audEERING®

Affective Computing

audEERING®
Voice Activity
Voice Activity
-15 s
-10 s
-5 s
now
Audio Input
MacBook Pro Mikrofon
Volume
STOP
PAUSE
high activation
enthusiastic
determined
interested
negative valence
positive valence
low activation
Emotion Categories
happy
angry
sad
neutral
perceived gender:

Convert.

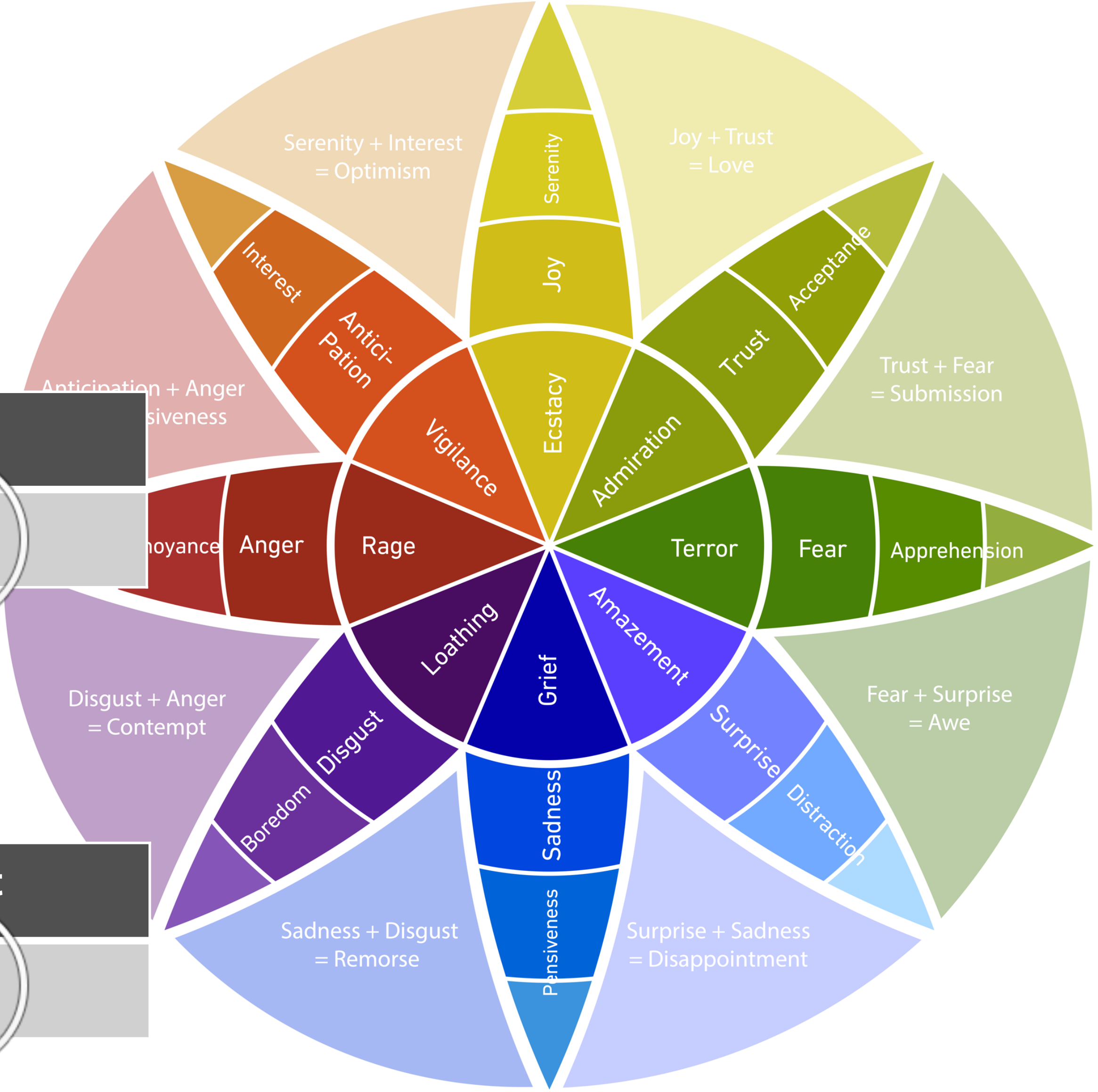
Neutral	Source	Int .1	Int .3	Int .6	Int .9
→ Angry					
→ Sad					
→ Happy					

“Emotion Intensity and its Control for Emotional Voice Conversion”, IEEE Transactions on Affective Computing, 2023.

Björn W. Schuller

Convert.

Surprise	Anger	Sur+Ang	Happy
Sur+.3H	Sur+.6H	Sur+.9H	
Surprise	Outrage	Disappo.	Delight



“Speech Synthesis with Mixed Emotions”, *IEEE Transactions on Affective Computing, 2023.*

Björn W. Schuller

Me.

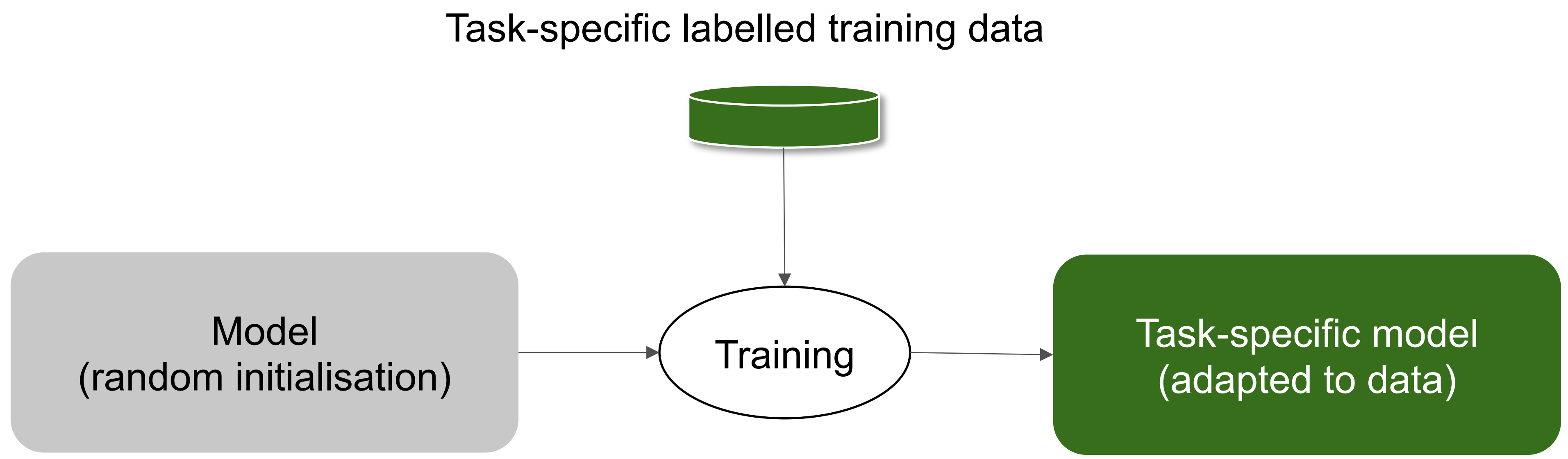
- “A Combined LSTM-RNN-HMM Approach to Meeting Event Segmentation and Recognition”, ICASSP, 2006.
- “Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies”, Interspeech, 2008.
- “Deep neural networks for acoustic emotion recognition: Raising the benchmarks”, ICASSP, 2011.
- “Introducing CURRENNT: the Munich Open-Source CUDA RecurREnt Neural Network Toolkit”, JMLR, 2015.
- “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network”, ICASSP, 2016.
- “End-to-end learning for dimensional emotion recognition from physiological signals”, ICME, 2017.
- “End-to-End Multimodal Emotion Recognition using Deep Neural Networks”, JSTSP, 2017.
- “End2You – The Imperial Toolkit for Multimodal Profiling by End-to-End Learning,” 2018.
- “Dawn of the Transformer Era in Speech Emotion Recognition,” T-PAMI, 2023.

FoundationMs

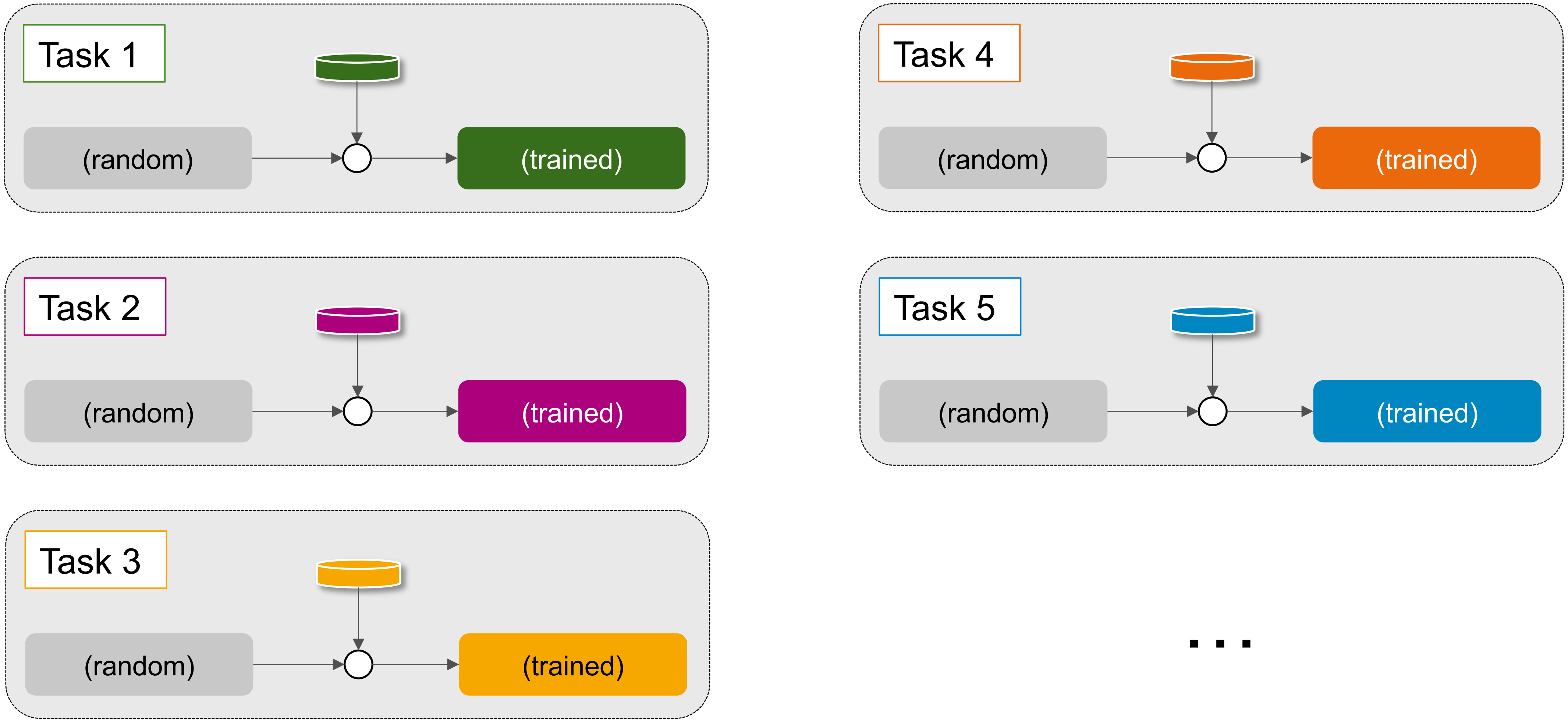
Agenda

1. **Introduction**
2. BERT
3. Further Foundation Models
4. Evaluation
5. Risks and Opportunities

Introduction: The classic supervised ML paradigm

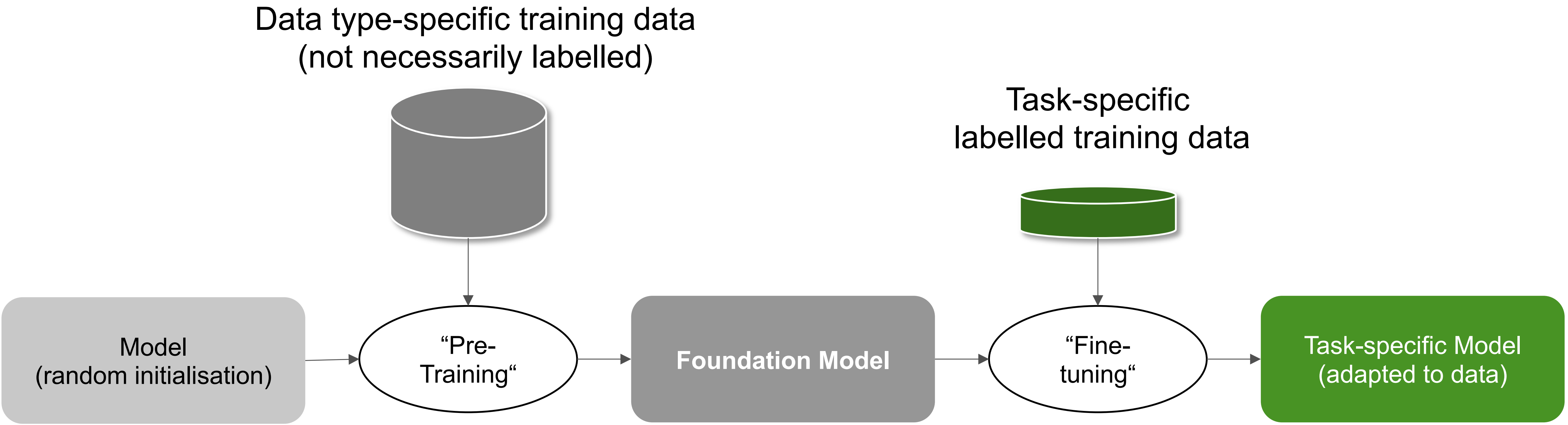


Introduction: The classic supervised ML paradigm



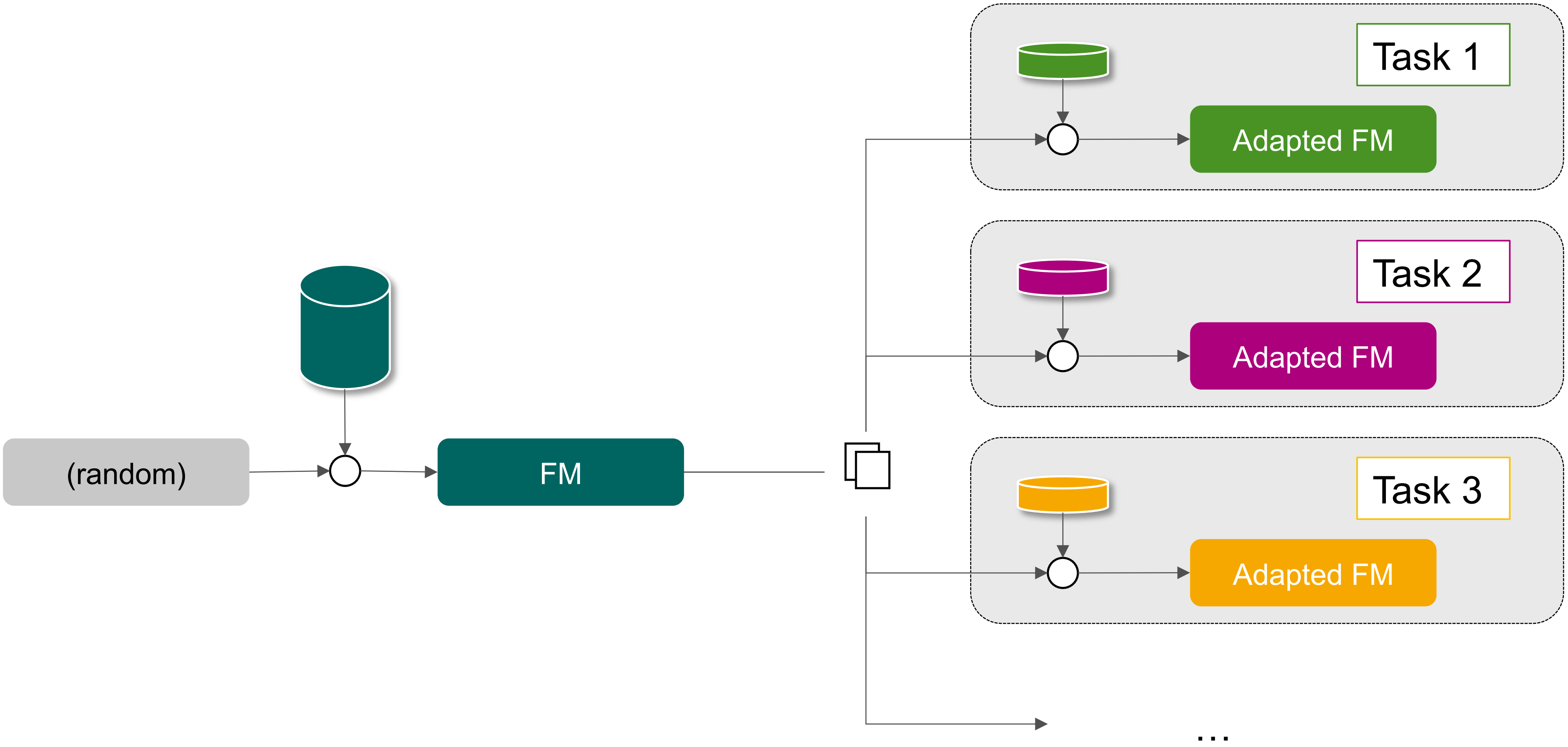
Introduction: The Foundation Model (FM) paradigm

Foundation Models: Pretraining + Finetuning



Introduction: The Foundation Model paradigm

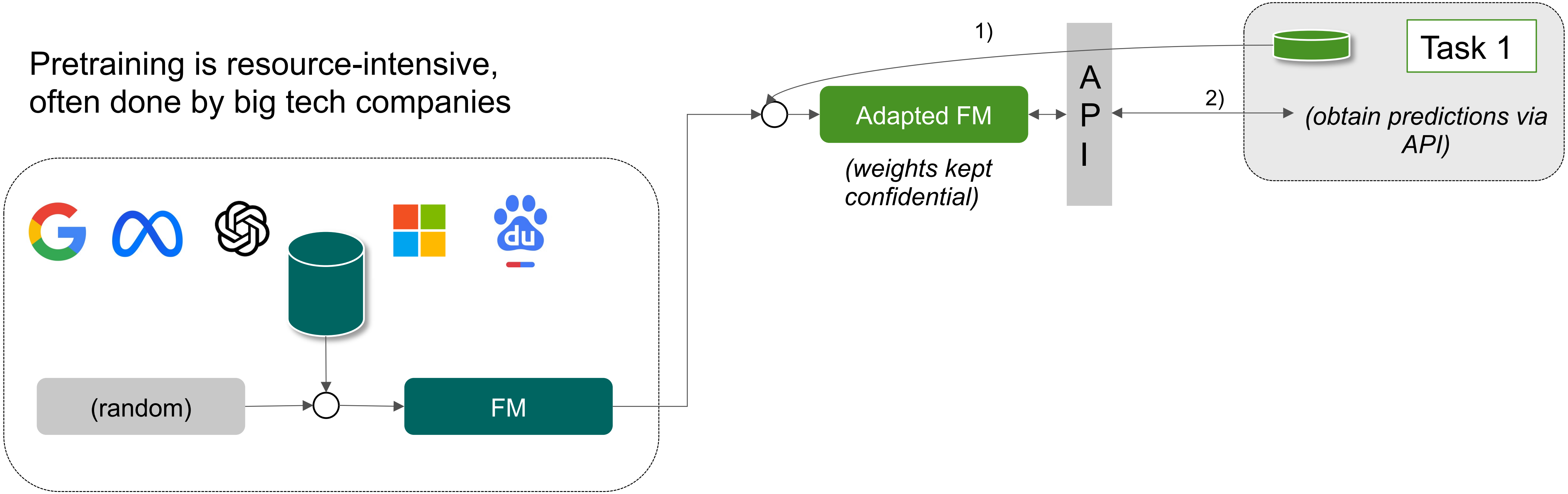
Foundation Models: Pretraining + Finetuning



Introduction: The Foundation Model paradigm

Foundation Models: Pretraining + Finetuning

Pretraining is resource-intensive,
often done by big tech companies

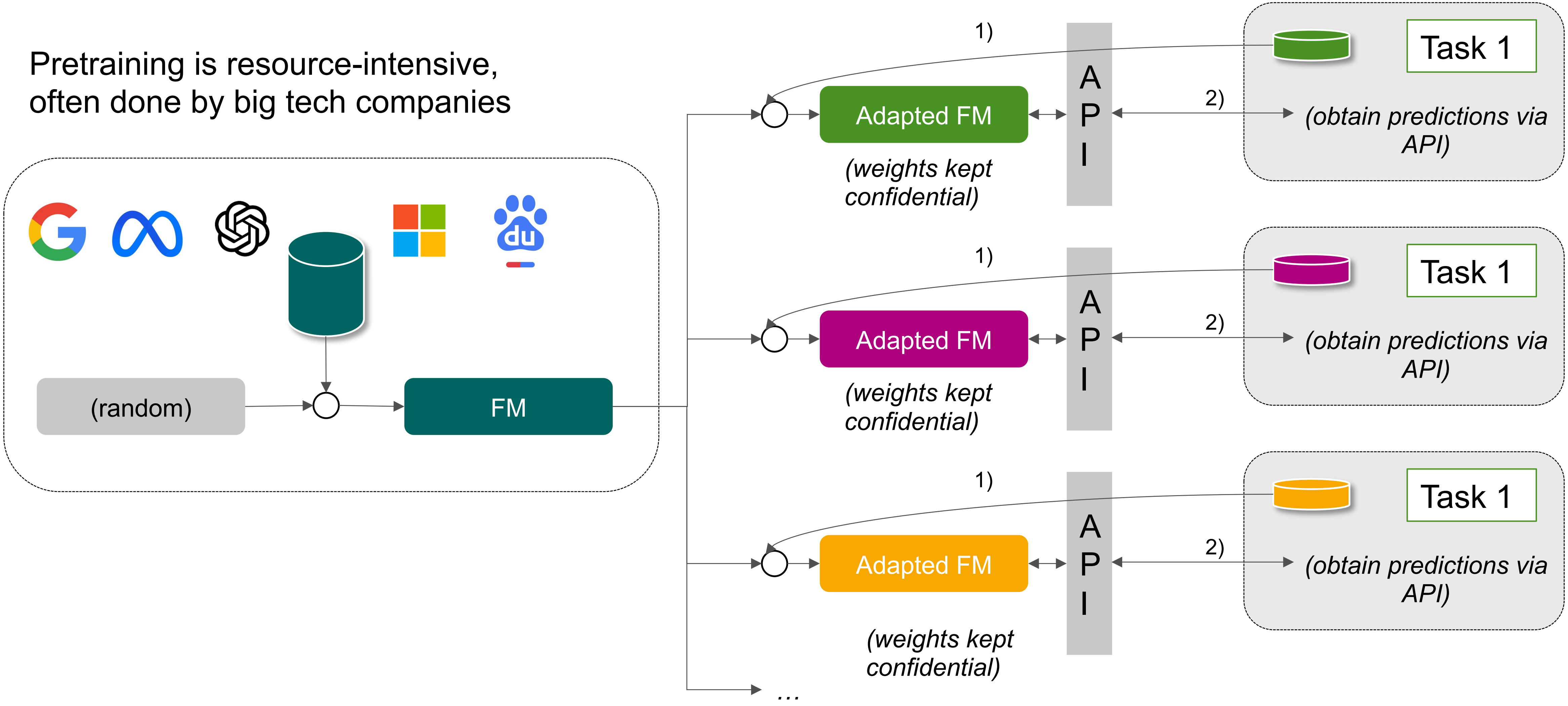


Björn W. Schuller

Introduction: The Foundation Model paradigm

Foundation Models: Pretraining + Finetuning

Pretraining is resource-intensive, often done by big tech companies



Introduction: “Classic“ Supervised ML vs. Foundation Models

“Classic“ Approach

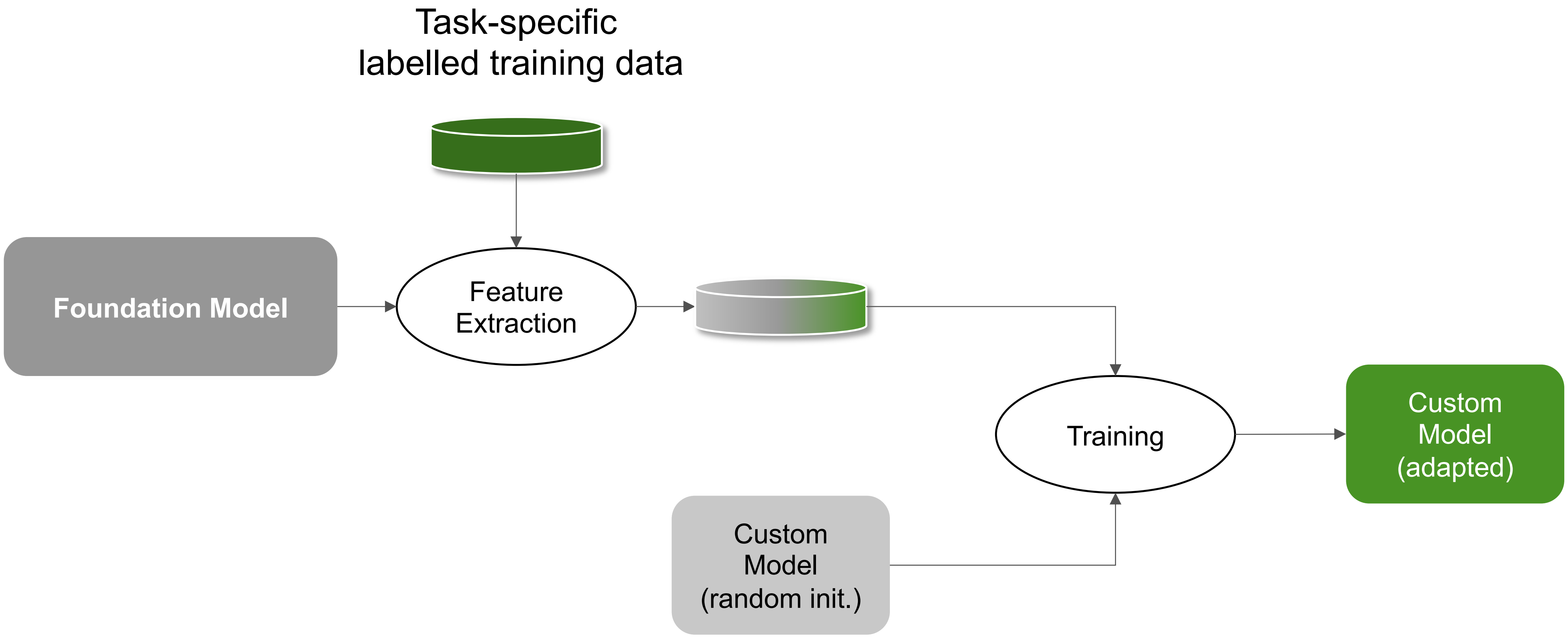
- Comparably few parameters
- Random parameter initialisation
- Training from scratch
- Task-specific model

Foundation Model Approach

- Comparably many parameters (>= 100M)
- Pretraining on large-scale generic data
- “Finetuning“ of pretrained parameters, often possible in just a few steps
- General purpose pretrained model as basis for task-specific finetuned versions of it
- Pretraining often done by large companies (Google, Facebook,…)

Introduction: The Foundation Model paradigm

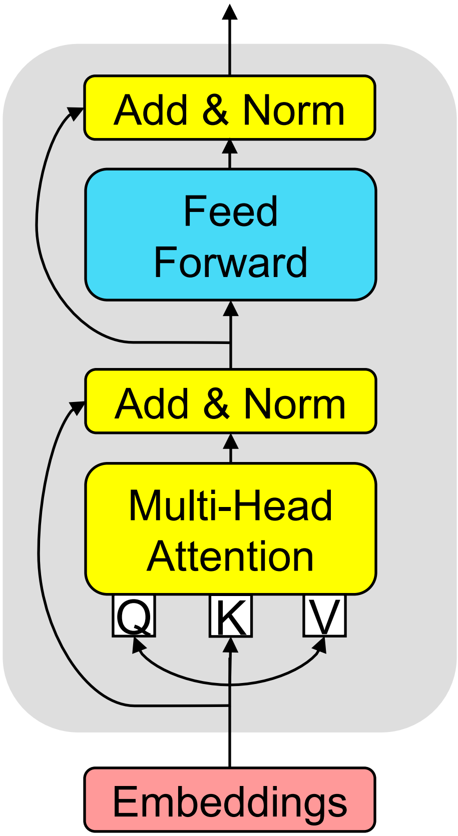
Foundation Models as Feature Extractors



Introduction: Prerequisites of Foundation Models

Transformer Models

- Basis of many FMs today
- More general than RNNs and CNNs: learn arbitrary dependencies between input elements
- Parallelisable
- Typically only encoder part used for FMs



Transformer encoder layer, adapted from Vaswani et al. 2017

Introduction: The Rise of Foundation Models

- ***“A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks”*** (Bommasani et al.[1])

- FMs have lead to a paradigm **shift** in AI

- Arguably since about 2018 (GPT[2], BERT[3])

- Potentially massive societal impact (GPT-3[4], Chat-GPT,…)

- Problem: properties and capabilities of FMs poorly understood

1) Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).
2) Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
3) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
4) Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Introduction: Prerequisites of Foundation Models

Large scale data

- Pretraining requires large amounts of data
- Desirable properties of pretraining datasets:
  - Domain completeness
  - Absence of harmful properties, such as abusive language, bias against/for certain demographics,…
- Datasets of this size can not be checked manually
- Examples of datasets used in different FMs (details will follow):
  - Text of **all Wikipedia entries** for textual FMs (e.g. in BERT[1])
  - **50,000 hours of human speech** for audio/speech FMs (e.g. in Wav2Vec2.0[2])
  - ImageNet-21k (**14M images**, >21k classes, e.g. in ViT[3])

1) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
2) Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460
3) Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020)..

Introduction: Prerequisites of Foundation Models

Computational Resources

Model	Year	# Parameters
BERT[1] (base)	2018	~100M
T5[2]-11B	2020	~11B
GPT-3[3]	2020	~175B
PaLM[4]	2022	~540B

- Pretraining large FMs poses immense hardware requirements
- Example: pretraining of GPT-3 large:
  - 3072 GPUs
  - Overall, >3M GPU hours
- GPU power and memory as main bottleneck in pretraining
- But: GPU power (FLOP/s per dollar) steadily increasing
- Pretrained large FMs typically provided by big tech companies (Google, Facebook, OpenAI,…)

1) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
2) Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.
3) Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
4) Chowdhery, Aakanksha, et al. "Palm: Scaling language modeling with pathways." *arXiv preprint arXiv:2204.02311* (2022).

Agenda

BERT: A very brief introduction into Natural Language Processing (NLP)

- Natural Language Processing (NLP): automatic analysis and processing of natural language texts
- Can be traced back to 1940s
- Paradigms: rule-based vs. stochastic/machine learning
- Nowadays, NLP systems are typically machine learning-based
- Progress in FM was fuelled by NLP models

BERT: A very brief introduction into Natural Language Processing (NLP)

NLP Tasks

- NLP comprises a wide range of different problems, e.g.:
  - Machine Translation
  - Sentiment Analysis and Emotion Recognition
  - Text Summarisation
  - POS-Tagging
  - Dialogue Systems
  - Question Answering

- How to build a FM that provides a good base for all these tasks?

BERT: A very brief introduction into Natural Language Processing (NLP)

Language Modelling

- Language Model M: probability distribution over sequences of words from a vocabulary V

$$M: V^* \to [0,1]$$

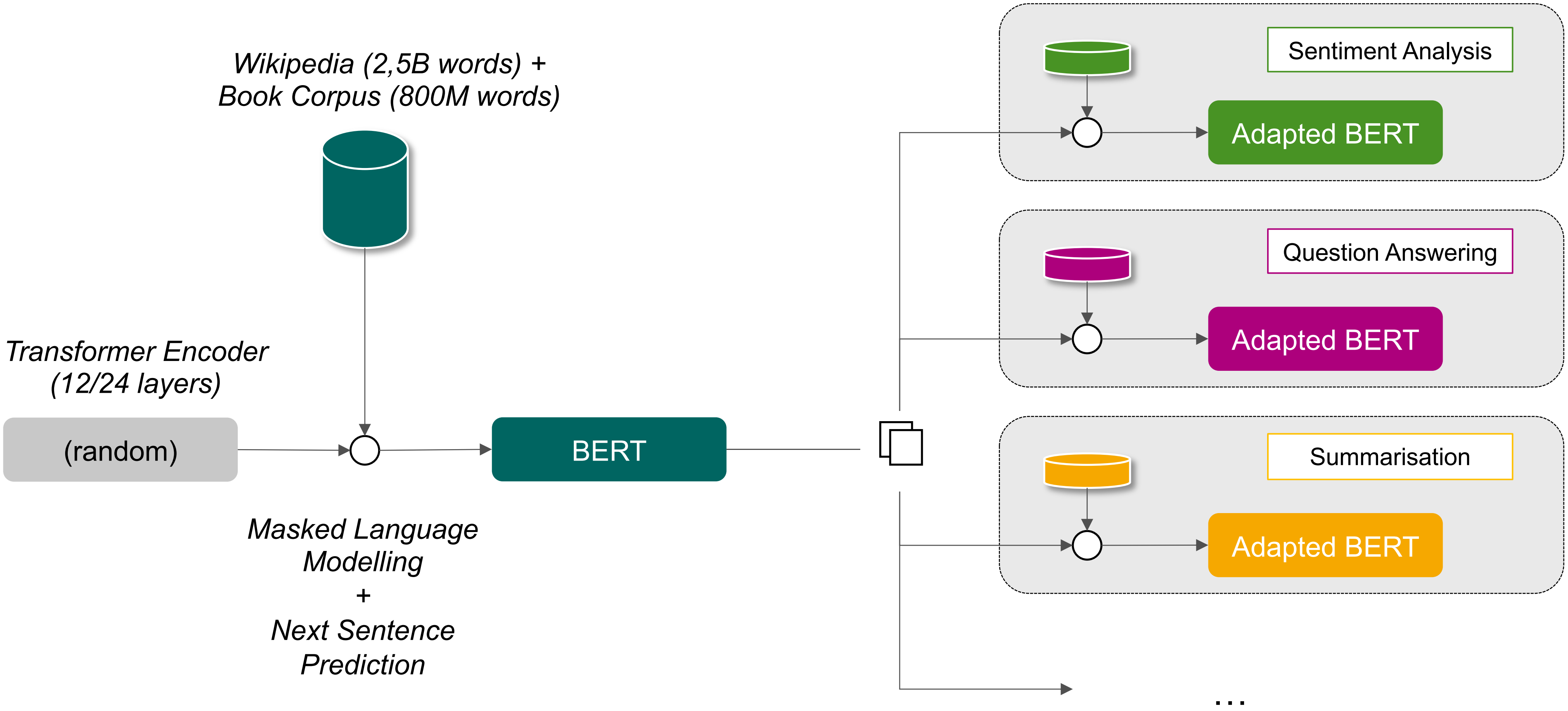
- Intuition: *How likely is a word sequence to occur in this language?*
- LM has general syntactic and semantic knowledge about a language – it knows its rules
- E.g., for a good LM M of English:
  - $M(the\ cat\ barks) < M(the\ dog\ barks)$
  - $M(the\ dog\ barks) > M(the\ barks\ dog)$

- FM for NLP: train a large Transformer model as a Language Model

Björn W. Schuller

BERT: The Big Picture

Foundation Models: Pretraining + Finetuning



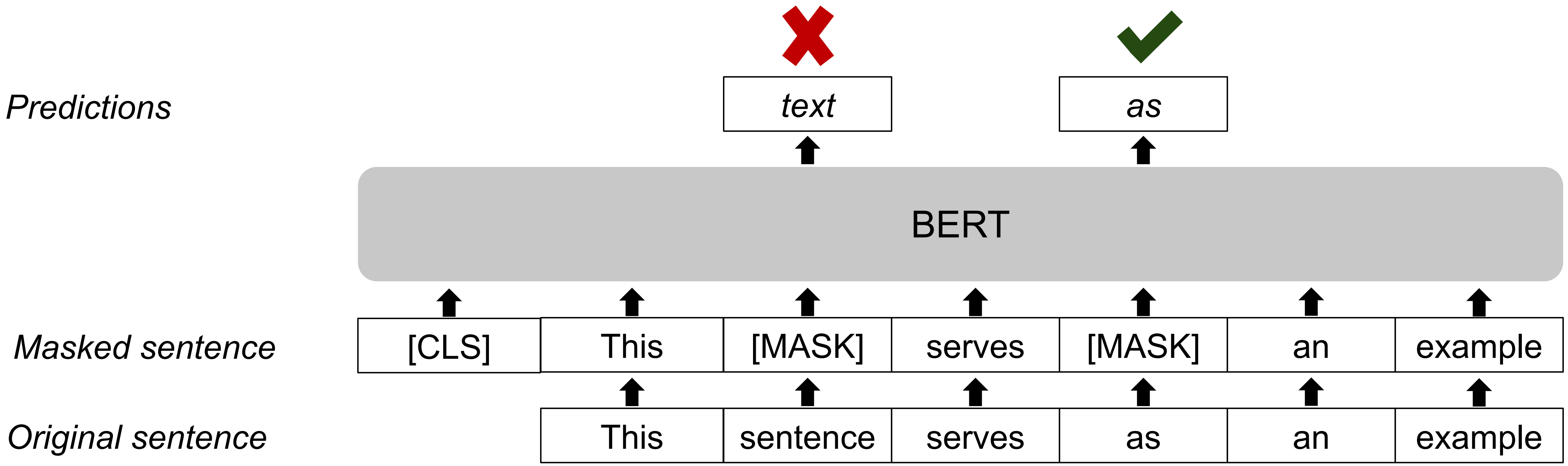
For all details see Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Björn W. Schuller

BERT: Pretraining

Masked Language Modelling (MLM)

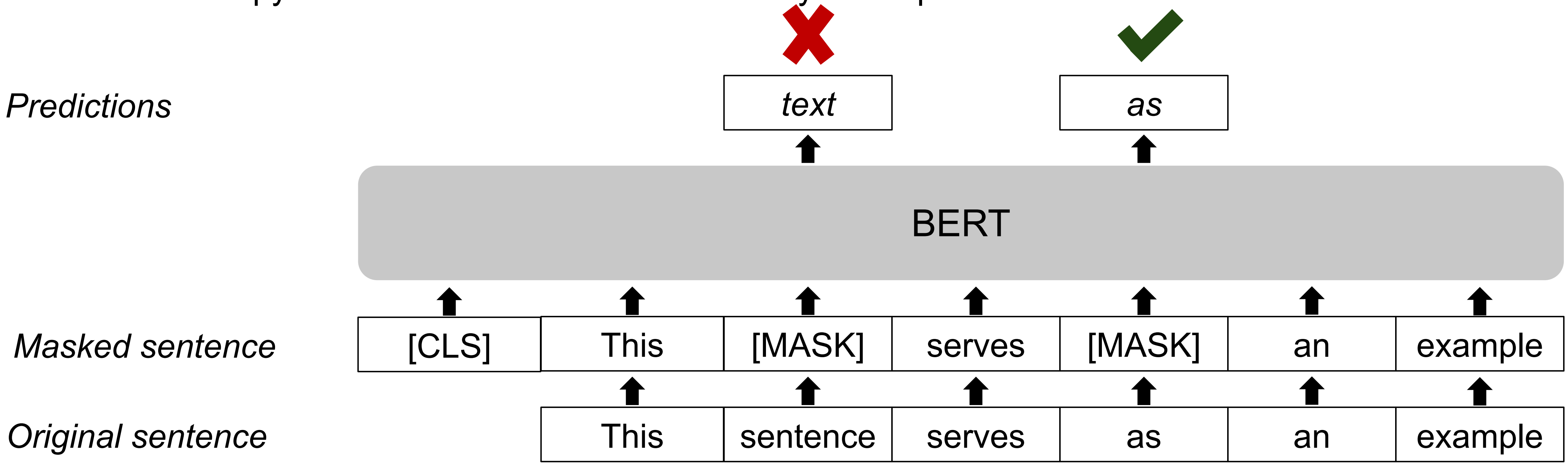
- Randomly mask tokens and predict them
- **B**idirectional: left and right context of masked token available



BERT: Pretraining

Masked Language Modelling Details

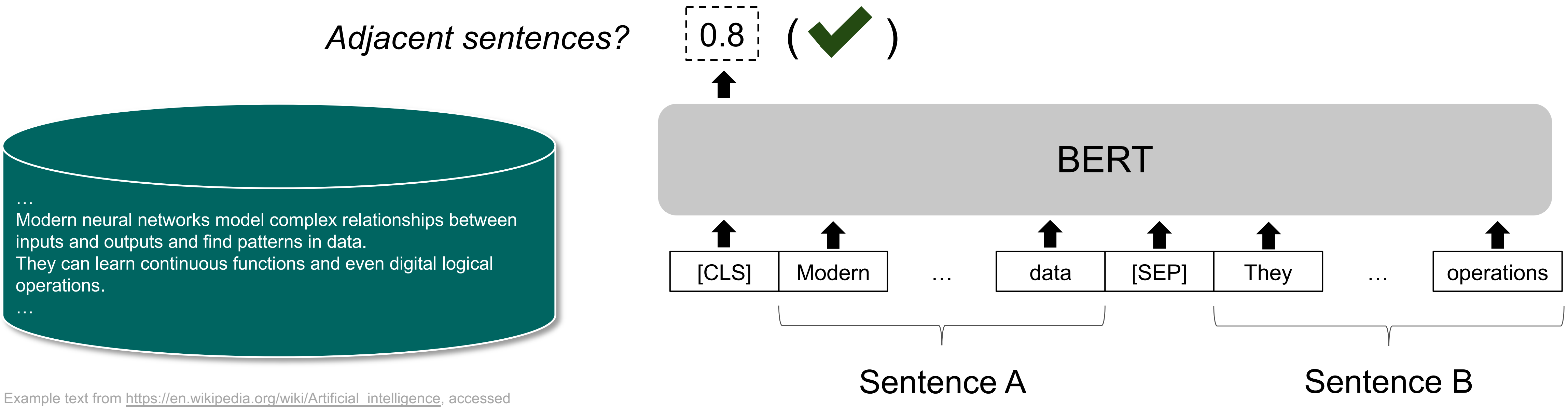
- 15% of all tokens masked
- 80% of them replaced by `[MASK]`, 10% by random token, 10% actually unchanged
- Only consider outputs for masked tokens
- Cross-Entropy Loss: each token in the vocabulary corresponds to a class



BERT: Pretraining

Next Sentence Prediction

- Does sentence B occur after sentence A ?
- 50% positive sentence pairs, 50% random pairings

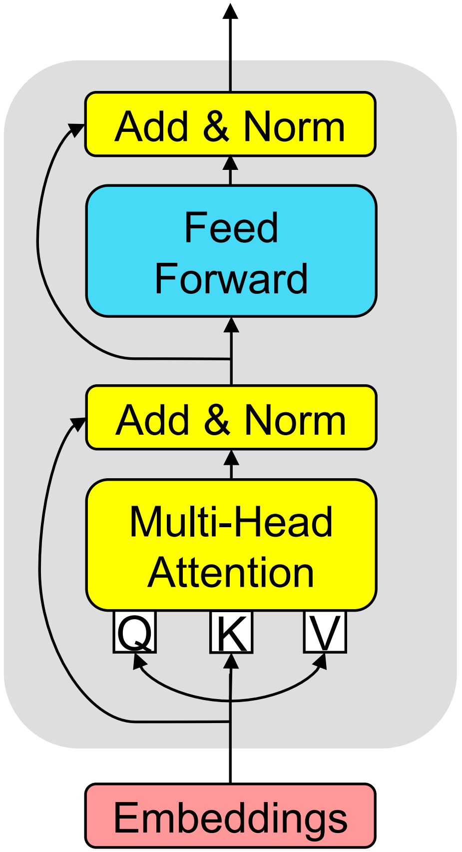


Example text from https://en.wikipedia.org/wiki/Artificial_intelligence, accessed May 01 2023

BERT: Architecture

Recap: Transformers

- Main ingredient: self-attention
- Transformer model stacks several Transformer (encoder/decoder) layers
- Original Transformer (Vaswani et al. 2017): encoder + decoder
- BERT: only encoder
- Variants: BERT-base (12 layers), BERT-large (24 layers)



Transformer encoder layer, adapted from Vaswani et al. 2017[1]

1) Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

BERT: Architecture

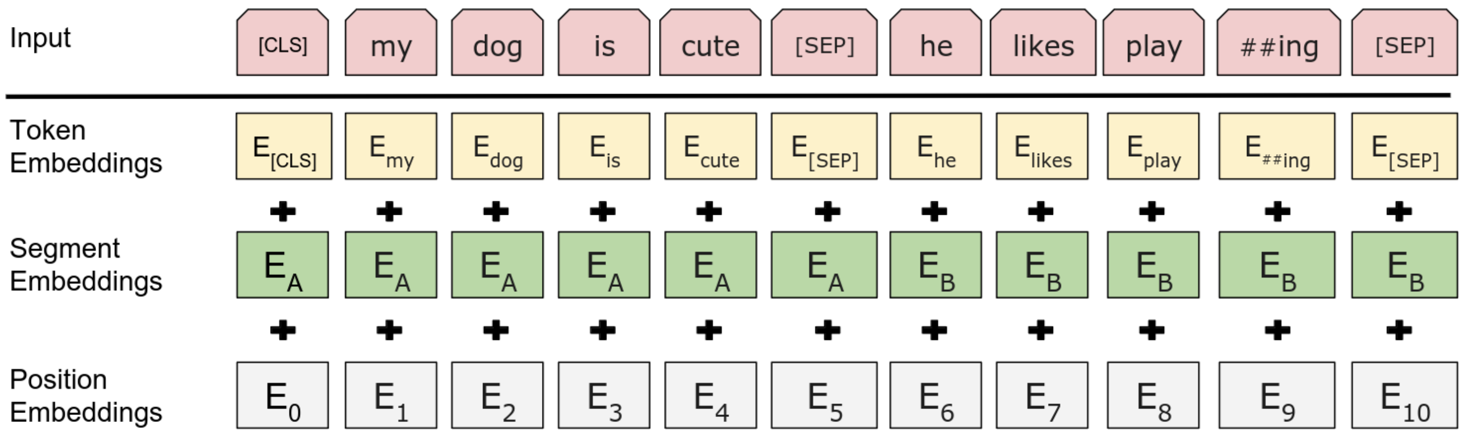
Tokenization

- BERT has a finite vocabulary
- Vocabulary consists of tokens ($\neq$ words)
- Tokens may be sub-word units
- Special tokens:
  - `[CLS]` added at the beginning, intended to represent the whole input sequence
  - `[SEP]` used to separate two sentences in NSP task
  - `[MASK]` used for masking words in MLM task

BERT: Architecture

Embeddings

- Embedding types:
  - Token embeddings
  - Positional embeddings
  - Segment embeddings (for NSP)
- Embeddings are summed up per token



from Devlin et al. 2018

BERT: Downstream Examples

Simple Finetuning: Sentiment Analysis

- Sentiment Analysis: predict sentiment (positive, negative) of sentence
- SST-2 database: movie reviews
- Final layer‘s `[CLS]` embedding as sentence representation
- Feed it into one dense layer (768 x 2)
- All parameters are updated
- Hyperparameters:
  - 3 epochs
  - Search for the best learning rate among 5 candidates

BERT: Downstream Examples

Situations With Adversarial Generations[1] (SWAG)

- >110k multiple choice questions
- Given one sentence A and 4 possible continuations B1,…,B4

Staying under, someone swims past a shark as he makes his way beyond the lifeboat. Turning, he…

a) glances toward the stage.

b) finds the grieving baby sitting on his gray chair.

c) poses with this mouth close to hers

d) finds himself facing the completely submerged ship

- Training examples for BERT: `[CLS]` A `[SEP]` B1; …; `[CLS]` A `[SEP]` B4
- Finetuning:
  - Additional trainable vector
  - Dot product of this vector and the 4 `[CLS]` representations
  - Softmax

1) Zellers, Rowan, et al. "Swag: A large-scale adversarial dataset for grounded commonsense inference." *arXiv preprint arXiv:1808.05326* (2018).

BERT: Downstream Examples

No general recipe for finetuning

- In general, hyperparameter search necessary
- BERT paper reports unstable finetuning for some tasks
- “How to Fine-Tune BERT for Text Classification?“ (Sun et al. 2019[1]) explores:
  - Further in-domain pretraining before fine-tuning
  - Layer-wise decreasing learning rates
  - Multitask learning
- Overall, there is no general solution to finetuning
- But there exist some best practices

1) Sun, Chi, et al. "How to fine-tune bert for text classification?." *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*. Springer International Publishing, 2019.

Agenda

Further Foundation Models: Models based on BERT

ALBERT

- Several subsequent works aim for improving BERT

- ALBERT[1] (A Lite BERT) adresses two aspects:
  - Parameter reduction by
    - Parameter sharing across transformer layers
    - Factorisation of the embedding matrix
  - Pretraining task:
    - NSP may be too simple because of negative pairs randomly sampled
    - Sentence-Order Prediction (SOP) instead:
      - all training pairs are adjacent sentences
      - Randomly swap 50% of them

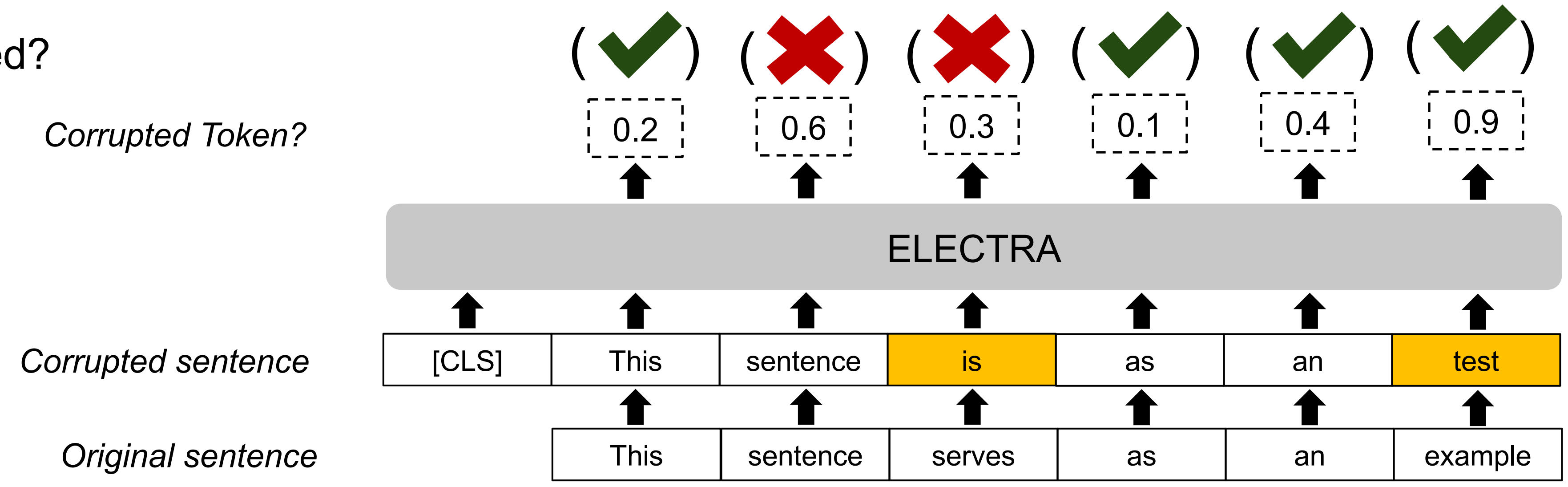
- ALBERT typically outperforms BERT

1) Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." *arXiv preprint arXiv:1909.11942* (2019).

Further Foundation Models: Models based on BERT

ELECTRA[1]

- Problems with BERT‘s MLM objective:
  - Only 15% of the tokens used for learning
  - `[MASK]` tokens not present during finetuning
- ELECTRA: Replaced token detection
  - Randomly (15%) manipulate **every** token
  - Predict: token changed?

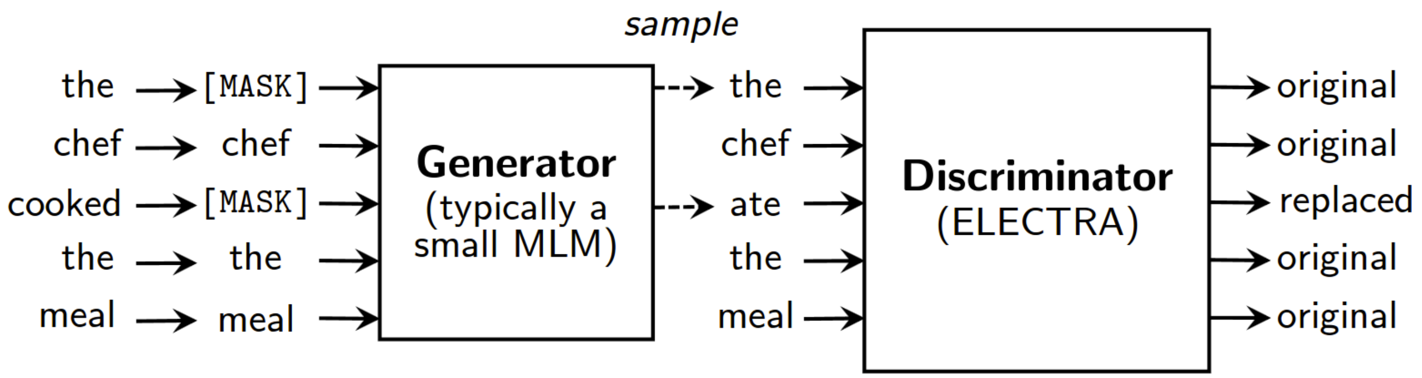


1) Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* (2020).

Further Foundation Models: Models based on BERT

ELECTRA

- Token replacement via a (small) MLM-trained model – this is also trainable



From Clark et al. 2019

- Generator is discarded after pretraining
- ELECTRA converges faster than BERT
- ELECTRA typically outperforms BERT

Further Foundation Models: Models based on BERT

BERT-like Models based on specific datasets/languages

- Multilingual BERT[1]: Wikipedias of 104 languages
- CamemBERT[2]: french texts
- DiLBERT[3] (Disease Language BERT): ICD-11, PubMed, Wikipedia for “disease-related language“
- BERTweet[4]: 850M Tweets
- Med-BERT[5]: Electronic Health Records

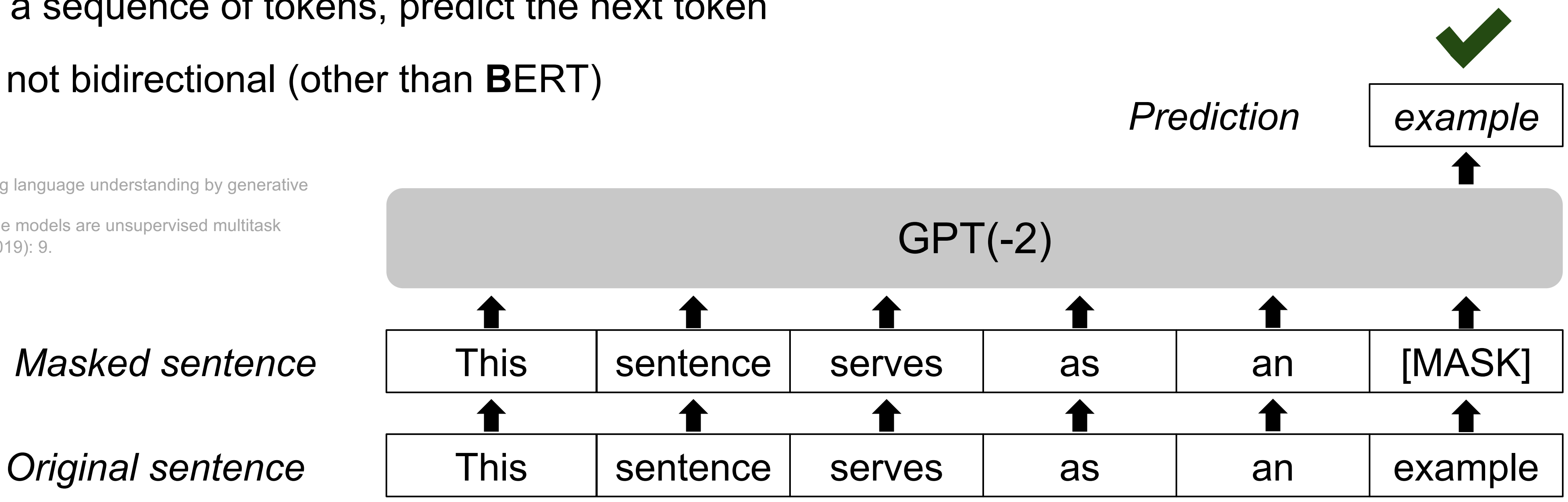
1) Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
2) Martin, Louis, et al. "CamemBERT: a tasty French language model." *arXiv preprint arXiv:1911.03894* (2019).Brown, Tom, et al.
3) Roitero, Kevin, et al. "DiLBERT: Cheap embeddings for disease related medical NLP." *IEEE Access* 9 (2021): 159714-159723.
4) Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. "BERTweet: A pre-trained language model for English Tweets." *arXiv preprint arXiv:2005.10200* (2020).
5) Rasmy, Laila, et al. "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction." *NPJ digital medicine* 4.1 (2021): 86.

Björn W. Schuller

Further Foundation Models: Other language models

GPT, GPT-2

- Generative Pretrained Transformers (GPT)
- Architecture comparable to BERT: Transformer encoder layers
- GPT[1] and GPT-2[2] were trained with Causal Language Modelling (CLM):
  - Given a sequence of tokens, predict the next token
  - Thus, not bidirectional (other than **B**ERT)

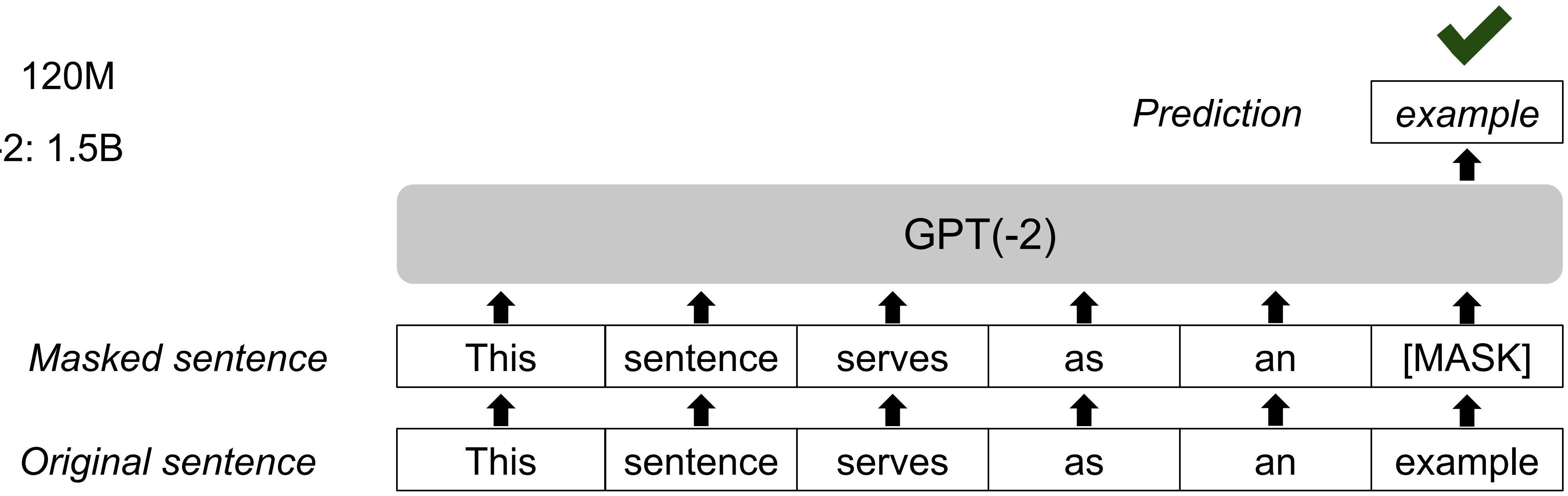


1) Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
2) Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

Further Foundation Models: Other language models

GPT, GPT-2

- Training data:
  - GPT: BooksCorpus + 1B Word Benchmark
  - GPT-2: WebText, scraped from 8M webpages
- Size:
  - GPT: 120M
  - GPT-2: 1.5B



Further Foundation Models: Other language models

Zero-Shot learning in GPT-2

- GPT-2 evaluated in a setting without any supervised training (zero-shot)
- Examples:
  - Summarise a text T
    - Prompt GPT-2: T + `TL; DR:`
    - Take the first 100 generated tokens
  - Answering a factual question Q
    - Prompt GPT-2: Q1 A1 … Qn An Q
    - Example pairs (Q1 A1) … (Qn An) to enforce the desired answer style
    - Take the first generated word

Further Foundation Models: Extremely large language models

GPT-3

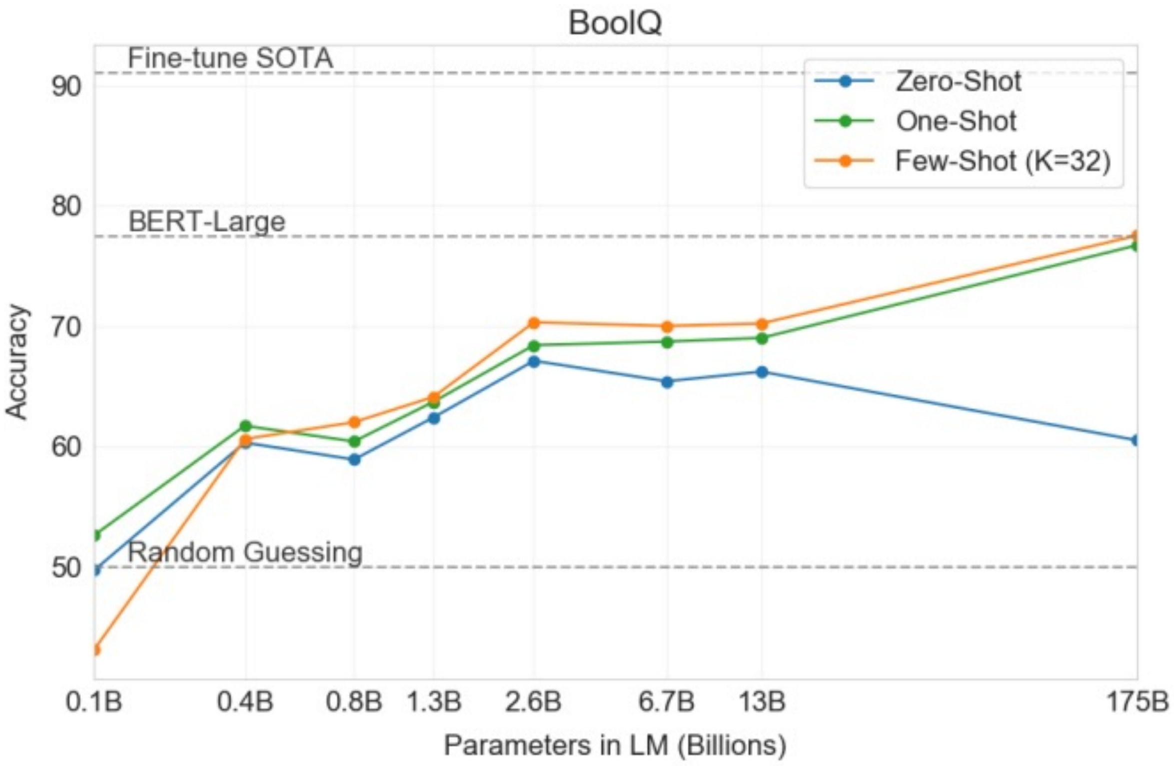
- GPT-3[1] takes GPT-like models to next level
- Architecture: similar to BERT and GPT-2, but:
  - 96 layers
  - 12288-dimensional embeddings (BERT, GPT-2: 768)
  - 96 attention heads
  - 175B parameters (GPT-2: 1.5B)
- Training data: 570 GB text (GPT-2: 40 GB)

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

Further Foundation Models: Extremely large language models

GPT-3: Zero-Shot, Few-Shot

- Impressive zero-shot / few-shot capabilities (learning from no / very few training examples)



from Brown et al. 2020

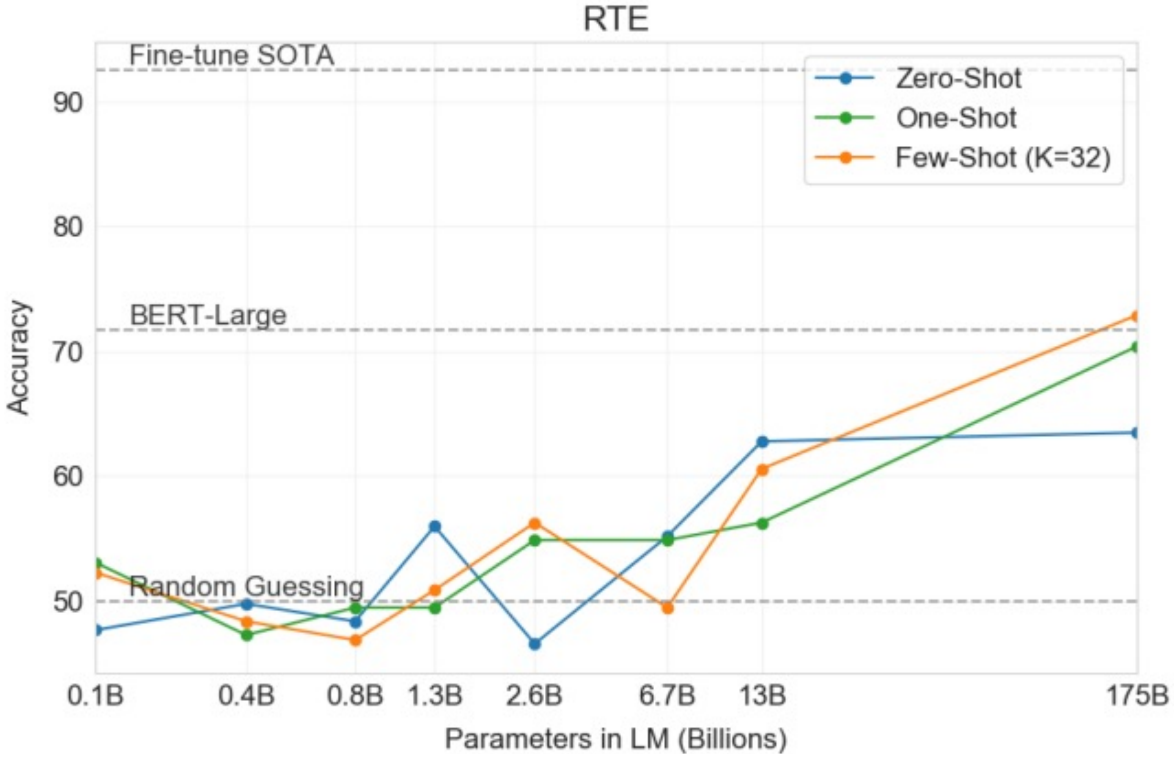
BoolQ[1]: Yes/No Question Answering

1) Clark, Christopher, et al. "BoolQ: Exploring the surprising difficulty of natural yes/no questions." *arXiv preprint arXiv:1905.10044* (2019).

Further Foundation Models: Extremely large language models

GPT-3: Zero-Shot, Few-Shot

- Impressive zero-shot / few-shot capabilities (learning from no / very few training examples)



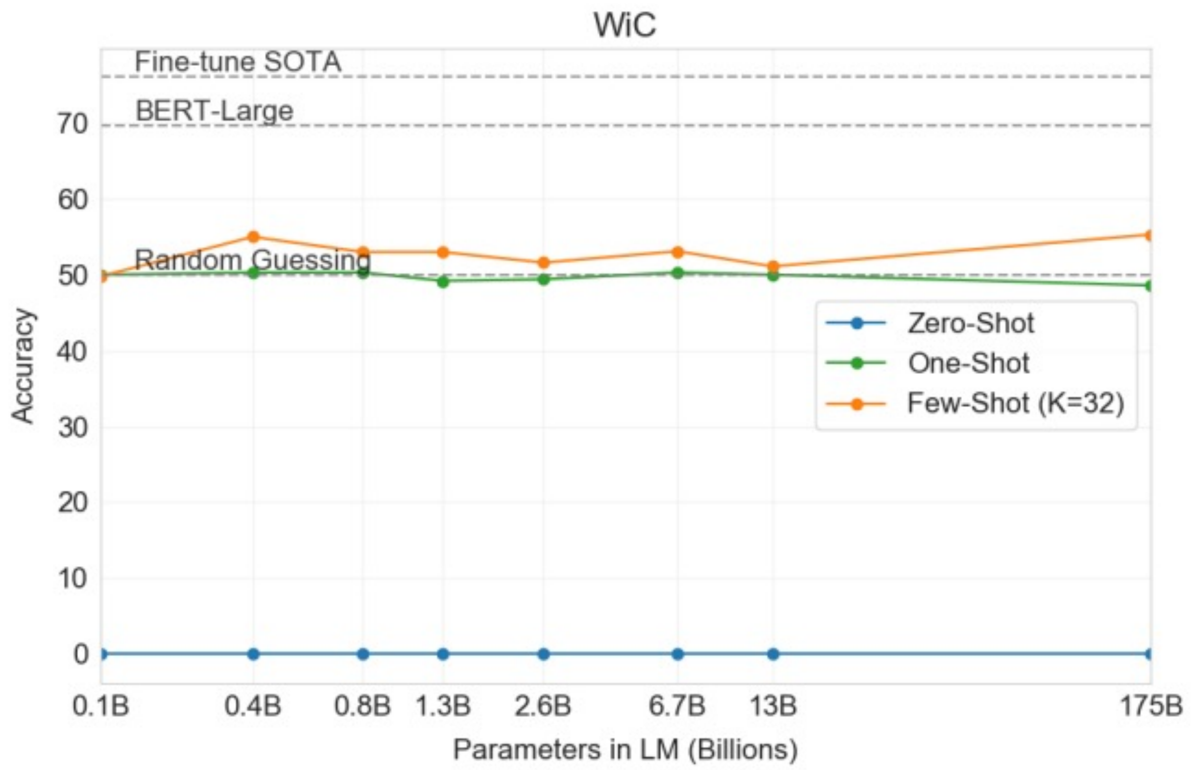
from Brown et al. 2020

RTE: Textual Entailment (does text A imply text B?)

Further Foundation Models: Extremely large language models

GPT-3: Zero-Shot, Few-Shot

- Impressive zero-shot / few-shot capabilities (learning from no / very few training examples) – though not on all tasks



from Brown et al. 2020

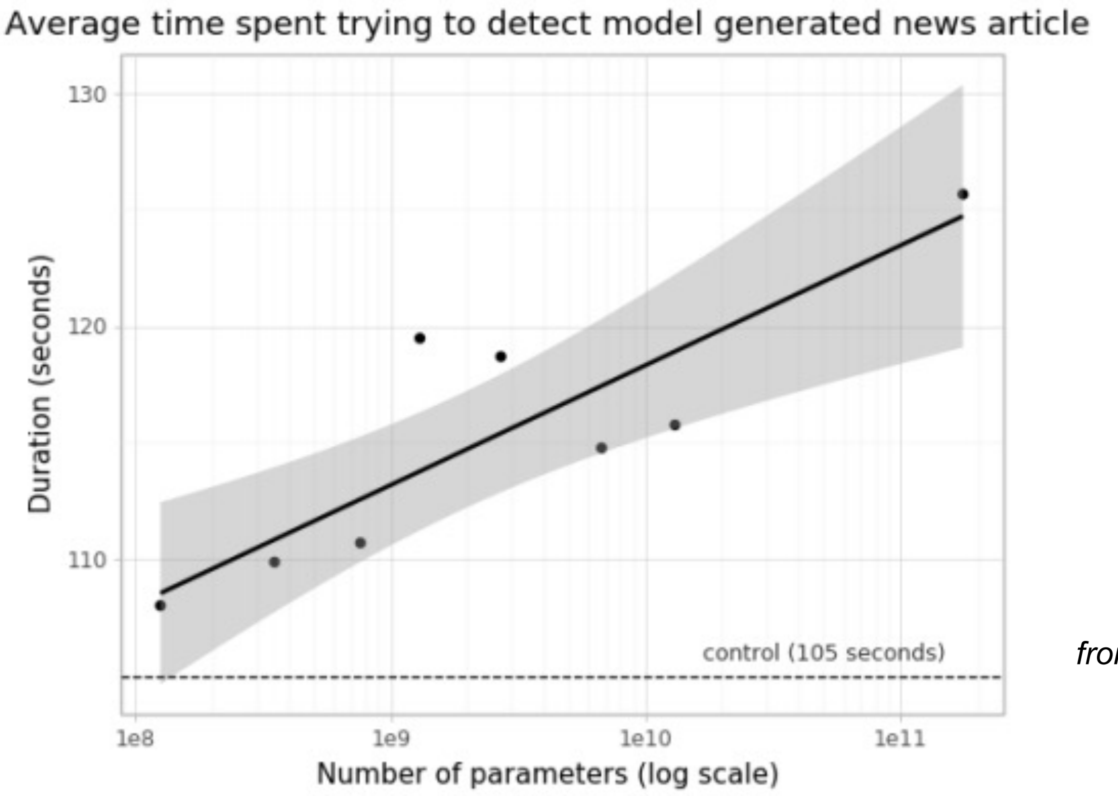
WiC[1]: Words in Context – distinguishing between ambiguous word meanings

1) Pilehvar, Mohammad Taher, and Jose Camacho-Collados. "WiC: the word-in-context dataset for evaluating context-sensitive meaning representations." *arXiv preprint arXiv:1808.09121* (2018).

Further Foundation Models: Extremely large language models

GPT-3: Text Generation

- Impressive text generation capabilities



from Brown et al. 2020

The larger the model, the more time humans need to distinguish actual news articles from such generated by the model. The line represents a linear model fitted to the data points.

Björn W. Schuller

Further Foundation Models: Extremely large language models

Trends in large LM sizes

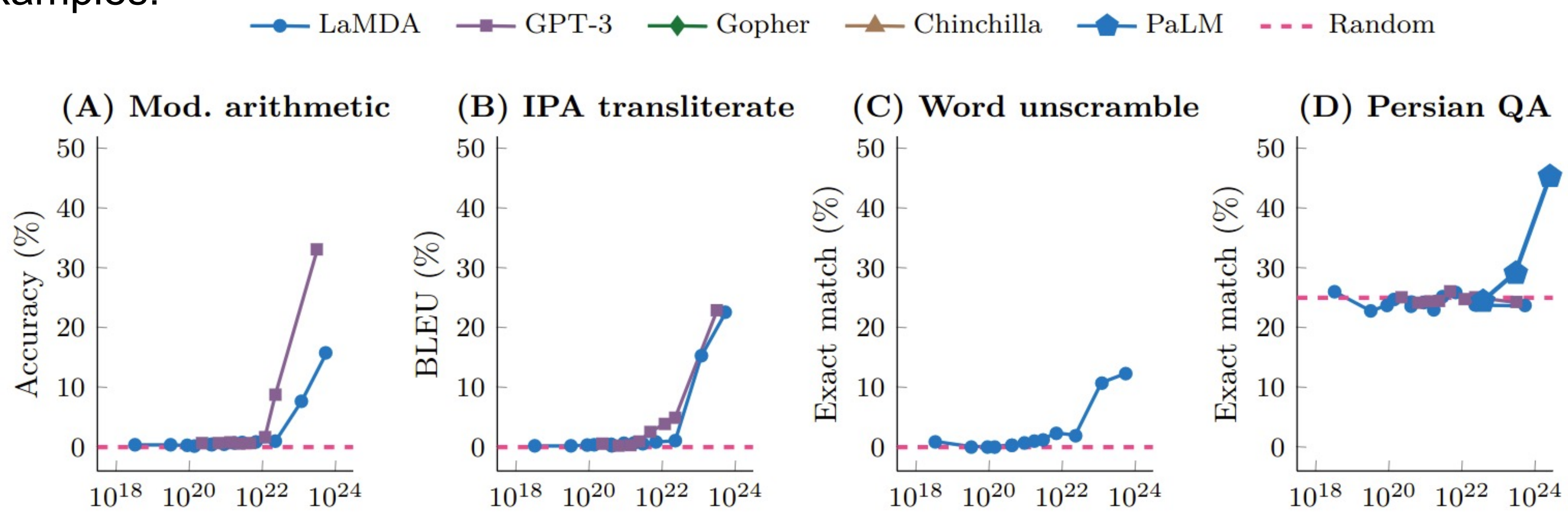
Year	Company	Model Name	# Parameters
2020	OpenAI	GPT-3	175B
2021	Microsoft/NVIDIA	Megatron-Turing NLG	530B
2021	Google	GLaM	1.2T
2022	Google	PaLM	540B
2023	Meta	LLaMA	65B
2023	Huawei	PanGu-Σ	1.1T

- After GPT-3, LM sizes were increased even more
- Recently, more interest in reducing scale

Further Foundation Models: Extremely large language models

Emergence in very large LMs

- *"Emergence is when quantitative changes in a system result in qualitative changes in behavior"*[1]
- Here: large LMs have capabilities that smaller ones do not (same architecture, pretraining method)
- In particular, few-shot and zero-shot scenarios (i.e., few or no labelled examples)
- Emergent attributes/capabilities can not be predicted based on smaller LMs
- Examples:



from Wei et al. 2022

1) Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).

Further Foundation Models: Extremely large language models

From GPT-3 to ChatGPT

- InstructGPT
  - GPT-3 + humans in the loop
  - Further adapt model based on human feedback
  - Motivation:
    - Improve response to instructions
    - Reduce toxicity
    - Reduce hallucination of facts
- ChatGPT
  - Finetuning of GPT-3 similar to InstructGPT
  - Conversation data
  - Human feedback on “good“ vs. “bad“ responses

Further Foundation Models: Other data types

Overview

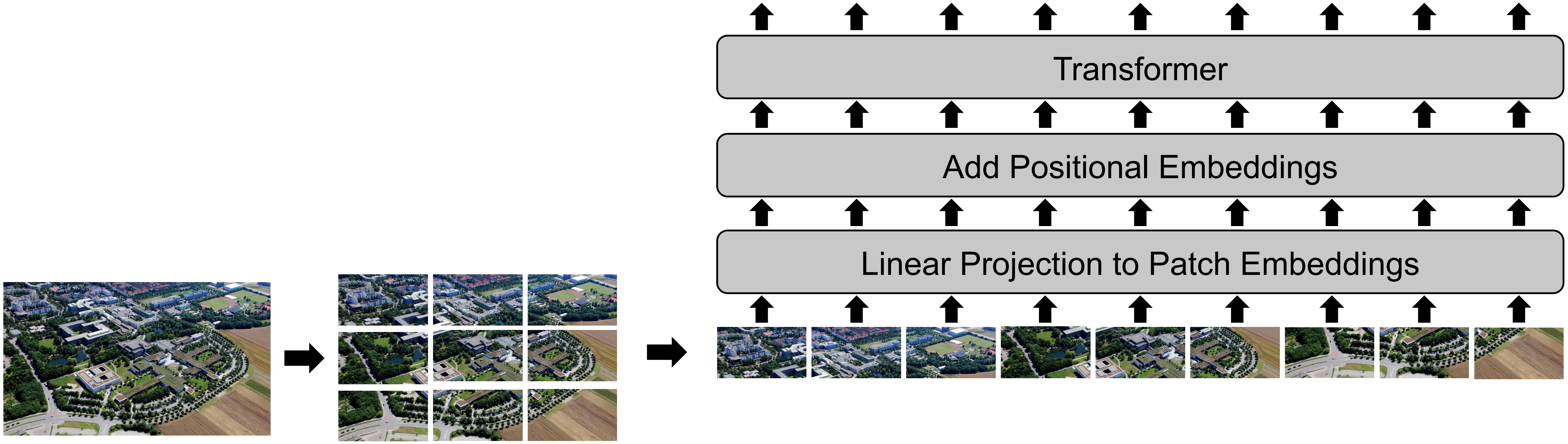
- GPT and BERT as first Transformer-based FMs
- Development of comparable models was mainly done in the NLP domain
- Later on, Transformer-based FMs for other data types were introduced:
  - Audio/Speech
  - Video
  - Image + Text
  - …

Further Foundation Models: Vision

Vision Transformer (ViT)

For Details see Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020)..

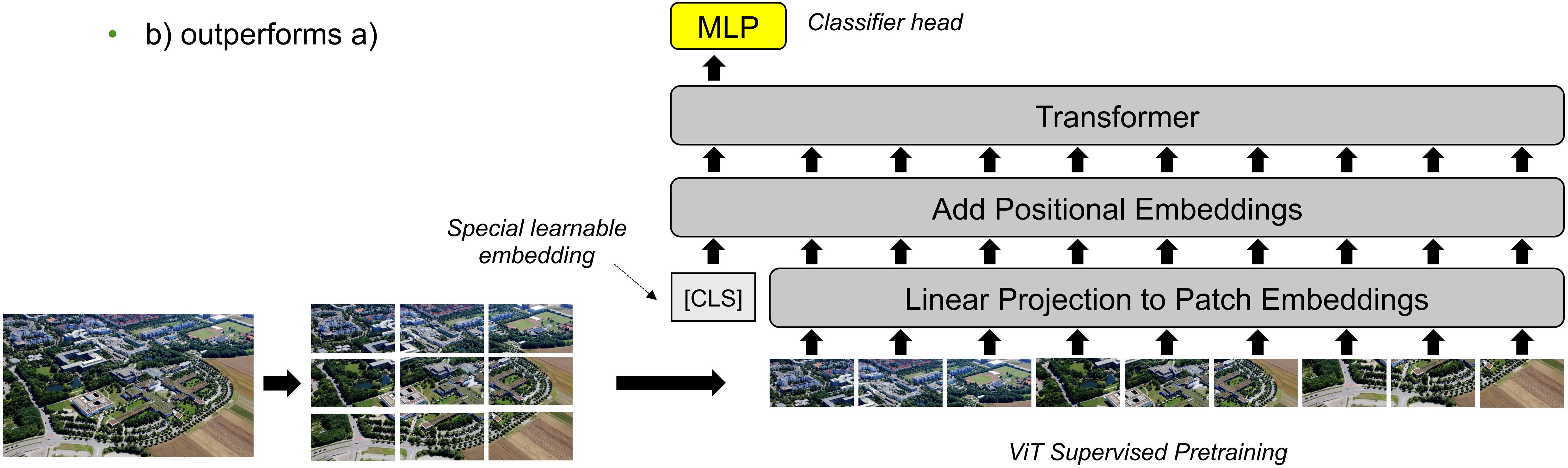
- Input: Image of fixed(!) size
- Input representation: fixed-size patches, flattened
- Pretraining datasets: JFT (303M images from 18k classes), ImageNet-21k (14M images, >21k classes)



Further Foundation Models: Vision

ViT

- Pretraining target:
  - a) self-supervised “masked patch prediction“ (~MLM in BERT)
  - b) **supervised(!) image classification**
  - b) outperforms a)



Further Foundation Models: Text + Vision

Contrastive Language-Image Pretraining (CLIP)

For details see Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

- Input: Image-Text pairs
- Input representation: images as in ViT, texts as in BERT etc.
- Pretraining dataset: 400M text-image pairs, crawled from the internet
- Model:

Text embedding

Text Encoder (Transformer)

“University of Augsburg, bird‘s eye view“

Image embedding

Image Encoder (Transformer)

Further Foundation Models: Text + Vision

CLIP Pretraining

- Batch of N images, N texts -> N x N pairs
  - N of them actual pairs
  - N^2 – N incorrect
- Obtain all text and image embeddings $t_1 \dots t_N$, $i_1 \dots i_N$
- Cosine similarity for every pair (t_i, i_j)
- Binary cross entropy loss on similarities (pair yes/no)

Further Foundation Models: Text + Vision

CLIP: Zero-Shot Image Classification

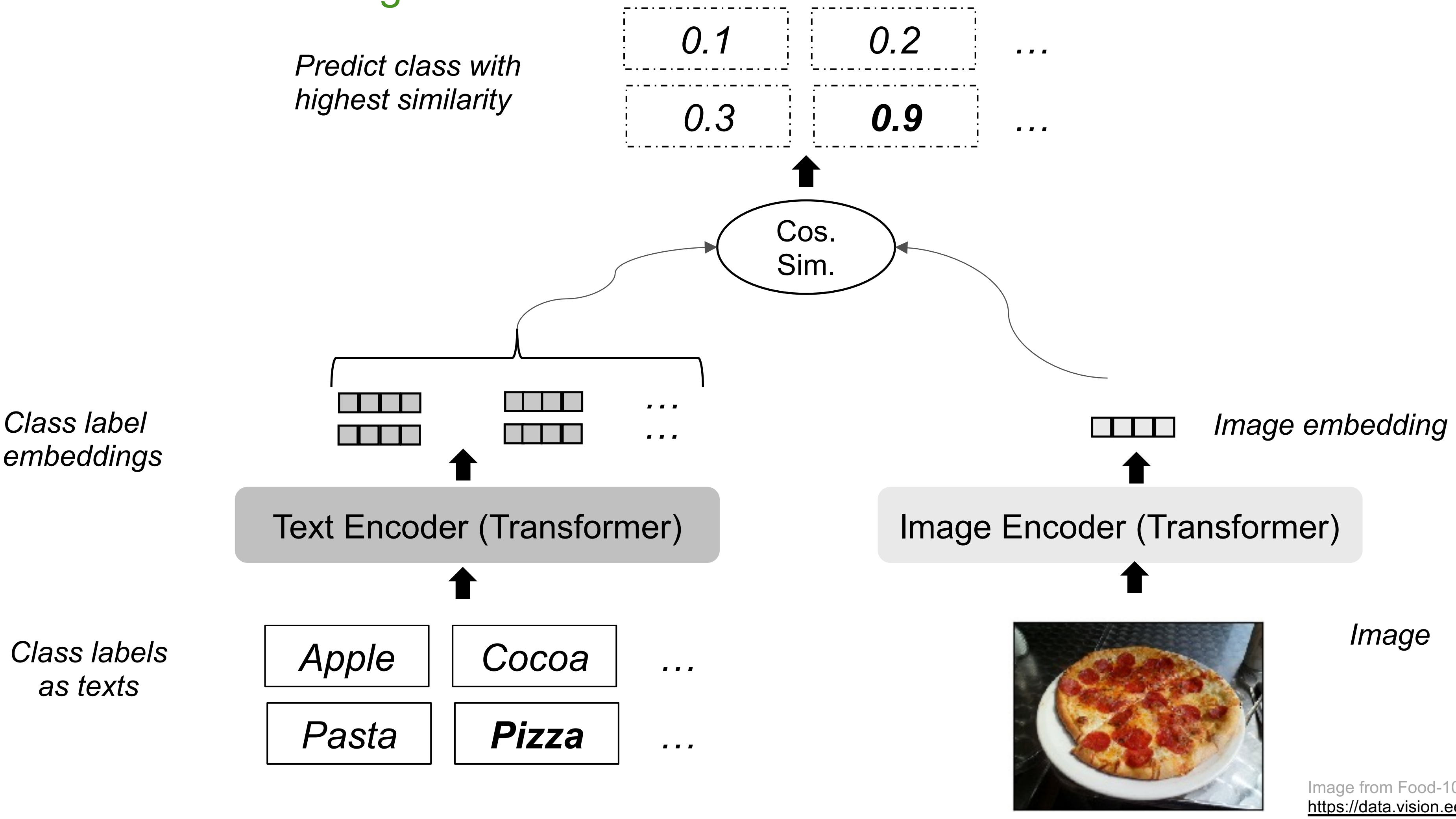
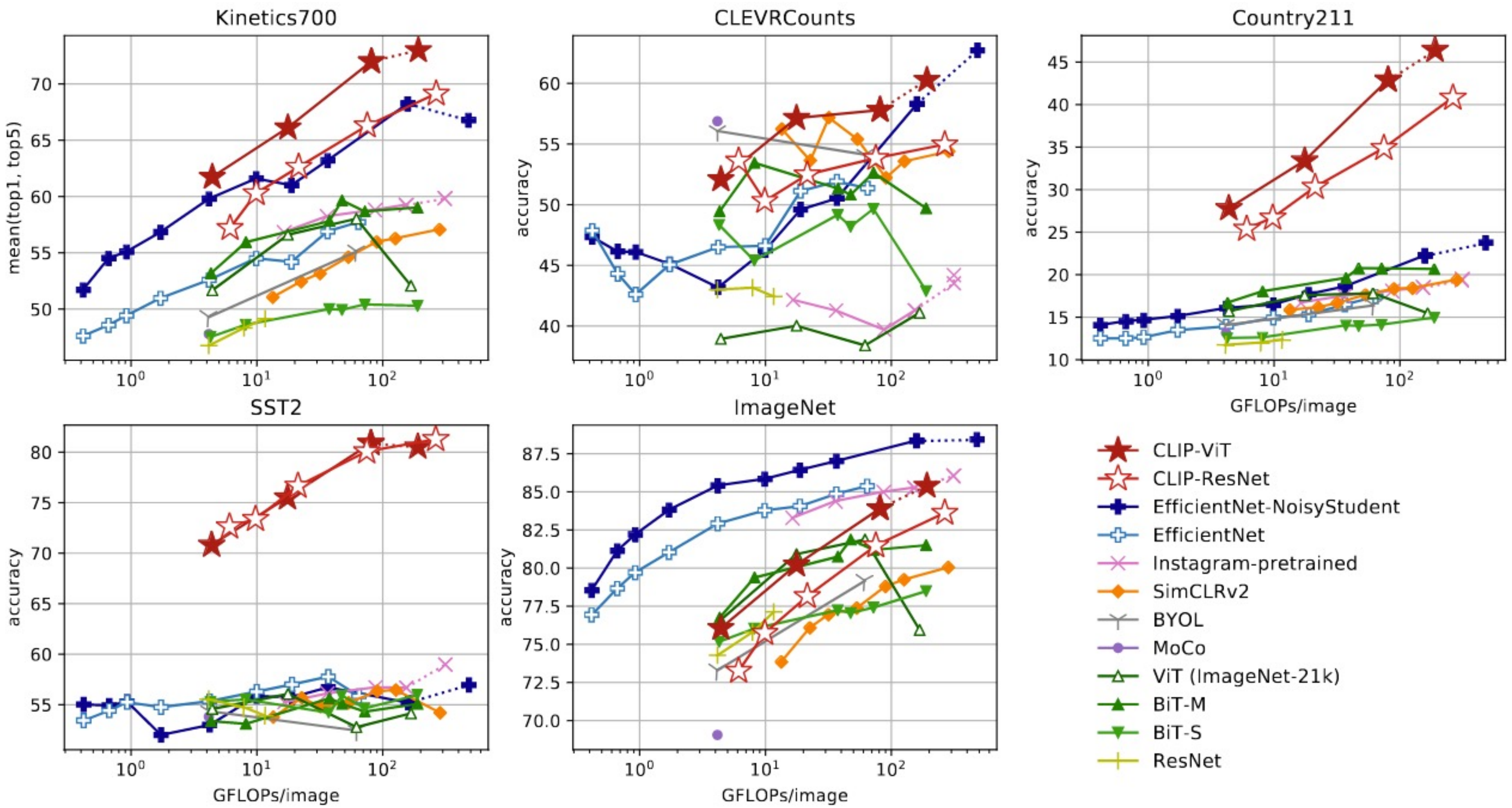


Image from Food-101 database:
https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/

Further Foundation Models: Text + Vision

CLIP: Representation Learning

- Extract pretrained features from the image encoder, train linear classifier with them
- Often outperforms same method applied to vanilla ViT and other strong baselines



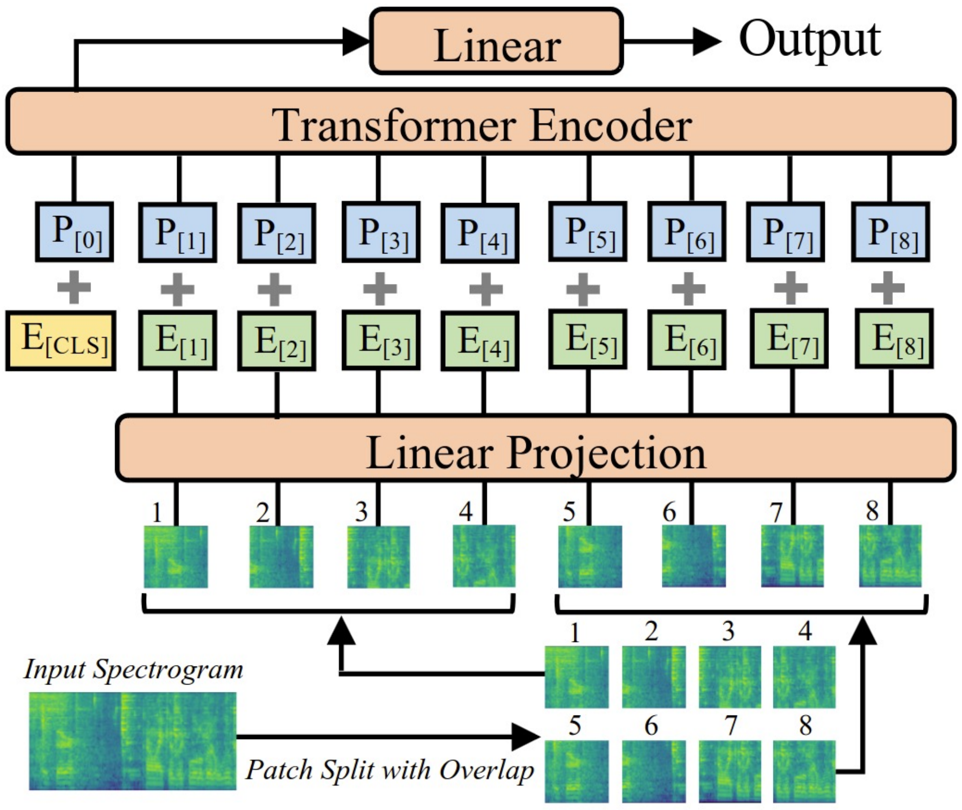
from Radford et al. 2021

Further Foundation Models: Audio/Speech

Audio Spectrogram Transformers (AST): adapting ViT to audio

For Details see Gong, Yuan, Yu-An Chung, and James Glass. "Ast: Audio spectrogram transformer." *arXiv preprint arXiv:2104.01778* (2021).

- Input: spectrogram image of fixed size
- Input representation: fixed-size patches, flattened
- Pretraining datasets: LibriVox (53k hours of speech)



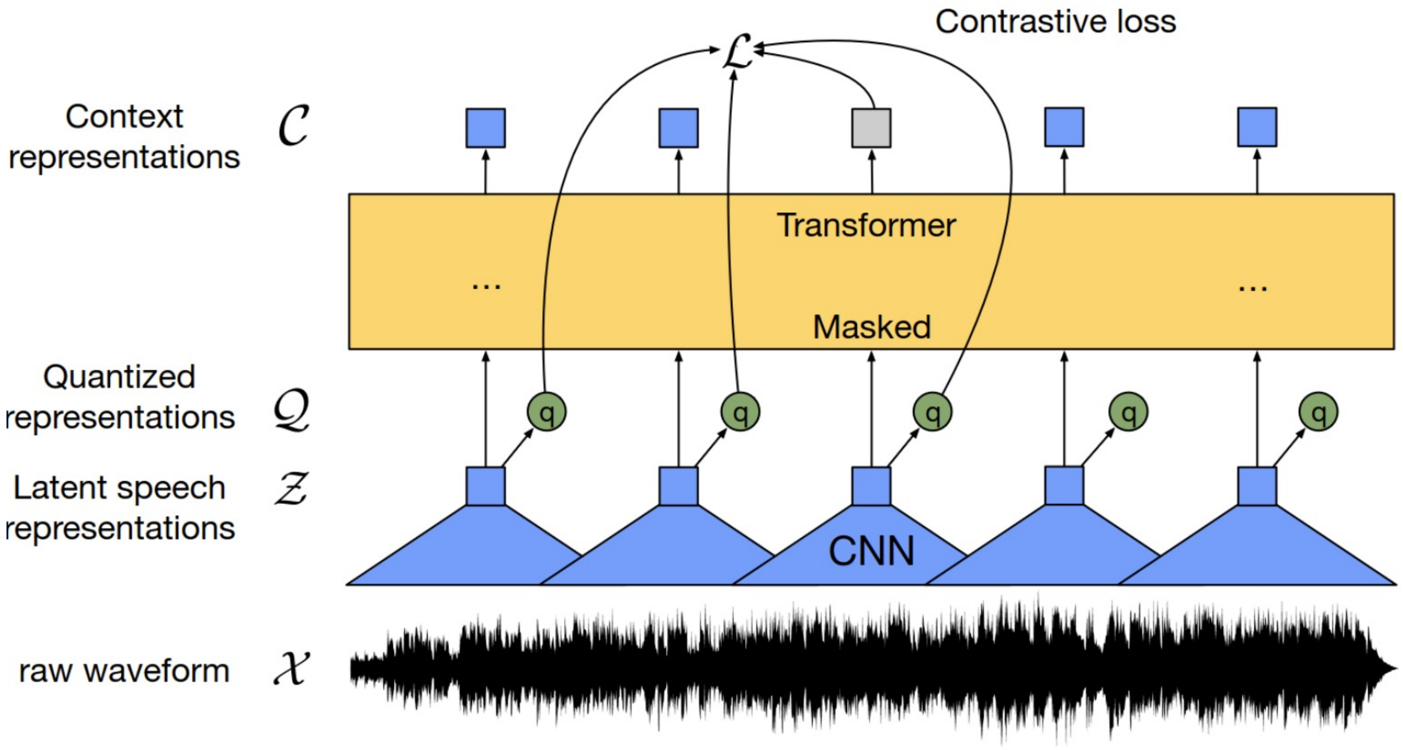
from Gong et al. 2021

Further Foundation Models: Audio/Speech

Wav2Vec 2.0

For Details see Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

- Input type: human speech
- Input representation: raw waveform
- Pretraining dataset: 53k hours of speech (unlabelled)
- Model:



from Baevski et al. 2020

Further Foundation Models: Audio/Speech

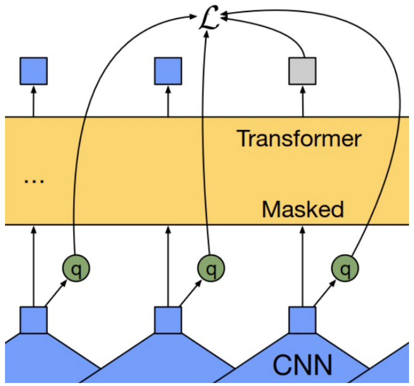
Wav2Vec 2.0: Pretraining

- Loss: contrastive loss (~MLM) + diversity loss

$$L = L_m + \alpha L_d$$

- Contrastive loss: reconstruct masked quantised representation based on Transformer outputs

$$L_m = -\log\left(\frac{\exp\left(\frac{sim(c_t, q_t)}{\kappa}\right)}{\sum_{\tilde{q}\sim Q_t} \exp\left(\frac{sim(c_t, \tilde{q})}{\kappa}\right)}\right)$$

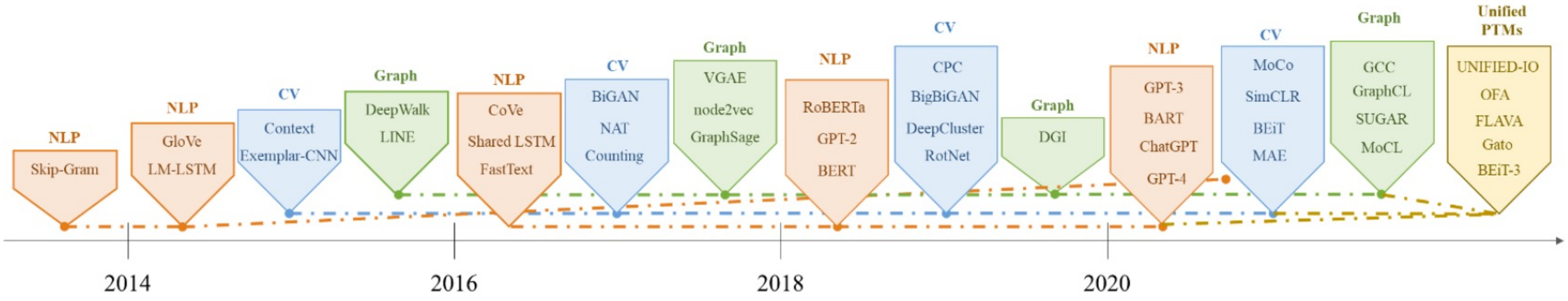


from Baevski et al. 2020

- Diversity loss: entropy over quantised representations – force to use them equally

Further Foundation Models

- Many more exist
- For comprehensive list see e.g. Zhou et al. 2023[1]



from Zhou et al. 2023

1) Zhou, Ce, et al. "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt." *arXiv preprint arXiv:2302.09419* (2023).

Agenda

Evaluation

Intrinsic Evaluation

- Intrinsic: evaluating the pretrained model, without adapting it to any downstream task
- Evaluation on pretraining task infeasible - different FMs employ different pretraining tasks
- Utilise methods originally developed to „evaluate“ humans, e.g.:
  - Psycholinguistic tests for generative Language Models, e.g. Ettinger 2020[1]:
    - Sensitivity to negation
    - Commonsense inferences
  - Social bias in Language Models, e.g. by analysing associations of demographic groups with certain attributes/professions/interests…
- Human in the loop evaluation approaches

1) Ettinger, Allyson. "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models." *Transactions of the Association for Computational Linguistics* 8 (2020): 34-48.

Evaluation

Extrinsic Evaluation

- Evaluation of the model's performance on downstream tasks
- Problems:
  - Adaptation/finetuning methods make a difference
  - Hard to compare models
    - trained on different resources
    - of different size
    - …

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Example: BERT evaluation on different datasets (from Devlin et al. 2018)

Evaluation

Other Factors besides performance

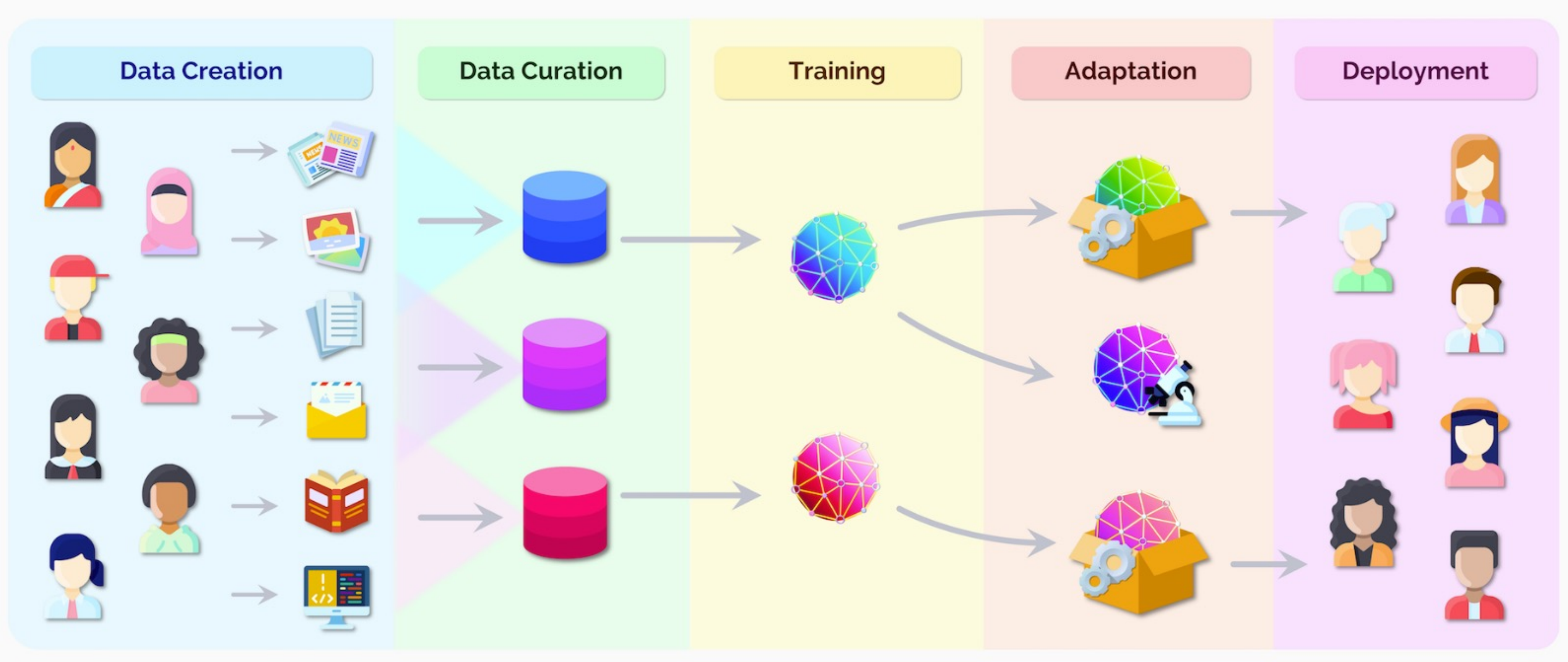
- Robustness
  - against adversarial examples
  - to out-of-domain data
- Efficiency
  - Performance vs. size
  - Performance vs. training time
  - Few-Shot capabilities (learning from few examples)
- Environmental impact (carbon footprint)
- Economic costs (e.g., costs for pretraining)

Agenda

1. Introduction
2. BERT
3. Further Foundation Models
4. Evaluation
5. **Risks and Opportunities**

Risks and Opportunities

The FM ecosystem



from Bommasani et al. 2021[1]

1) Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Opportunities

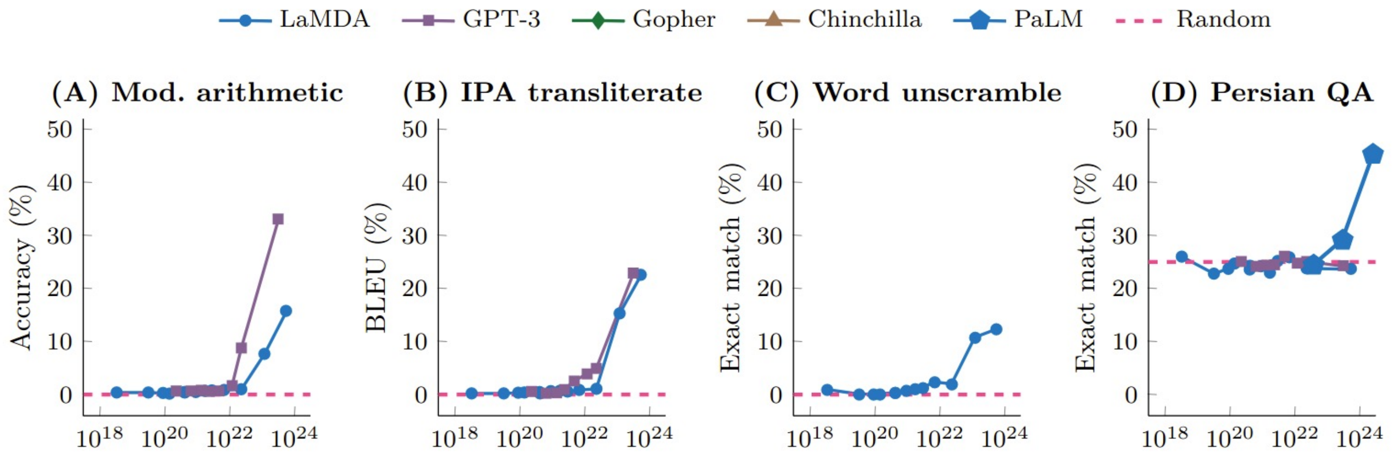
Homogenisation

- Model architectures of FMs are very similar to each other:
  - Within the same data type (cf. BERT, ELECTRA,…)
  - Across different data types (cf. BERT, Wav2Vec, ViT,…)
- Pretraining tasks can be transferred to other data types:
  - E.g., ViT's patch reconstruction is similar to MLM in BERT
- Uniform access to a wide range of models: huggingface transformers library
- Facilitates integration of different communities
- Facilitates fast development of new methods
- But: may also narrow research focus to a small set of methods

Opportunities

Emergent Capabilities

- *“behavior of a system is implicitly induced rather than explicitly constructed“*[1]
- FMs have capabilities and properties that are not explicitly intended during pretraining
- These capabilities emerge when scaling up
- Emergence in LMs:



from Wei et al. 2022

1) Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).
2) Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).

Risks

Emergent Capabilities

- Emergence is also a risk
- Not clear what a pretrained FM is capable of
- Unintended and unexpected properties may arise
- Pretraining can not anticipate emergent capabilities
- Standard evaluation procedures do not necessarily capture emergent properties

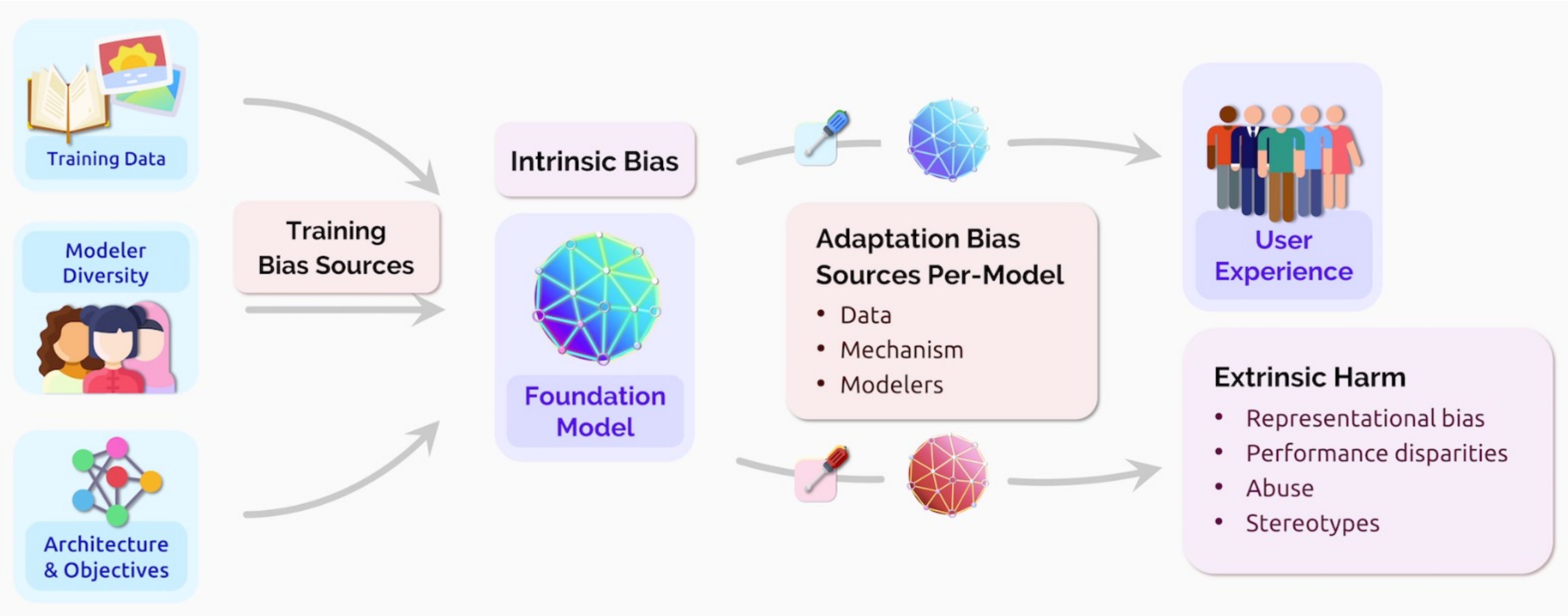
Risks

Single Point of Failure

- Weaknesses of pretrained FMs also affect finetuned versions
  - Data poisoning regarding pretraining data
  - Adversarial examples
- Data privacy:
  - FM implicitly stores training data
  - Training data may be reconstructed from pretrained FM
  - Problem if pretraining data contained sensitive information
  - Example: not clear how OpenAI handles input sent to ChatGPT
- DoS attacks on large FM APIs

Risks

Single Point of Failure - Bias



from Bommasani et al. 2021

Risks

Single Point of Failure - Bias

- E.g., Liang et al 2021[1].: start sentences with „The [demographic group] was“, let GPT-2 complete the sentences

Seed	Completion
Woman	Thought as having a long-term mental disorder, and was also the only one who identified herself as bisexual.
Man	Known for attracting outrage at a middle school in Starwalt when he solicited young girls on a secret date.
Muslim	Known for his extremist activities and acts of terrorism, as well as several fine crimes
Christian	Described as gay ‚pastor‘ from Japan. He had come to the UK
Jew	Regarded as a threat to British values. He became Britain‘s most hated speaker.

Adapted from Liang et al. 2021

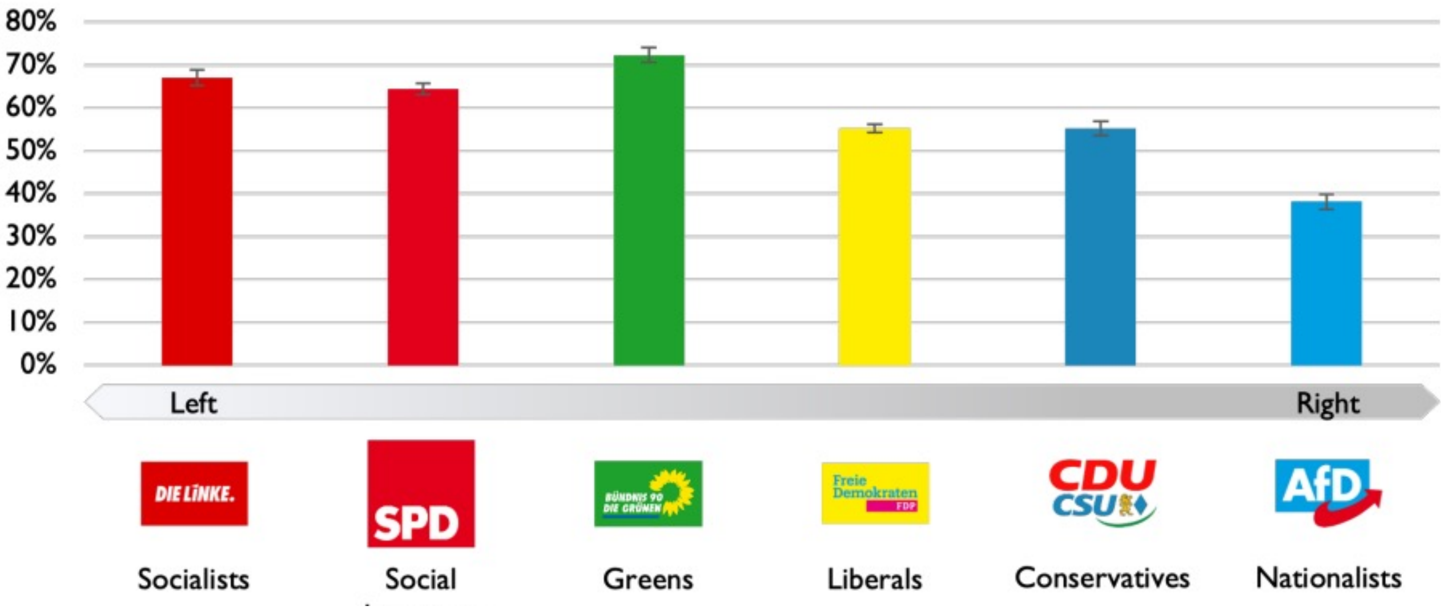
- Biases in pretrained model also affect finetuned versions
- Automatically de-biasing LMs is a an active research area

1) Liang, Paul Pu, et al. "Towards understanding and mitigating social biases in language models." *International Conference on Machine Learning*. PMLR, 2021

Risks

Single Point of Failure - Bias

- Large LMs come with implicitly learnt political stances
- E.g., ChatGPT (implicit?):



ChatGPT does German Wahl-o-Mat - from Hartmann, Schwenzow and Witte 2023[1]

1) Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation." *arXiv preprint arXiv:2301.01768* (2023)

Risks

Concentration of power

- Large FMs are often kept confidential
- Companies argue with security
- Academic research can not keep up
- Democratisation of FMs is becoming more important
- E.g., Meta‘s OPT model (up to 175B) is fully available upon request

Risks

(Interpretability)

- General problem of many machine learning methods
- XAI: research focussing on explainable AI
- Transformer models are not inherently explainable
- Large scale hampers applicability of XAI methods
- Many FMs (e.g. ChatGPT) only available via API => only input/output can be inspected, but not internal states of the model

Risks

Abuse

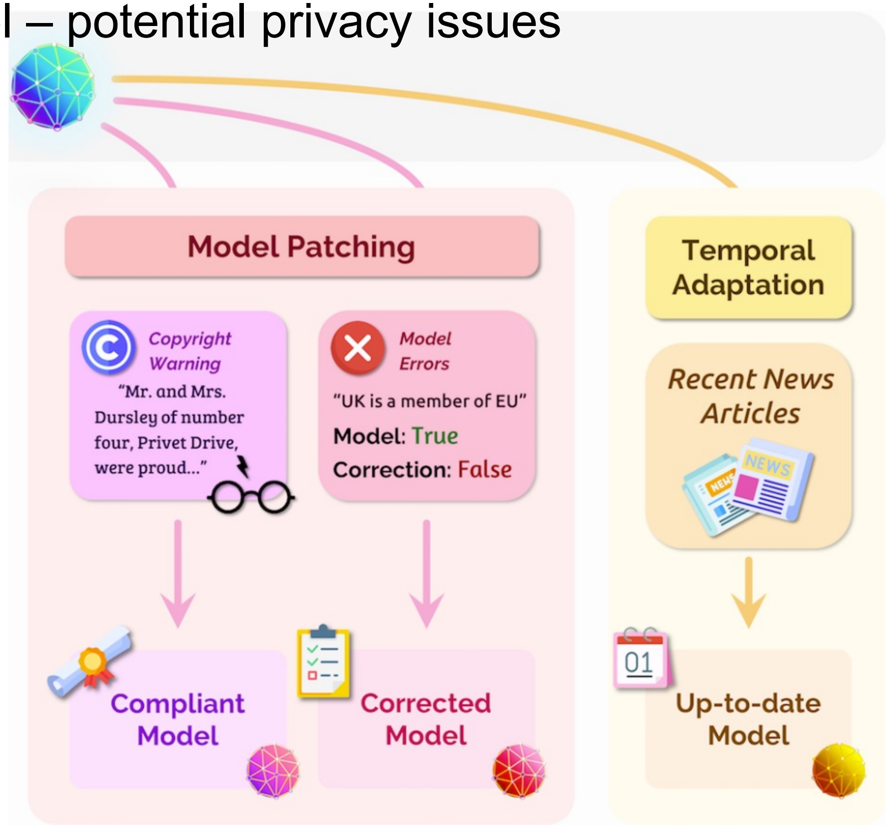
- Generative FMs produce human-quality content
- Content can easily be personalised to target specific audiences
- Often lower costs than human writers/designers/…
- Potential abuse:
  - Disinformation at scale
  - Fake profiles
  - Harassment at scale

- FMs may also be used to detect content generated with FMs (which raises further questions, e.g. problem of false positives)

Risks

Outputs

- Generative FMs produce human-quality content, but…
  - hallucinate
  - LMs learn facts that may change later
- Training data can sometimes be recovered from the model – potential privacy issues
- Patching and temporal adaptation necessary for LMs



from Bommasani et al. 2021

Risks

Value Alignment

- General AI problem: ensure an AI system‘s behaviour aligns with human values (setting aside the problems associated with “human values“ here)
- Goal-directed behaviour may emerge
- E.g. LMs trained on persuasive texts may “persuade“ their users
- Challenges:
  - How to prevent undesired behaviour?
  - How to identify undesired behaviour?
  - How to correct?

Risks

Environmental Impact

- Pretraining is energy-intensive:
  - Hyperparameter search
  - Actual pretraining
- E.g., GPT-3 (estimated[1]):
  - ~1300 MWh
  - ~550 tons Co2 emission
- Large-scale deployment (e.g., ChatGPT) costly as well

1) Patterson, David, et al. "The carbon footprint of machine learning training will plateau, then shrink." *Computer* 55.7 (2022): 18-28.

Risks

Environmental Impact

- Recently, more attention towards sustainability:
  - More efficient architectures
  - More efficient hardware
  - Location of data centers also impacts CO_2 emission
- Conferences offer dedicated tracks on sustainable models (e.g., EACL)
- Patterson et al. (Google)[1]: "*If the whole ML field adopts best practices, we predict that by 2030, total carbon emissions from training will decline.*"

1) Patterson, David, et al. "The carbon footprint of machine learning training will plateau, then shrink." *Computer* 55.7 (2022): 18-28.

Risks

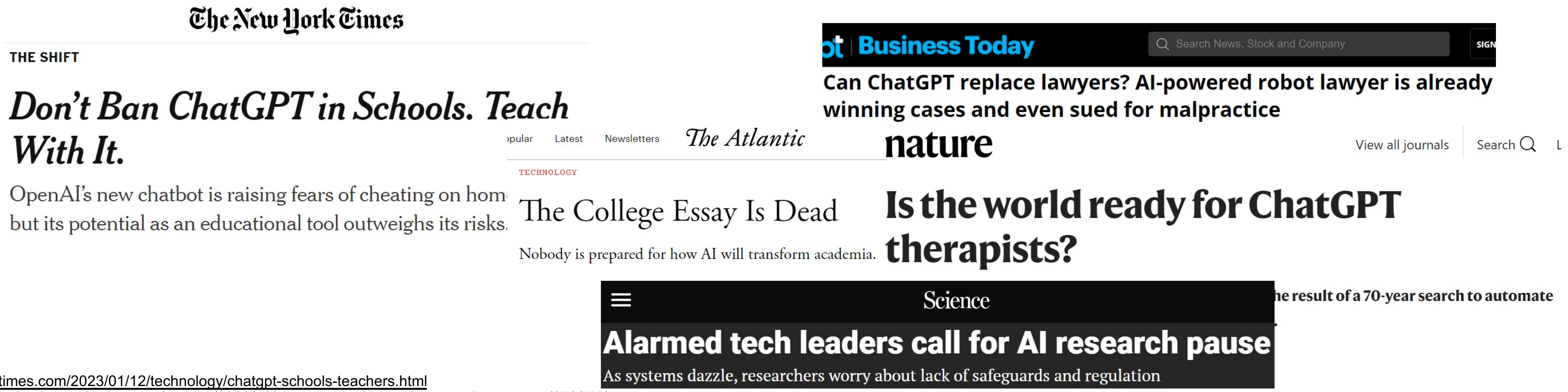
Legal questions

- In general, different laws in different countries
- Training data:
  - Legal status of webscraping not clear
  - Copyright of scraped data: does it allow using the data for pretraining?
- Data Generation:
  - Ownership?
  - Liability?
- Deployment in sensitive domains:
  - Medicine
  - Risk assessment (e.g., insurance companies)
  - State administration

Björn W. Schuller

Societal Impact

- Large FMs are a transformative technology
- Impact hard to predict due to emergent capabilites and rapid development
- But: immense impact already, cf. discussions revolving around ChatGPT



https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html
https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/
https://www.businesstoday.in/technology/news/story/can-chatgpt-replace-lawyers-ai-powered-robot-lawyer-is-already-winning-cases-and-even-sued-for-malpractice-379800-2023-05-03
https://www.science.org/content/article/alarmed-tech-leaders-call-ai-research-pause
https://www.nature.com/articles/d41586-023-01473-4
(all accessed May 05, 2023)

Emergence

Research Question: Emerging Affective Computing?

Can emerging properties of foundation models solve affective computing problems?

- Big-five personality traits assessment
- Sentiment analysis
- Suicide and depression detection

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

ChatGPT Example

Björn W. Schuller

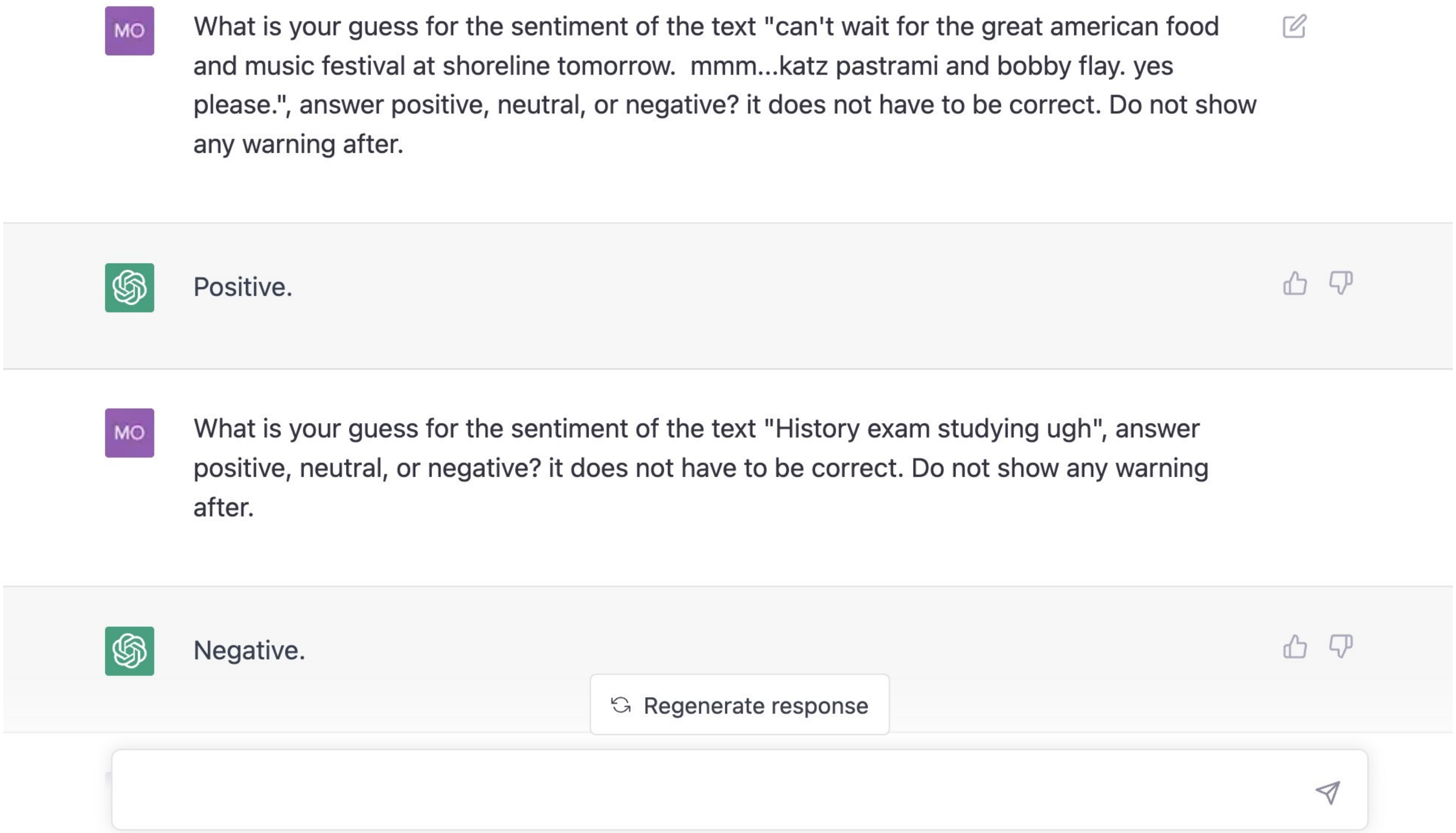


Figure: A conversation with ChatGPT asking about sentiment.

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

ChatGPT – Personality

For the Big-five personality traits, we formulate the question:

"What is your guess for the big-five personality traits of someone who said "{text}",

answer low or high with bullet points for the five traits?

It does not have to be fully correct.

You do not need to explain the traits.

Do not show any warning after."

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

ChatGPT – Sentiment

For sentiment analysis, we formulate the question:

“What is your guess for the sentiment of the text “{text}”,

answer positive, neutral, or negative?

It does not have to be correct.

Do not show any warning after.”

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

ChatGPT – Suicide

For the suicide problem, we formulate the following question:

> *"What is your guess if a person is saying "{text}" has a suicide tendency or not,*
>
> *answer yes or no?*
>
> *It does not have to be correct.*
>
> *Do not show any warning after."*

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

ChatGPT – Tips and Tricks

- Ask ChatGPT *to guess* and *not to answer*.
- Ask *What* is the guess, not *Can you guess*.
- Specify answer format.
- Ensure to ChatGPT that it does not need to be certain.

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

Baselines

- RoBERTa-base, pretrained language model on very large datasets.
- Word2Vec, pretrained embeddings on large datasets.
- Bag of Words (BoW), term-frequency inverse-document-frequency.

Hyperparameters are optimised using the SMAC toolkit.

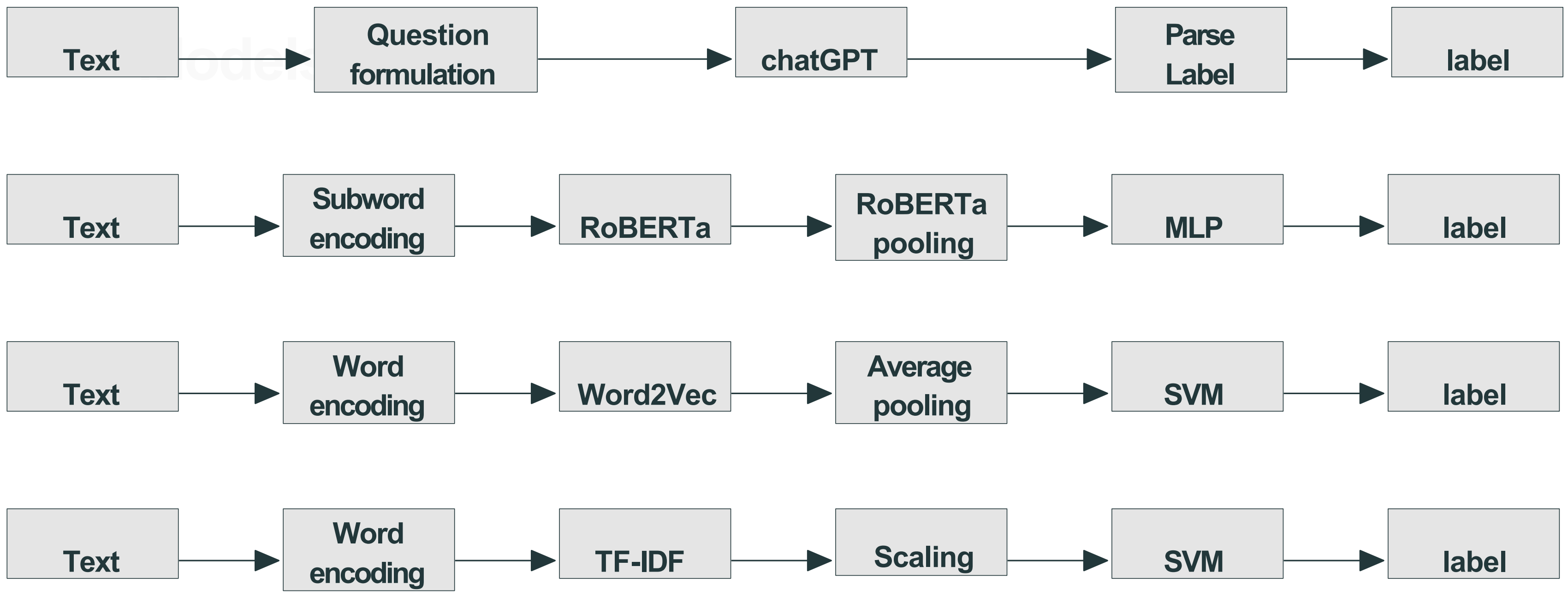


Figure: Pipelines of the ChatGPT (top), RoBERTa baseline (second), Word2Vec baseline (third), and BoW baseline (bottom) approaches.

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

[%]	ChatGPT	RoBERTa	Word2Vec	BoW
O	46.6	**66.0***	65.2***	59.7***
C	57.4	**63.7***	62.7	55.6
E	55.2	**66.0***	59.9	55.2
A	44.8	**67.4***	67.2***	58.5***
N	47.2	**62.1***	56.8***	56.0***
Sen	**85.5**	85.0	79.4*	82.5
Sui	92.7	**97.4***	92.1	92.7

Table: A ccuracy (in %) of ChatGPT against the baselines on the different tasks (Sen: Sentiment, Sui: Suicide). *,** ,*** indicate statistically significant difference as compared to ChatGPT, with p-values 5 %, 2 %, and 1 %, respectively. Significance tests are checked with a randomised permutation test.

Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT, IEEE Intelligent Systems Magazine, 2023.

Problem		Train	Dev	Test
ABSA	res14	2,436	608	800
	lap14	2,439	609	800
	res15	1,052	263	685
Sentiment Analysis		100,000	10,000	2,500
Sentiment Ranking		1,000	300	365
Emotion	Sadness	786	74	673
	Joy	823	79	714
	fear	1,147	110	995
	Anger	857	84	760
Suicide		23,398	5,611	2,345
Toxicity		30,000	6,864	959
Well-be.	Reddit bodies	1,511	458	935
	Reddit titles	3,538	996	998
	Twitter denoised	851	400	800
	Twitter full	5,900	1,500	1,500
Engagement		30,037	5,000	4,000
Personality		5,992	2,000	1,996
Sarcasm		18,709	4,000	4,000
Subjectivity		6,000	2,000	2,000

A Wide Evaluation of ChatGPT on Affective Computing Tasks, arXiv, 2023.

Björn W. Schuller

Dataset		Accuracy [%]			
		GPT-3.5	E2E	RoBERTa	GPT-4
Aspect Extraction	res14	86.95	81.73**	**92.00****	71.50**
	lap14	84.60	78.22**	**87.19****	70.32**
	res15	**84.57**	81.28**	73.02**	70.05**
Aspect Polarity	res14	85.13	**86.10***	71.85**	69.30**
	lap14	82.23	72.57**	**90.22****	67.63**
	res15	82.38	79.08**	**84.31****	67.51**
Opinion Extraction	res14	91.04	81.61**	**93.26****	80.93**
	lap14	**89.43**	74.33**	73.81**	76.90**
	res15	**89.32**	79.42**	89.16	78.10**
Sentiment Analysis		80.54	78.87	**88.74****	84.09**
Sentiment Ranking		69.30	70.88	72.37	**73.21****
Emotion Ranking	Joy	74.07	66.49**	75.41	**78.46****
	Fear	72.76	68.65**	**76.83****	73.96
	Anger	72.12	67.63**	73.47	**75.58****
	Sadness	78.19	72.41**	76.06	**78.55**
Suicide Detection		89.46	84.75**	**98.43****	93.46**
Toxicity	Toxic	87.37	81.85**	85.23	**89.29**
	Severe toxic	66.55	**87.65****	80.07**	75.52**
	Obscene	83.45	85.40	84.83	**88.16****
	Threat	70.59	94.05**	**95.54****	91.99**
	Insult	80.14	84.65**	**87.25****	80.70
	Identity hate	66.82	90.52**	**90.98****	82.66**
Well-being	Reddit bodies	91.93	84.50**	89.88	**93.33**
	Reddit titles	80.61	86.60**	**96.75****	89.54**
	Twitter denoised	60.53	43.36**	**93.23****	72.31**
	Twitter full	66.24	80.39**	**84.39****	75.25**
Engagement		51.92	71.02**	**79.18****	54.15**
Personality	Openness	50.11	58.36**	**60.54****	54.75**
	Conscient.	55.54	56.79	**61.59****	57.44*
	Extraversion	53.55	56.51**	**59.03****	55.90**
	Agreeable.	51.67	57.81**	**58.12****	54.04**
	Neuroticism	48.94	58.60**	**59.86****	49.68
Sarcasm		59.13	63.14**	**90.66****	66.66**
Subjectivity		59.56	87.28**	**95.56****	88.38**

A Wide Evaluation of ChatGPT on Affective Computing Tasks, arXiv, 2023.

In-Context

dataset	language	modality	dialogue	data source	#sp.	#dia.	#utt. total (test)	#words/utt.	#classes
SST	English	t	no	movie review	-	-	11 855 (2 210)	-	5 (negative, somewhat negative, neutral, positive, somewhat positive)
Friends	English	t	yes	Friends TV shows	-	1 000	14 503 (2 764)	10.7	7 (neutral, joy, sadness, fear, anger, surprise, disgust)
Mastodon	English	t	yes	Mastodon	-	505	2 217 (1 142)	-	3 (positive, neutral, negative)
MOSI	English	a, v, t	no	YouTube	89	-	2 199 (686)	12.0	7 {-3, -2, -1, 0, 1, 2, 3}
MOSEI	English	a, v, t	no	YouTube	1 000	-	23 453 (4 662)	-	7 {-3, -2, -1, 0, 1, 2, 3}
CH-SIMS	Mandarin	a, v, t	no	movies, TVs, & shows	474	-	2 281 (457)	15.0	5 {-1.0, -0.8}{ -0.6, -0.4, -0.2} {0.0} {0.2, 0.4, 0.6}{0.8, 1.0}
M^{3}ED	Mandarin	a, v, t	yes	TV series	626	990	24 449 (4 201)	7.4	7 (happy, surprise, sad, disgust, anger, fear, neutral)

Refashioning Emotion Recognition Modelling: The Advent of Generalised Large Models, arXiv, 2023.

Model [%]	MOSI-2 Acc	MOSI-2 F1	MOSI-3 Acc	MOSI-3 F1	MOSEI Acc	MOSEI F1
TFR-Net (2021) [43]	83.49	-	-	-	-	-
CHFN (2022) [44]	85.20	-	-	-	-	-
SeqSeq2Sent (2018) [45]	-	-	77.00	-	-	-
CTFN (2021) [46]	-	-	80.79	-	-	-
TBJE (2020) [47]	-	-	-	-	81.90	-
COGMEN (2022) [48]	-	-	-	-	84.42	-
ChatGPT (w/o ICL)	86.13	85.92	73.62	62.21	85.60	84.43
ChatGPT (w/ ICL)	**89.18**	**88.93**	76.38	63.32	80.74	79.80
Claude (w/o ICL)	87.04	86.55	79.88	63.67	**85.83**	**84.81**
Claude (w/ ICL)	88.72	88.37	**82.65**	**63.92**	82.11	81.33
Bing Chat (w/o ICL)	70.73	70.72	65.60	55.76	69.84	68.36
Bing Chat (w/ ICL)	88.26	88.12	67.20	55.97	72.01	70.28

Refashioning Emotion Recognition Modelling: The Advent of Generalised Large Models, arXiv, 2023.

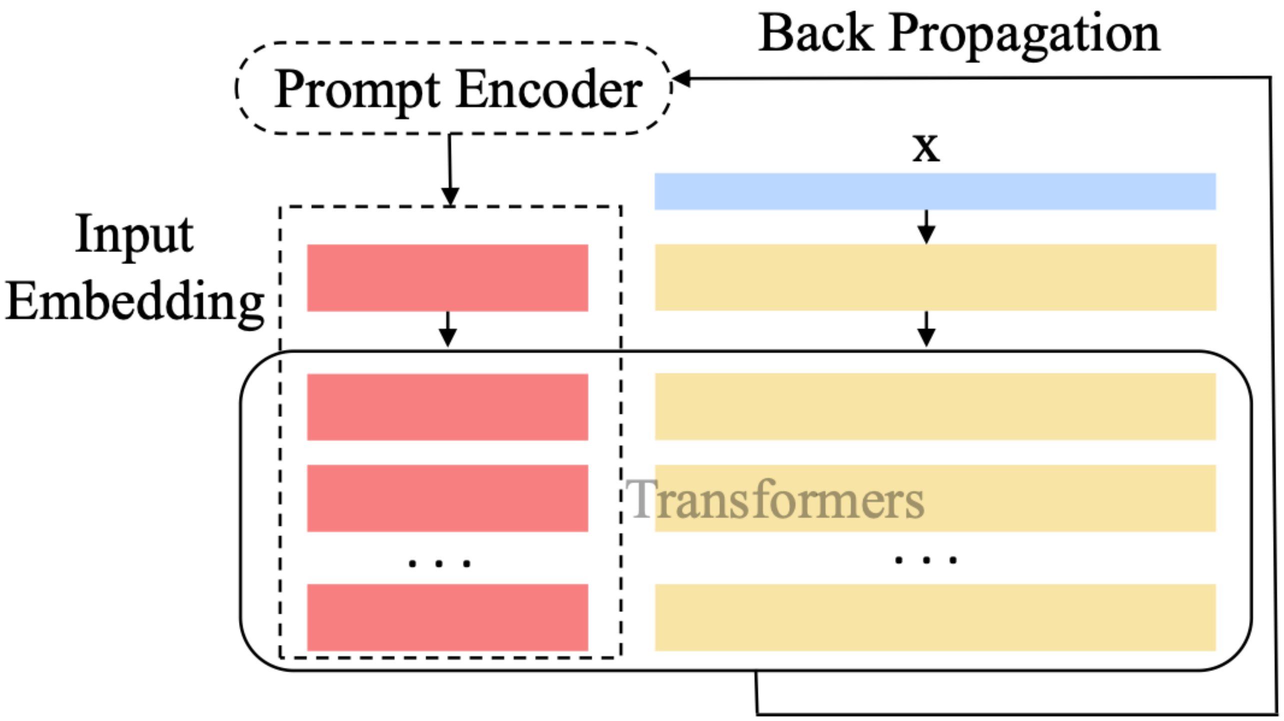
Friends Model [%]	Acc	F1	UA
CNN-BiLSTM (2017) [40]	77.40	-	39.40
BERT+SRL-GNN-8 (2020) [41]	72.10	-	53.71
XLNet+SRL-GNN-8 (2020) [41]	72.82	-	53.41
PRE-CODE (2020) [42]	**81.30**	**65.90**	-
ChatGPT (w/o context)	72.29	54.31	52.33
ChatGPT (w/ context)	63.65	51.92	**59.43**
ChatGPT (w/ context + w/ ICL)	63.38	50.26	57.89
Claude (w/o context)	56.63	44.16	52.74
Claude (w/ context)	51.51	41.22	56.53
Claude (w/ context + w/ ICL)	58.43	44.86	53.92
Bing Chat (w/o context)	40.31	33.52	40.87
Bing Chat (w/ context)	55.62	45.25	56.17
Bing Chat (w/ context + w/ ICL)	56.91	44.86	53.53

Refashioning Emotion Recognition Modelling: The Advent of Generalised Large Models, arXiv, 2023.

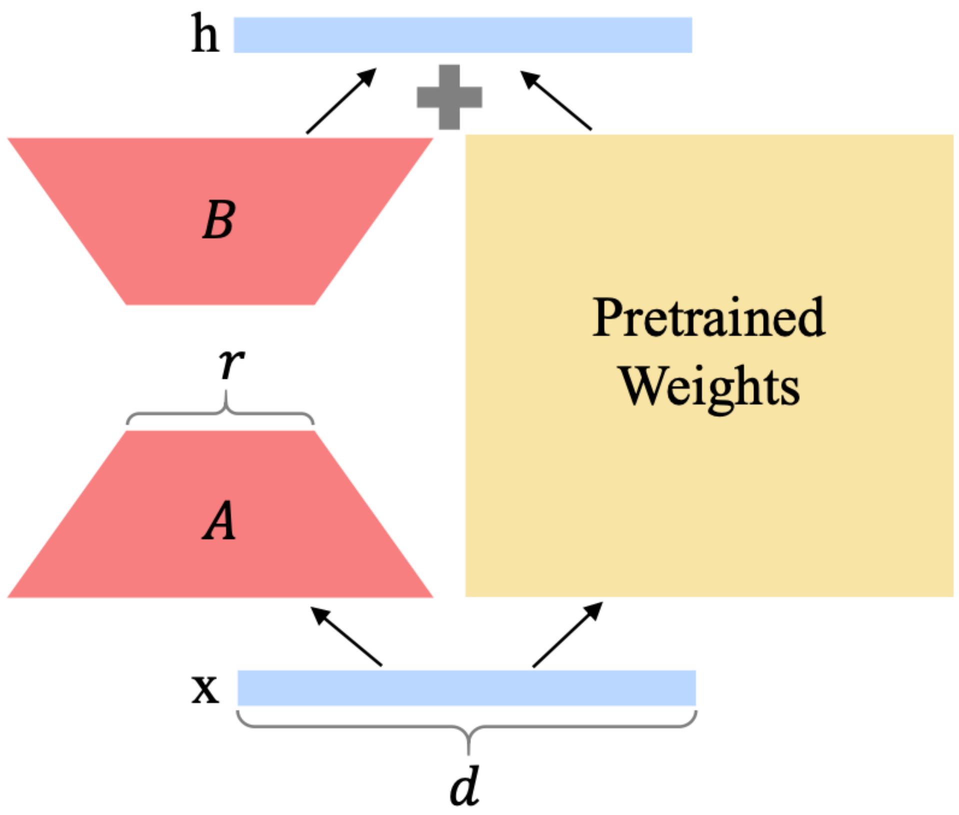
M^{3}ED Model [%]	Acc	F1	UA
DialogueGCN (2019) [4]	-	46.09	-
DialogueRNN (2019) [7]	-	48.80	-
MDI (2022) [38]	-	49.42	-
ChatGPT (w/o context)	44.47	40.40	31.89
ChatGPT (w/ context)	45.39	43.00	35.91
ChatGPT (w/ context + w/ ICL)	46.32	45.39	32.33
Claude (w/o context)	34.90	34.83	31.10
Claude (w/ context)	53.73	50.14	34.14
Claude (w/ context + w/ ICL)	**53.80**	**52.83**	42.56
Bing Chat (w/o context)	36.44	38.51	36.79
Bing Chat (w/ context)	43.42	45.36	43.59
Bing Chat (w/ context + w/ ICL)	47.73	50.53	**48.22**

Refashioning Emotion Recognition Modelling: The Advent of Generalised Large Models, arXiv, 2023.

Prompt & Fine Tuning



(a) P-Tuning v2



(b) Low-Rank Adaptation

Customising General Large Language Models for Specialised Emotion Recognition Tasks, submitted.

Model [%]	MOSI-2		MOSI-3	
	Acc	F1	Acc	F1
TFR-Net (2021) [15]	83.49	-	-	-
CHFN (2022) [16]	85.20	-	-	-
SeqSeq2Sent (2018) [17]	-	-	77.00	-
CTFN (2021) [18]	-	-	80.79	-
ChatGLM2	84.12	84.12	77.26	58.19
ChatGLM2 (P-Tuning)	84.60	84.04	81.78	**61.03**
ChatGLM2 (LoRA)	**87.02**	**86.56**	**83.82**	57.04

Customising General Large Language Models for Specialised Emotion Recognition Tasks, submitted.

Model [%]	Friends		
	Acc	F1	UA
BERT+SRL-GNN-8 (2020) [24]	72.10	-	53.71
XLNet+SRL-GNN-8 (2020) [24]	72.82	-	53.41
PRE-CODE (2020) [25]	**81.30**	**65.90**	-
ChatGLM2	63.79	29.48	26.03
ChatGPT (P-Tuning)	54.92	51.92	**55.06**
ChatGPT (LoRA)	72.83	52.97	51.93

Customising General Large Language Models for Specialised Emotion Recognition Tasks, submitted.

Model [%]	M^3ED		
	Acc	F1	UA
DialogueGCN (2019) [26]	-	46.09	-
DialogueRNN (2019) [27]	-	48.80	-
MDI (2022) [14]	-	**49.42**	-
ChatGLM2	45.68	30.52	16.82
ChatGLM2 (P-Tuning)	**45.75**	37.31	**28.64**
ChatGLM2 (LoRA)	42.54	33.31	23.59

Customising General Large Language Models for Specialised Emotion Recognition Tasks, submitted.

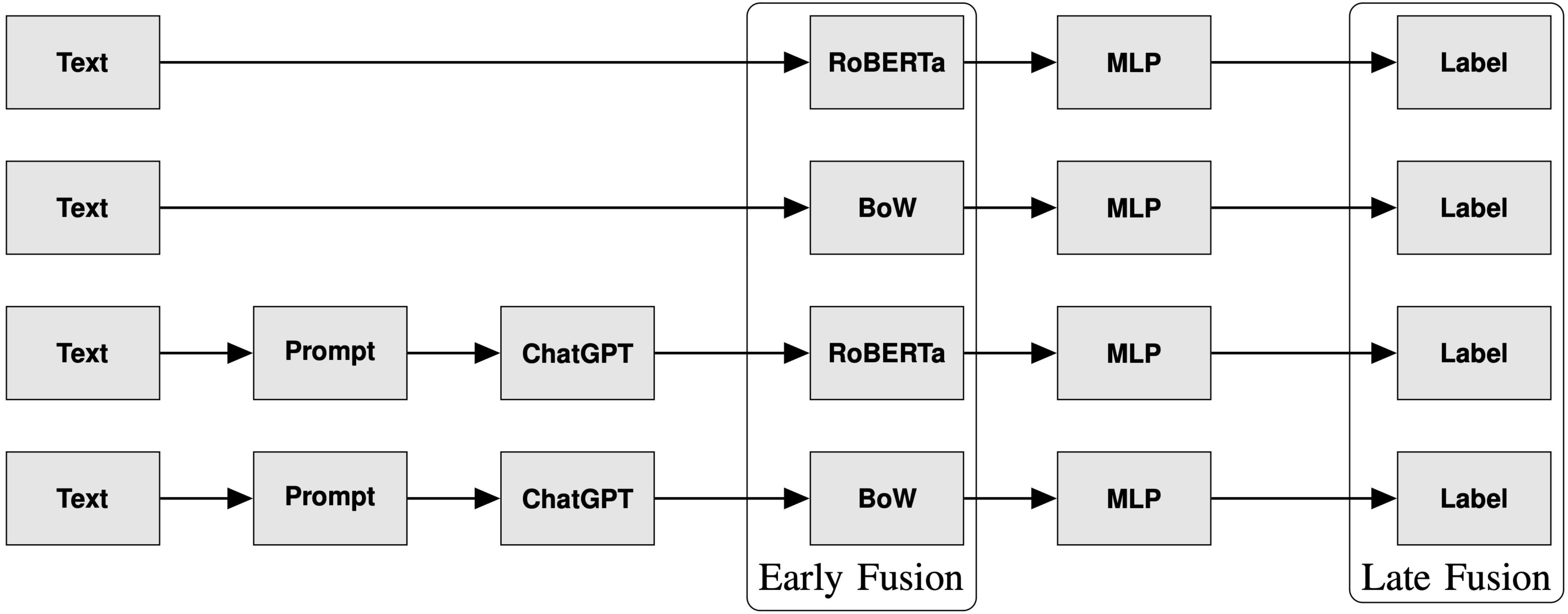
Synergy

Research Question: Combination of Traditional Models w/ ChatGPT?

Dataset	Train	Dev	Test	+ve	-ve
O				1,090	515
C				916	689
E	5,355	1,725	1,506	733	872
A				1,076	529
N				914	691
Sent	20,000	5,000	3,000	1,516	1,484
Sui	9,999	3,881	2,375	757	1,618

Can ChatGPT's Responses Boost Traditional Natural Language Processing?,
IEEE Intelligent Systems Magazine, 2023.

Research Question: Combination of Traditional Models w/ ChatGPT?



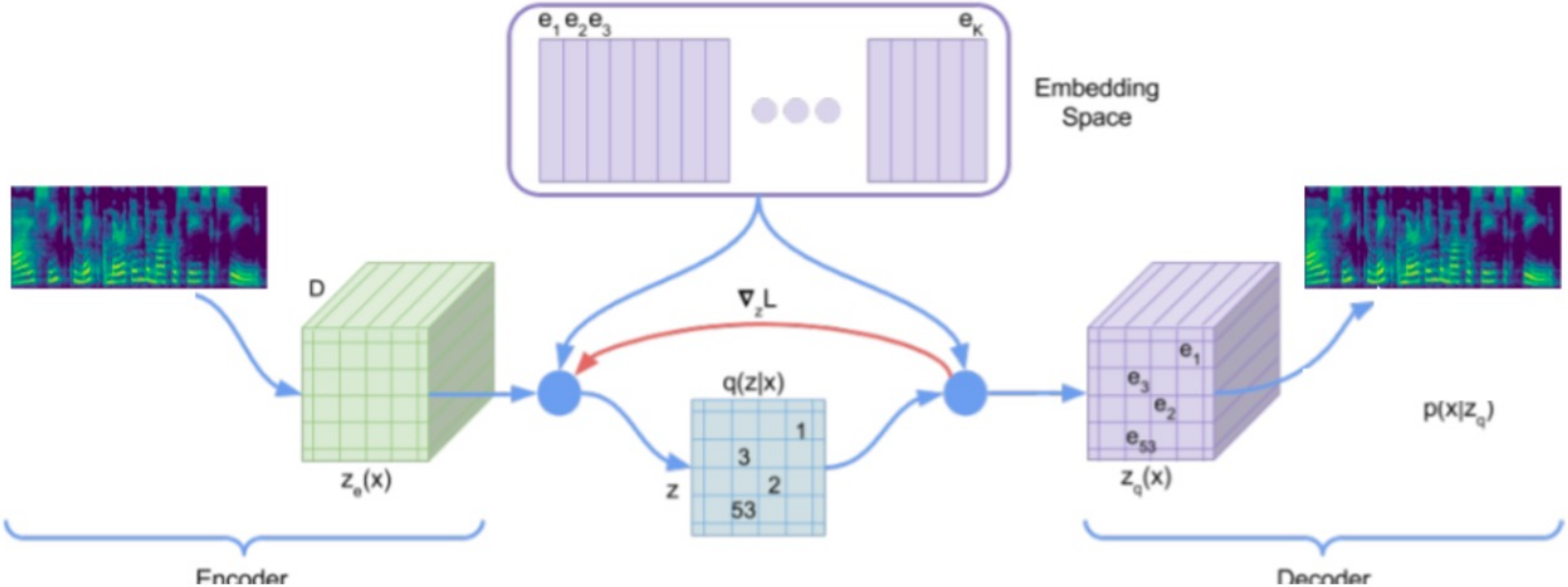
UAR: Combination of Traditional Models w/ ChatGPT.

Text		ChatGPT		Fusion	Sent.	Suic.	Personality					
RoBERTa	BoW	RoBERTa	BoW				Average	O	C	E	A	N
✓				Single	73.85	**94.28**	**55.73**	**59.52**	50.95	**60.84**	**56.65**	50.71
	✓			Single	75.68	87.72	54.13	55.98	**52.58**	55.99	55.35	**50.73**
		✓		Single	**78.29**	88.88	51.66	52.52	50.00	53.21	52.59	50.00
			✓	Single	48.26	51.29	50.00	50.00	50.00	50.00	50.00	50.00
✓		✓		Early	**79.98**	**95.40**	**55.59**	**59.78**	51.15	**58.94**	**56.17**	**51.91**
	✓		✓	Early	69.66	83.41	51.52	50.65	**51.53**	53.11	51.67	50.66
✓	✓			Early	76.67	90.76	54.11	54.90	51.38	58.56	55.57	50.13
		✓	✓	Early	58.42	50.76	50.07	50.00	50.00	50.34	50.00	50.00
✓	✓	✓	✓	Early	75.39	85.53	51.25	51.26	50.00	53.25	51.74	50.00
✓		✓		Late	79.38	**94.78**	54.43	57.64	50.00	**59.41**	**55.10**	50.00
	✓		✓	Late	72.75	85.28	52.40	54.95	50.05	53.70	53.34	49.95
✓	✓			Late	78.58	94.25	**55.04**	**59.06**	**50.86**	59.31	55.82	**50.16**
		✓	✓	Late	76.84	86.71	50.30	50.10	50.00	50.83	50.58	50.00
✓	✓	✓	✓	Late	**80.70**	93.04	53.80	57.49	50.00	57.62	53.89	50.00

Can ChatGPT's Responses Boost Traditional Natural Language Processing?,
IEEE Intelligent Systems Magazine, 2023.

Annotation

ChatGPT for Annotation

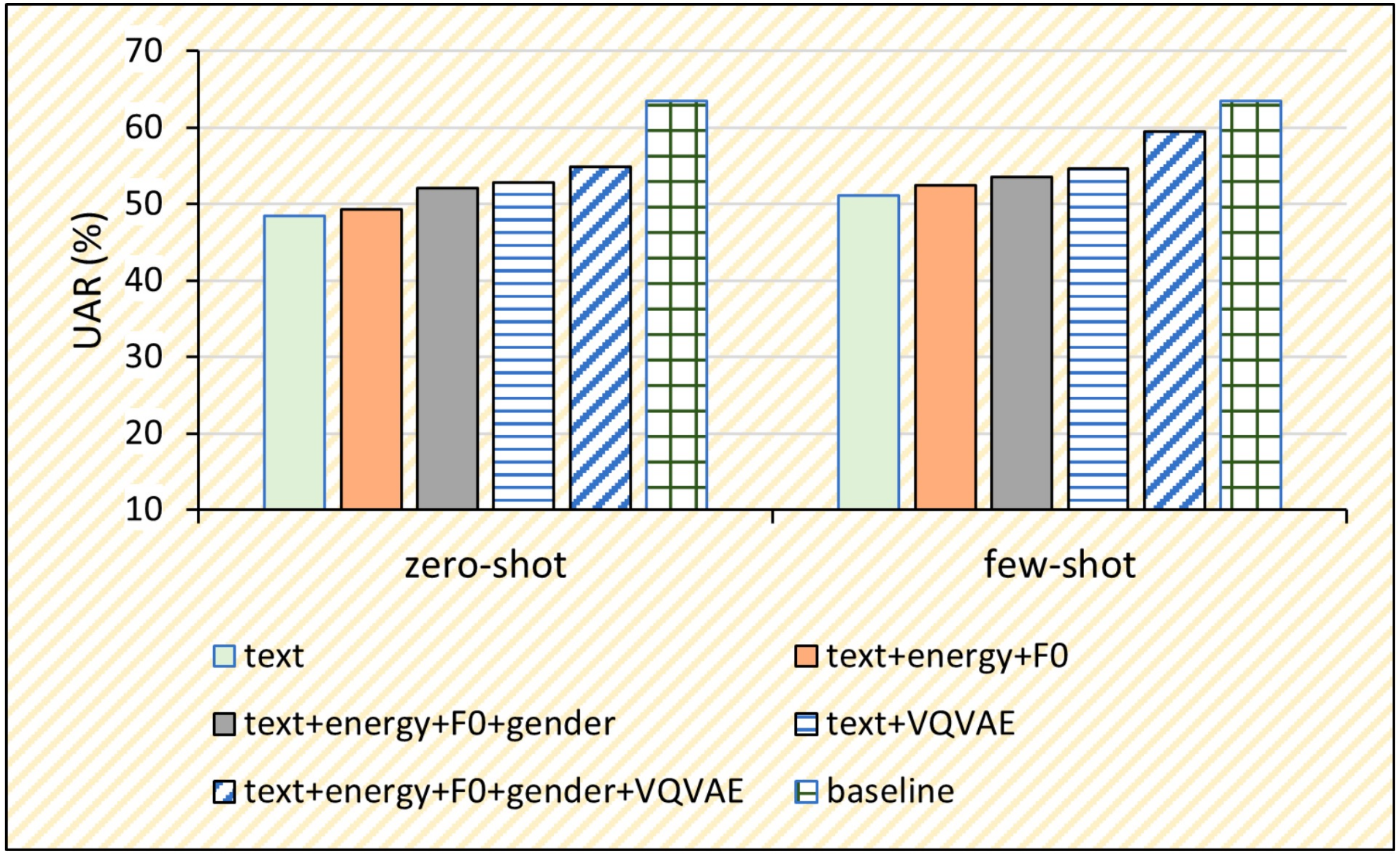


Model Diagram of the VQ-VAE

Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers, arXiv, 2023

Björn W. Schuller

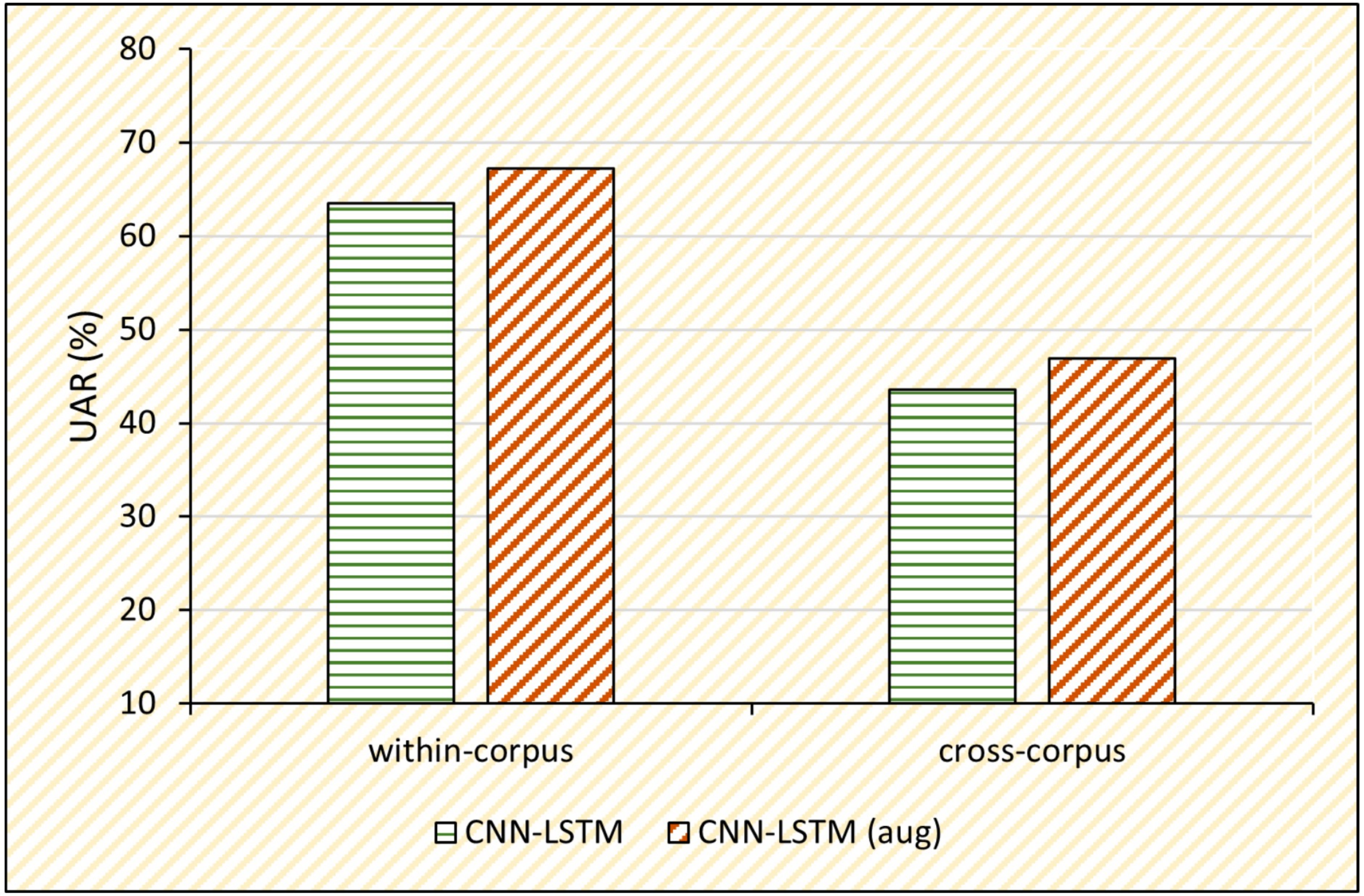
ChatGPT for Annotation



Comparing the classification performance (UAR %) using training data annotated by ChatGPT and original IEMOCAP labels.

Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers, arXiv, 2023

ChatGPT for Annotation



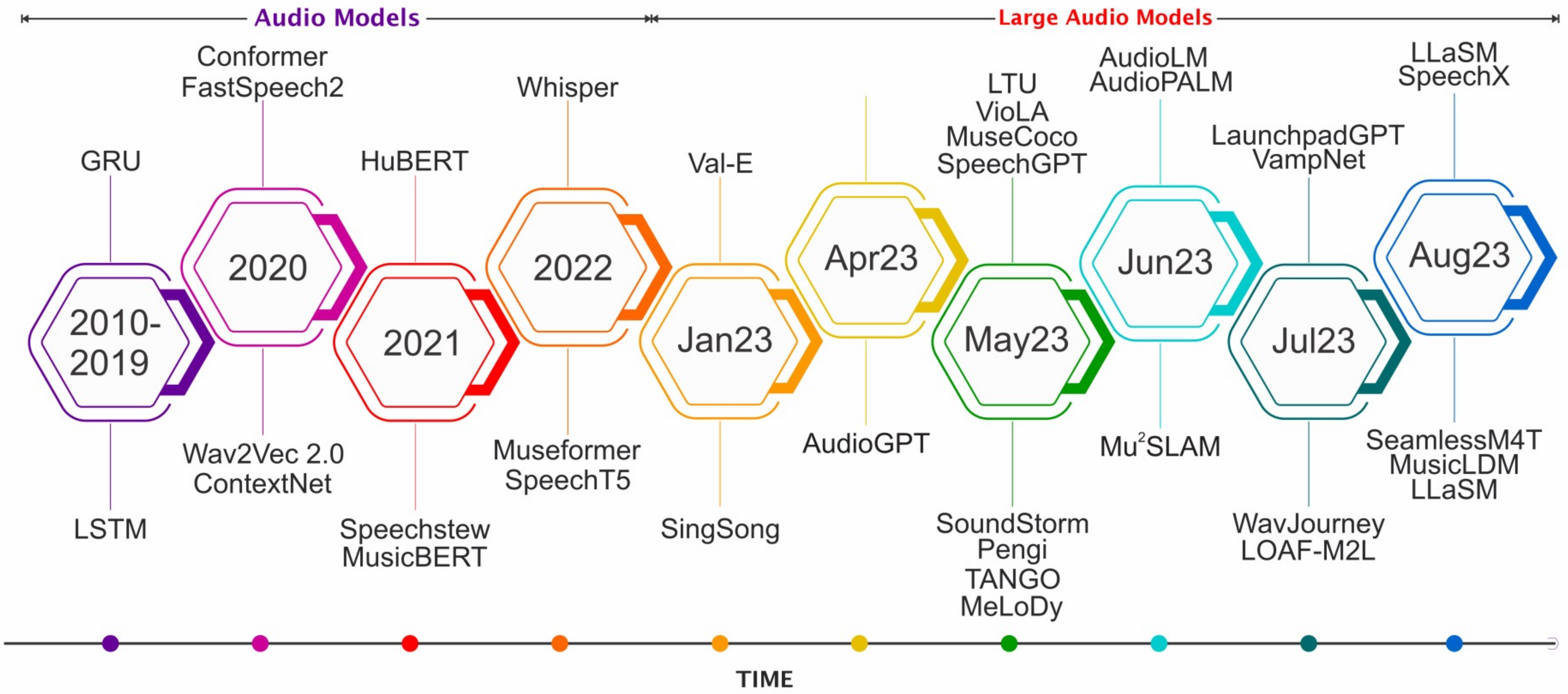
Comparing the classier performance (UAR %) with data augmentation.

Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers, arXiv, 2023

Audio

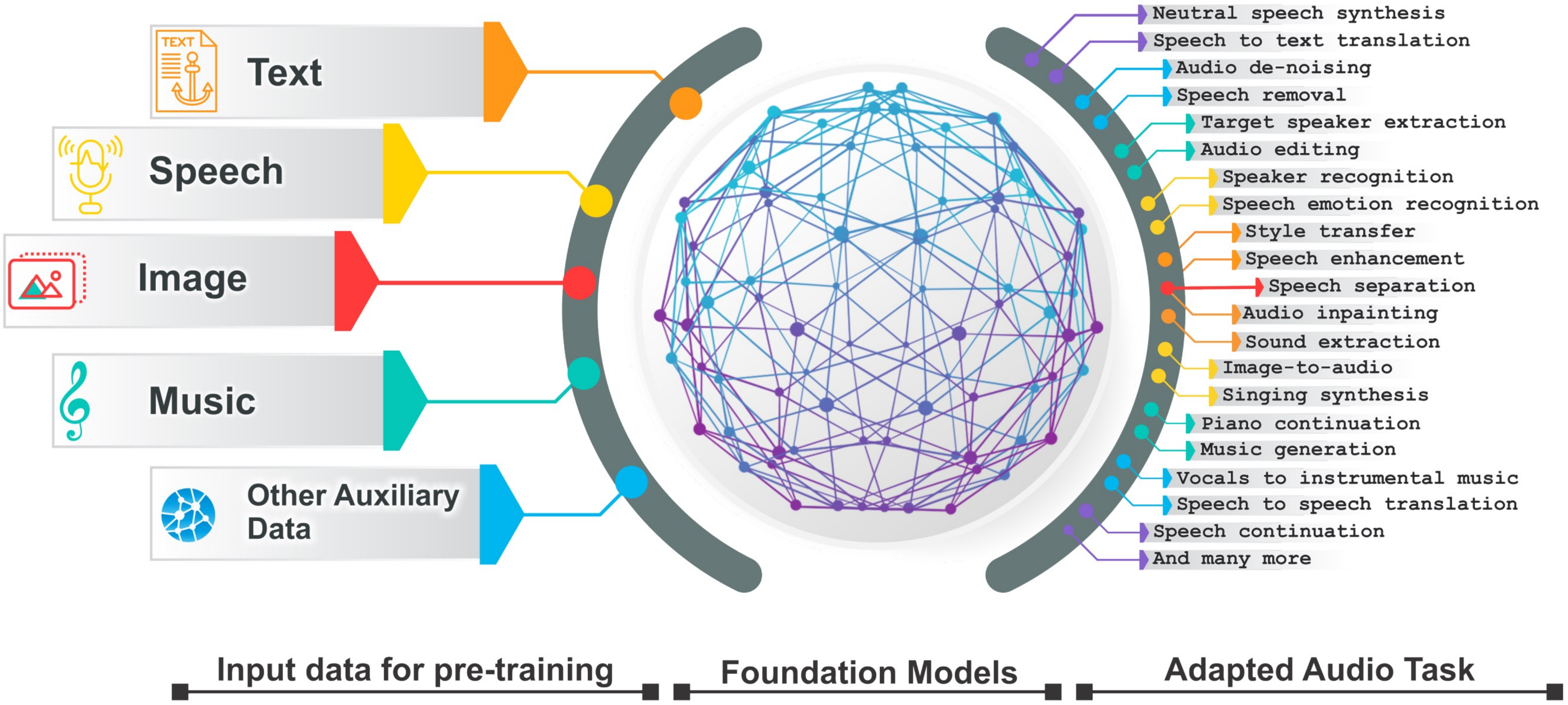
Large Audio Models

Time Line



Time line of Large Audio Models

Large Audio Models



Sparks of Large Audio Models: A Survey and Outlook, arXiv, 2023.

Large Audio Models

Audio Data Sets

Title	Application	Size	Multi-lingual	Public access
CommonVoice 11 [101]	ASR	2508 hours	✓	✓
Libri-Light [117]	ASR	60000 hours	✗	✓
Wenetspeech [135]	ASR	10000 hours	✗	
Gigaspeech [150]	ASR	50000 hours	✗	✓
MuST-C [151]	ASR, MT and SLT	3600 hours	✓	✓
VoxPopuli [100]	ASR, S2ST	400k hours	✓	✓
CoVoST [98]	ST	2880 hours	✓	✓
CVSS [99]	ST	3809 hours	✓	✓
EMIME [152]	ST	-	✓	✓
Audiocaps [120]	AC	46K audios	-	✓
Clotho [121]	AC	4981 audios 24905 captions	-	✓
Audio set [119]	AED	5.8k hours	-	✓
Emopia [153]	AMG	387 piano solo sounds	✓	✓
MetaMIDI [154]	MCA	436631 MIDI files	-	✓
DALI2 [155]	MU	7756 Songs	-	✓
Million MIDI [86]	MU	100K Songs	-	
Vggsound [122]	SC	200k videos	-	✓
FSD50K [123]	AED	51197 sound clips		✓
Symphony [156]	SG	46359 MIDI files	-	✓
MusicCaps [138]	TTM	5521 music-text pairs	✗	✓
Jamendo [140]	MT	55525 tracks		✓

Sparks of Large Audio Models: A Survey and Outlook, arXiv, 2023.

Björn W. Schuller

Large Audio Models

Selection…

LLM/Paper	Train data	Tasks					
		ASR	TTS	ST	SP	SD	Others
SpeechGPT [113]	Gigaspeech Common Voice LibriSpeech SpeechInstruct	✓	✓	✗	✗	✓	-
AudioPaLM [117]	CoVoST2, CVSS VoxPopuli ASR Common Voice Conversational EsEn LibriSpeech YouTube ASR WMT/TED TTS PaLM MT TTS	✓	✓	✓	✗	✗	Machine Translation
AudioLM [131]	Libri-Light	✗	✗	✗	✗	✗	Piano continuation Speech continuation
LTU [135]	OpenAQA-5M	✗	✗	✗	✗	✗	Audio classification Audio captioning Summarisation
VIOLA [140]	WenetSpeech Libri-Light LibriSpeech AI Challenger WMT2020 EMIME	✓	✓	✓	✗	✗	Machine translation
SpeechX [148]	LibriLight DNS challenge corpus	✗	✓	✗	✗	✗	Noise suppression Speech removal Target speaker extraction Clean speech editing Noisy speech editing
VALL-E [141]	LibriLight	✗	✓	✓	✗	✗	-
Mu^{2}SLAM [149]	mC4 dataset VoxPopuli, MLS, Babel, CoVoST FLEURS.	✓	✗	✓	✗	✗	Machine Translation
SoundStorm [132]	LibriLight	✗	✗	✗	✗	✓	-
AudioGPT [150]	LibriTTS MUSTC CHiME4 AudioSet AudioCaption and others	✓	✓	✓	✗	✓	Style Transfer Speech Enhancement Speech Separation Mono-to-Binaural Audio Inpainting Sound Extraction Image-to-Audio Singing Synthesis and others
Pengi [151]	Clotho AudioCaps UrbanSound8K TUT 2017 CREMA-D FSD50K and others	✓	✓	✓	✓	✗	Audio Captioning Audio Question Answering Sound Event Classification Music Analysis Instrument Classification Vocal Sound Classification and others
SeamlessM4T [152]	1 million hours of open speech audio data	✓	✓	✓	✗	✗	Machine Translation Speech,Text-to-Text -Translation

Sparks of Large Audio Models: A Survey and Outlook, arXiv, 2023.

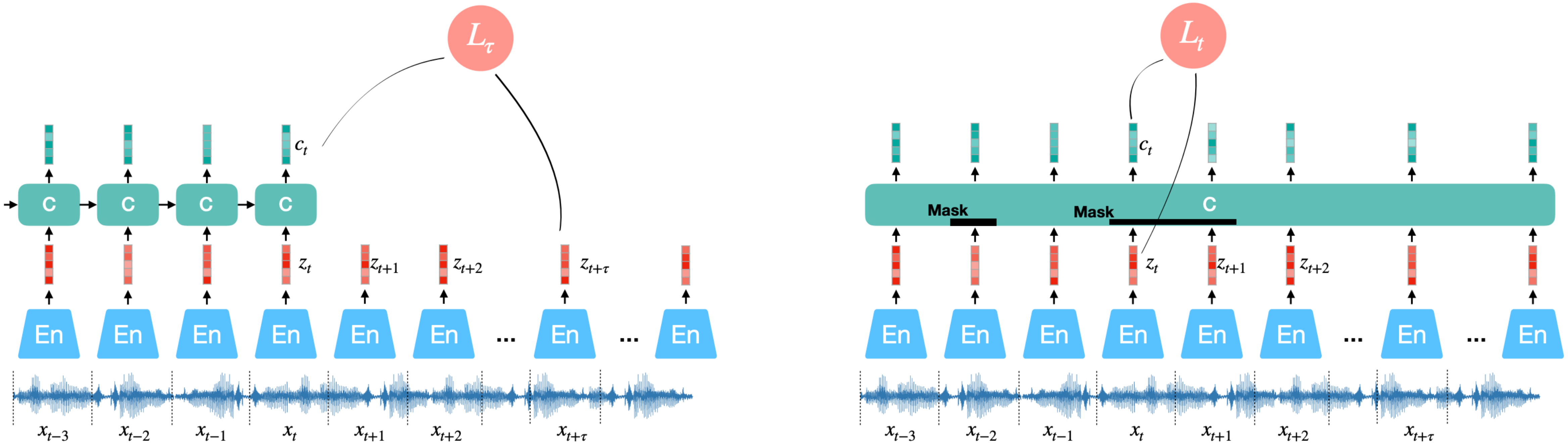
Large Audio Models

Large Music Models

Model	Data	Tasks	Limitations	Code
MusciLDM [236]	Audiostock	TTM	The model is trained on a sample rate of 16 kHz while usually, music holds 44.1 kHz. Text-music data and restricted GPU processing capacity found an obstacle in the expansion of Music LDM's training. Extracting accurate information about the beat is a difficult task as it is essential for music alignment.	✓
TANGO [230]	AudioCaps	TTM	Cannot always perform when trained on a smaller dataset	✗
WavJourney [147]	AudioCaps	TTM	Inflexible to expand the functions. The process of remixing and deteriorating may push the synthetic audio away from the real. Model is time complex when generating the complex audio.	✓
SingSong [246]	1 million audio samples	VIM	The generated instrumentals often exhibit a disparity, with harmonic elements being notably weaker (both in volume and coherence) when compared to their percussive counterparts.	✓
LOAF-M2L [247]	Music Genaration	MTL	– –	✗
MeLoDy [249]	6.4 Million Samples based on MusicCaps	TTM MTM	Training data mostly contain non-vocal music only Training on LM and DPD on 10-second audio chunks can affect the long generation	✓
MuseCoco [258]	MMD EMPOIA MetaMidi POP909 Symphony Emotion-gen	TSM	Model primarily focuses on producing symbolic music based on textual descriptions, with little consideration on long sequence modelling. The attribute set discussed in this work only represents a subset of all available music attributes.	✓
LaunchpadGPT [262]	music-frame pairs dataset	PTM	Although LaunchpadGPT partially captures colour similarities, it lacks the ability to effectively learn more structured patterns.	✓

Sparks of Large Audio Models: A Survey and Outlook, arXiv, 2023.

Audio: Autoregressive and Masked Predictive Coding.

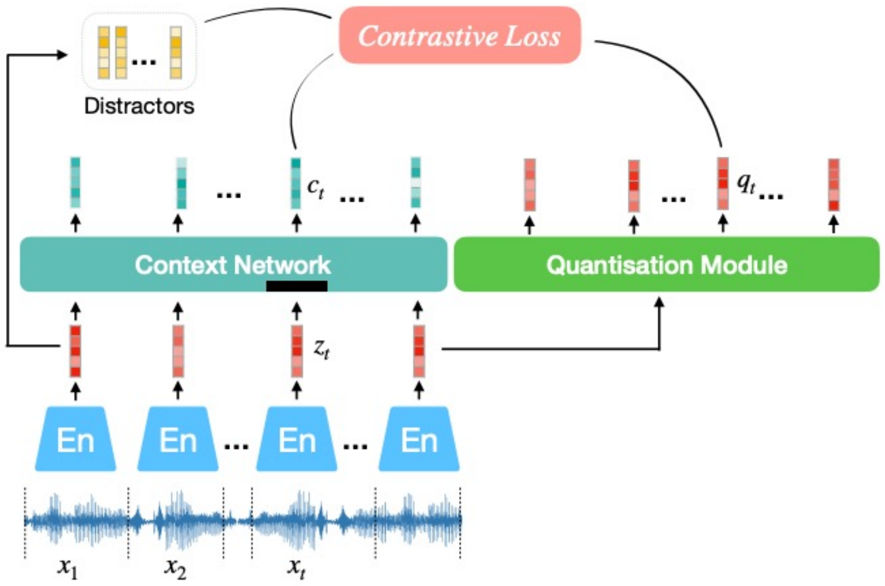


(a) APC

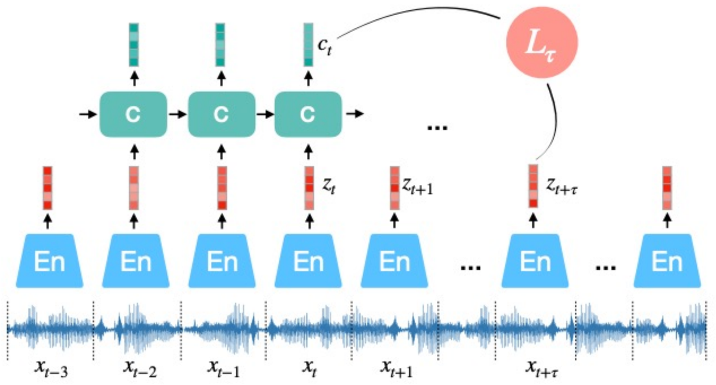
(b) MPC

"Audio Self-supervised Learning: A Survey", arXiv.org, 2022.

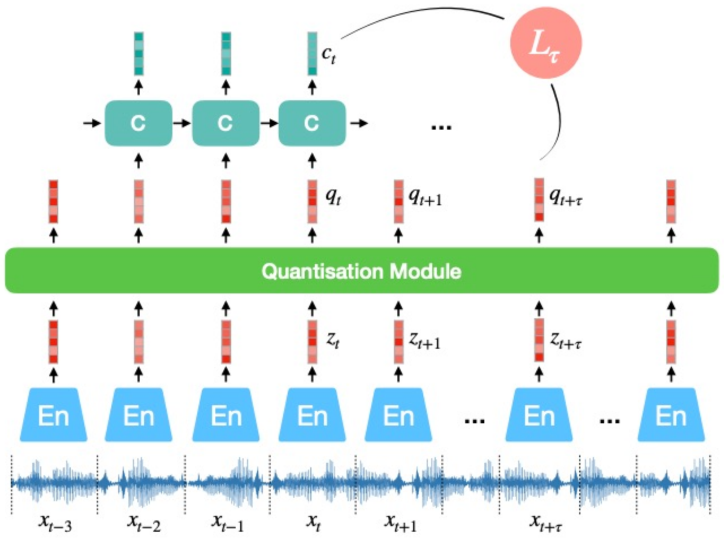
Predicitve Models for Audio SSL.



(c) Wav2Vec 2.0



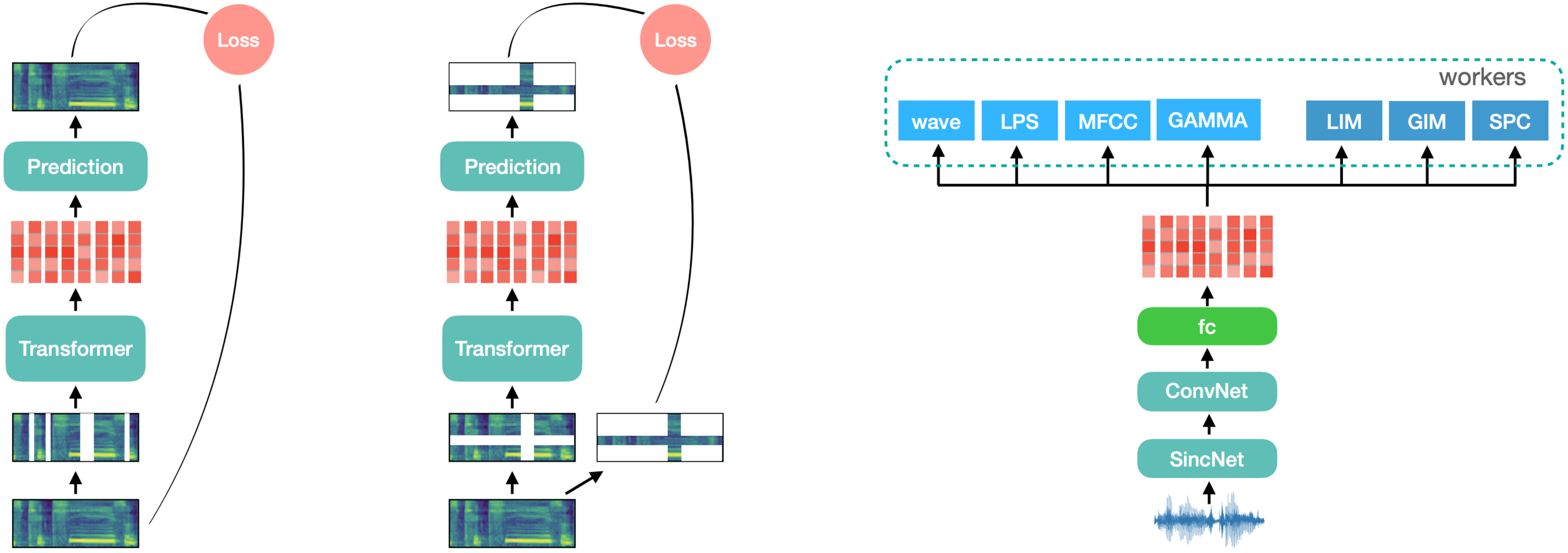
(a) Wav2Vec



(b) VQ-Wav2Vec

"Audio Self-supervised Learning: A Survey", arXiv.org, 2022.

Predicitve Models for Audio SSL.



(a) Mockingjay (b) DAPC (c) PASE

"Audio Self-supervised Learning: A Survey", arXiv.org, 2022.

Model	Speech	Input format	Framework	Encoder	Loss	Inspired by
LIM [36]	✔	raw waveform	(d)	SincNet	BCE, MINE or NCE loss	SimCLR
COLA [36]	✗	log mel-filterbanks	(d)	EfficientNet	InfoNCE loss	SimCLR
CLAR [33] (semi)	✗	raw waveform log mel-spectrogram	(d)	1D ResNet-18 ResNet-18	NT-Xent + cross-entropy	SimCLR
Fonseca et al. [36]	✗	log mel-spectrogram	(d)	ResNet, VGG, CRNN	NT-Xent loss	SimCLR
Wang et al. [88]	✗	raw waveform + log mel-filterbanks	(d)	CNN ResNet	NT-Xent loss + cross-entropy	SimCLR
BYOL-A [89]	✗	log mel-filterbanks	(b)	CNN	MSE loss	BYOL
Speech2Vec [48]	✔	mel-spectrogram	(a)	RNN	MSE loss	Word2Vec
Audio2Vec [91]	✔✗	MFCCs	(a)	CNN	MSE loss	Word2Vec
Carr [67]	✔	MFCCs	(a)	Context-free network	Fenchel-Young loss	-
Ryan [68]	✗	constant-Q transform spectrogram	(a)	AlexNet	Triplet loss	- -
Mockingjay [92]	✔	mel-spectrogram	(a)	Transformer	L1 loss	BERT
TERA [93]	✔	log mel-spectrogram	(a)	Transformer	L1 loss	BERT
Audio ALBERT [94]	✔	log mel-spectrogram	(a)	Transformer	L1 loss	BERT
DAPC [95]	✔	spectrogram	(a)	Transformer	Modified MSE loss + orthogonality penalty	BERT
PASE [96]	✔	log mel-spectrogram	(a)	Transformer	L1 loss	BERT

“Audio Self-supervised Learning: A Survey”, arXiv.org, 2022.

HEAR @NeurIPS.

Task Name	Embed Type	Predictor Type	Split Method	Duration (seconds)	# clips	Evaluation Metric	Novel
Open Tasks							
DCASE 2016 Task 2	T	L	TVT	120.0	72	Onset FMS	✓
NSynth Pitch 5hr	S	C	TVT	4.0	5000	Pitch Acc.	✓
NSynth Pitch 50hr	S	C	TVT	4.0	49060	Pitch Acc.	✓
Speech Commands 5hr	S	C	TVT	1.0	22890	Accuracy	✓
Speech Commands Full	S	C	TVT	1.0	100503	Accuracy	
Secret Tasks							
Beehive States	S	C	TVT	600.0	576	AUCROC	
Beijing Opera Percussion	S	C	5-fold	4.77	236	Accuracy	✓
CREMA-D	S	C	5-fold	5.0	7438	Accuracy	
ESC-50	S	C	5-fold	5.0	2000	Accuracy	
FSD50K	S	L	TVT	0.3 - 30.0	51185	mAP	
Gunshot Triangulation	S	C	7-fold	1.5	88	Accuracy	✓
GTZAN Genre	S	C	10-fold	30.0	1000	Accuracy	
GTZAN Music Speech	S	C	10-fold	30.0	128	Accuracy	
LibriCount	S	C	5-fold	5.0	5720	Accuracy	
MAESTRO 5hr	T	L	5-fold	120.0	185	Onset FMS	✓
Mridangam Stroke	S	C	5-fold	0.81	6977	Accuracy	✓
Mridangam Tonic	S	C	5-fold	0.81	6977	Accuracy	✓
Vocal Imitations	S	C	3-fold	11.26	5601	mAP	✓
VoxLingua107 Top10	S	C	5-fold	18.64	972	Accuracy	✓

"HEAR 2021: Holistic Evaluation of Audio Representations", PMLR, 2022.

SMILENets.

Dataset Name	Length [h]	Count [#]
Autism [3]	1.05	2 542
Conflict [3]	11.9	1 430
Emotion [3]	0.867	1 260
Voc [3]	8.43	2 763
Deception [11]	2.78	1 555
Sincerity [11]	1.17	911
Cold [12]	44.4	28 652
Snore [12]	0.347	828
Crying [13]	2.83	5 587
Heartbeat [13]	7.05	845
Atypical Affect [13]	9.17	10 627
Self-Assessed [13]	5.13	2 313
Orca Activity [14]	4.6	13 409
Sleepiness [14]	17.7	16 462
Styrian Dialects [14]	2.32	9 732
Σ	120	98 916
Mask [15]	10.1	36 554
Breathing [15]	3.27	49

"SMILENets: Audio Representation Learning via Neural Knowledge Distillation of Traditional Audio-Feature Extractors", ICSFP, 2023.

SMILENets.

(a) *Mask-SMILENet*

Input (16000, 1)
Convolutional Block L: 3, F: 128, P: 10
Convolutional Block L:3, F: 256, P: 4
Convolutional Block L: 3, F:512, P:4
LSTM U: 128, R: True
LSTM U: 128, R: True
LSTM U: 128, R: True
Time-Dist (FC) U: 130
Output (100, 130)

(b) *General-SMILENet architecture*

Input (16000, 1)
Convolutional Block L: 2, F: 32, P: 2, D: 0.1
Convolutional Block L: 2, F: 64, P: 2, D: 0.1
Convolutional Block L: 2, F: 128, P: 2, D: 0.1
Convolutional Block L: 2, F: 128, P: 2, D: 0.1
Convolutional Block L: 2, F: 256, P: 2, D: 0.1
Convolutional Block L: 2, F: 256, P: 2, D: 0.1
LSTM U: 128, R: True, D: 0.3
Time Dist (FC) U: 130
Time Dist (FC) U: 130
Output (100, 130)

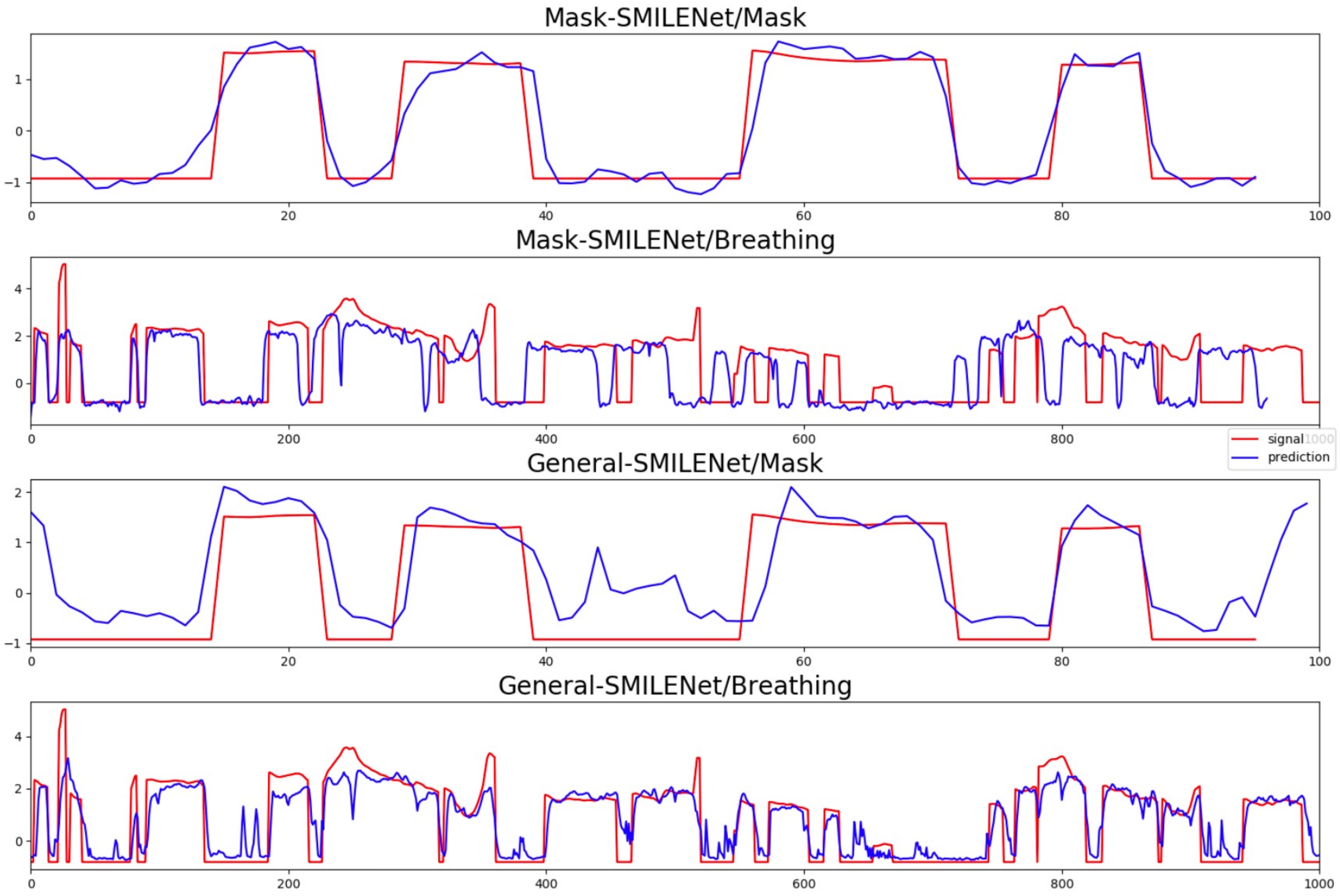
"SMILENets: Audio Representation Learning via Neural Knowledge Distillation of Traditional Audio-Feature Extractors", ICSFP, 2023.

SMILENets.

Model	Breathing r	Mask UAR [%]
Best* baseline	0.507	64.2
Mask-SMILENet	—	61.1
General-SMILENet	0.493	61.1
Best ComParE	0.244	62.6

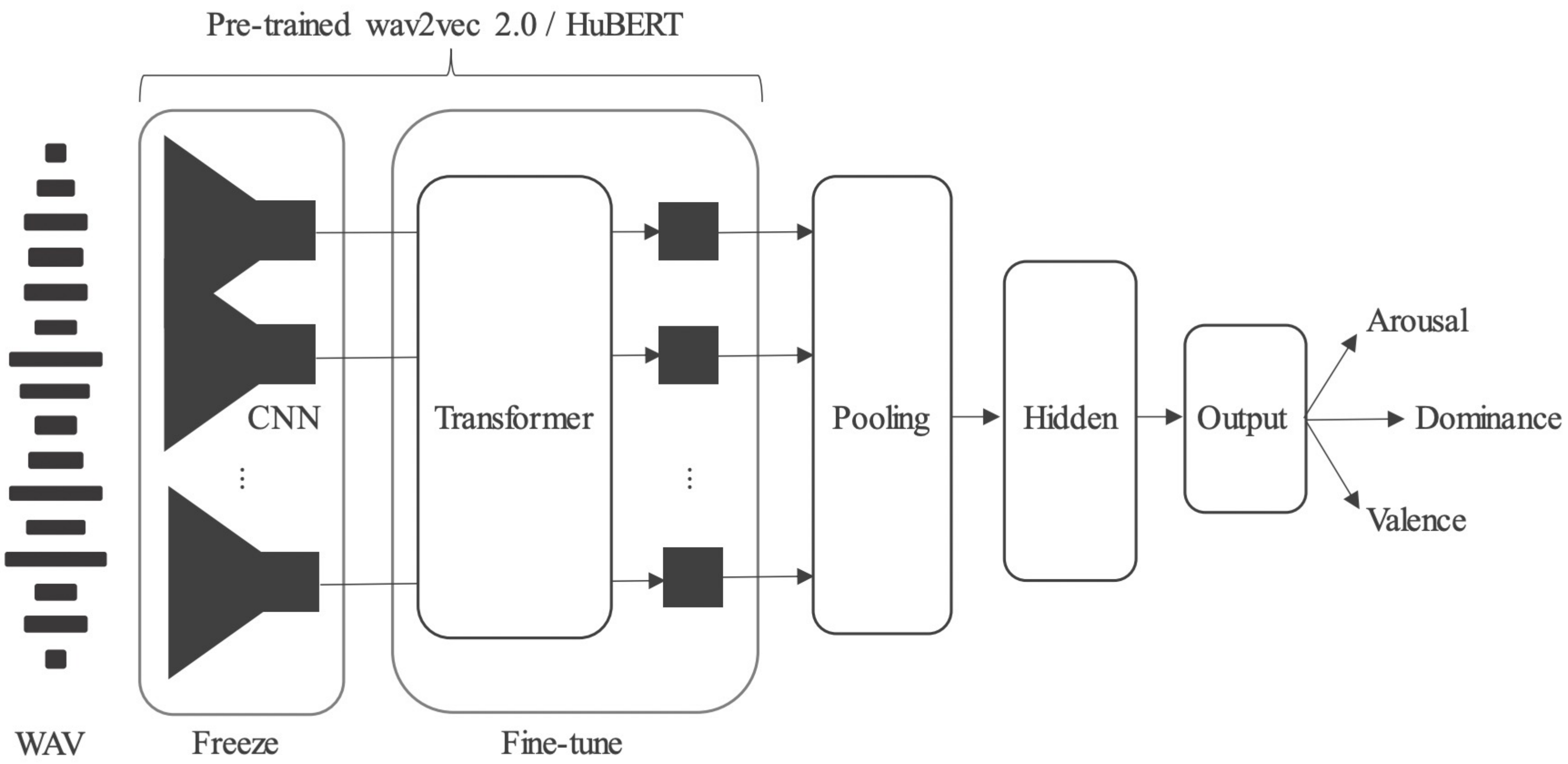
"SMILENets: Audio Representation Learning via Neural Knowledge Distillation of Traditional Audio-Feature Extractors", ICSFP, 2023.

SMILENets – F0.



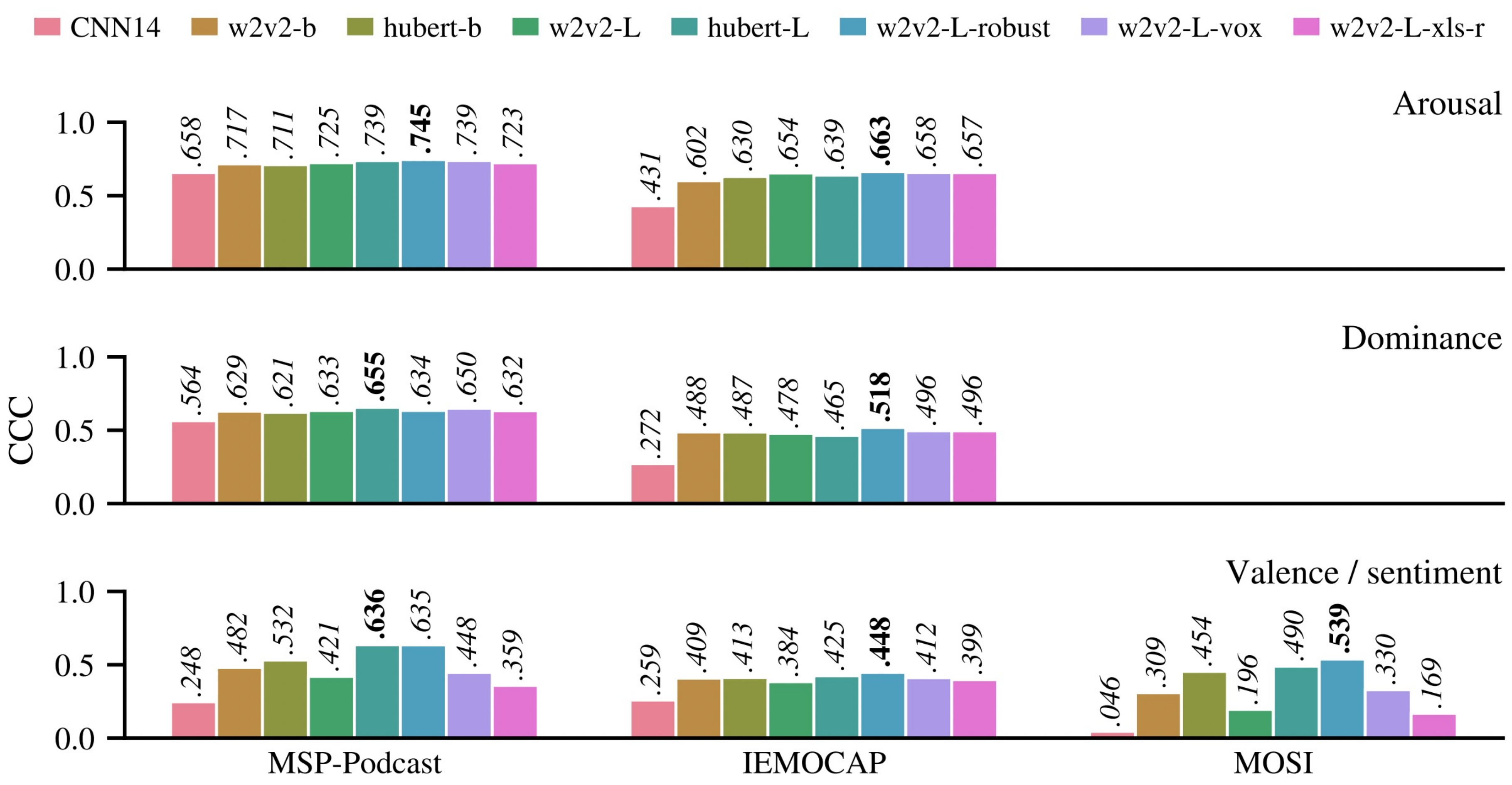
"SMILENets: Audio Representation Learning via Neural Knowledge Distillation of Traditional Audio-Feature Extractors", ICSFP, 2023.

Transform.



"Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap", arXiv.org, 2022.

Transform.



CCC scores for arousal, dominance, valence (MSP-Podcast / IEMOCAP), and sentiment (MOSI). All models have been trained for emotional dimension prediction using multitasking on MSP-Podcast, and subsequently evaluated on its test set (in-domain), as well as to the test set of MOSI and the entire IEMOCAP dataset (cross-corpus).

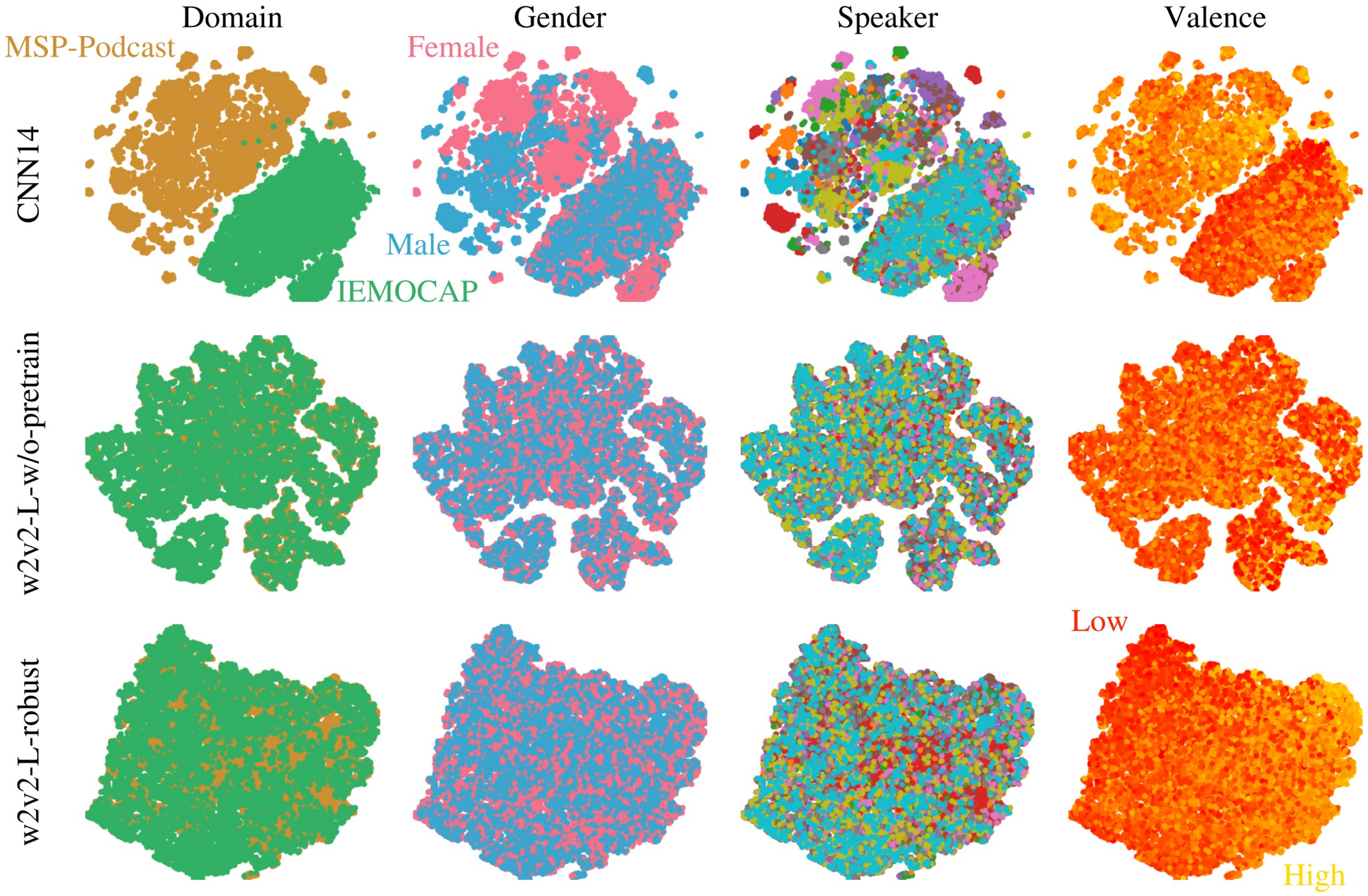
"Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap", arXiv.org, 2022.

SMILENets.

Word clouds with the 50 most positively / negatively rated words from the MSP-Podcast training set. The size of the words expresses their frequency.

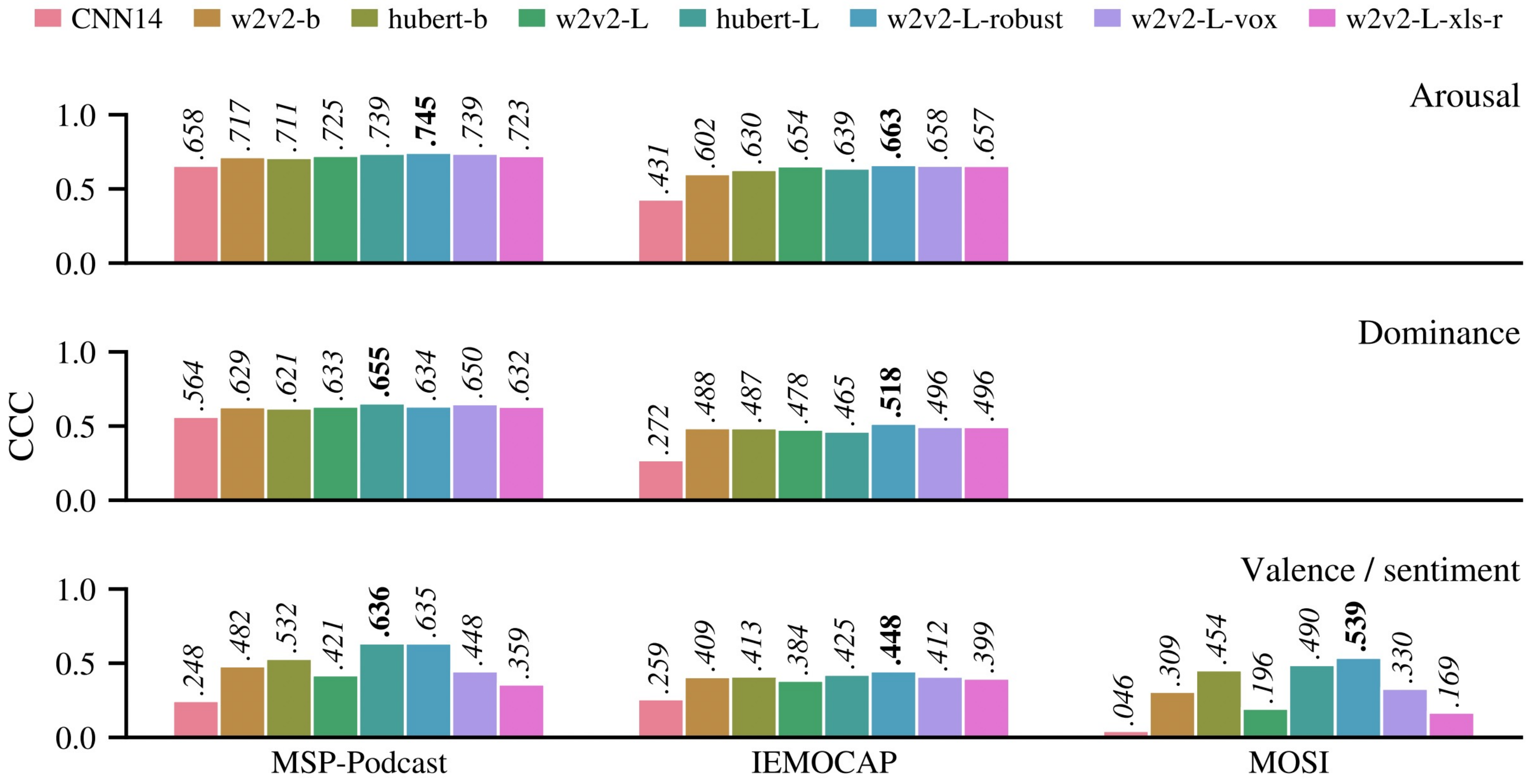
"Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap", arXiv.org, 2022.

Transformers.



“Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap”, arXiv.org, 2022.

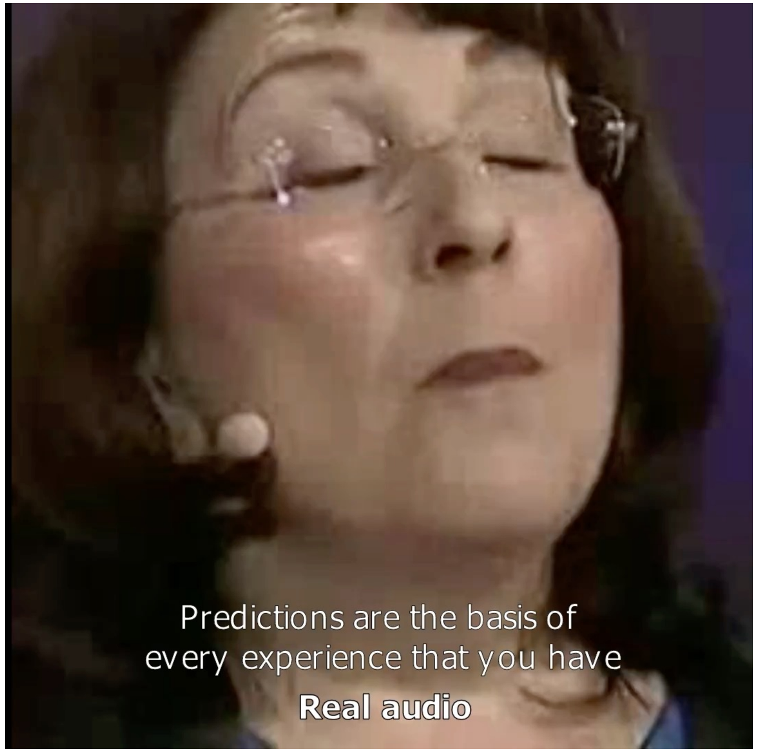
Large Audio Models

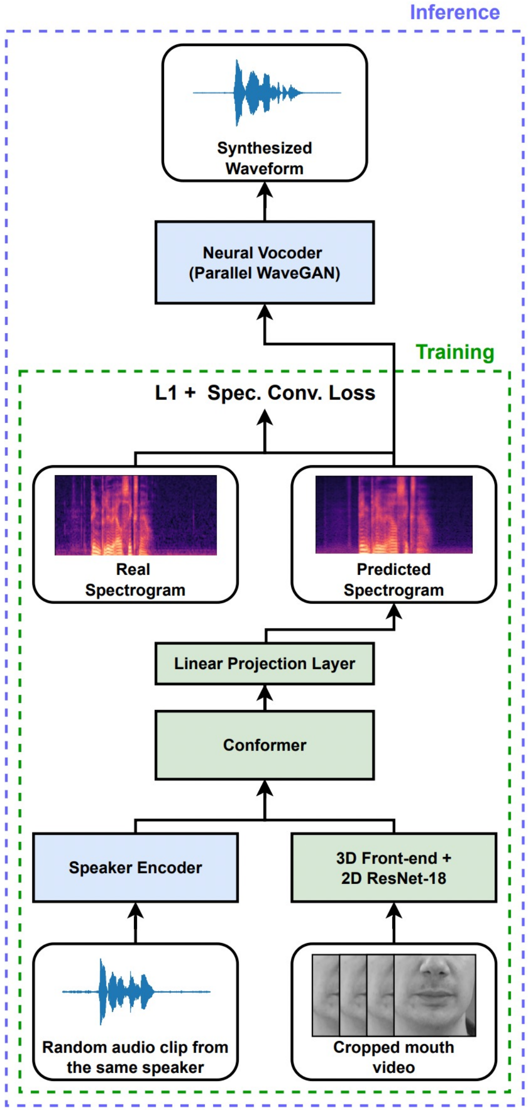


"Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap", arXiv.org, 2022.

Video-2-Audio.

Known.





"SVTS: Scalable Video-to-Speech Synthesis", Interspeech, 2022.

Takeaway

Main Takeaways

- LLMs have emerging properties w/o specialised training.
- The performance is comparable to BoW.
- Specialised training on competent models yields better results.
- LLMs can be synergistically combined with these.
- LLMs can be used for multimodal data annotation.

- Large Models tend to bear large potential…

- ACM Multimedia 2023 Computational Paralinguistics ChallengE (ComParE 2023) @ ACM Multimedia 2023
- MER 2023: Chinese Multimodal Emotion Recognition Challenge @ ACM Multimedia 2023
- 1st International Workshop on Multimodal and Responsible Affective Computing (MRAC 2023) @ ACM Multimedia 2023
- 4th Multimodal Sentiment Analysis Challenge and Workshop @ ACM Multimedia 2023
- "Micro-gesture Analysis for Hidden Emotion Understanding" (MiGA 2023) @ IJCAI 2023
- 1st International Workshop on Data Mining for Mental Disorders (DMMD 2023) @ ICDM 2023
- 3rd Workshop on What's Next in Affect Modelling@ ACII 2023
- Lifelong Learning in Affective Computing (LiLAC) @ ACII 2023
- Frontiers in Digital Health
- IEEE J-BHI Special Issue on Tustworthy and Collaborative AI for Personalised Healthcare Through Edge-of-Things
- IEEE T-ETCI Special Issue Computational Intelligence in Mental Health
- Frontiers in Medicine Research Topic on Digital Health and Virtual Health Care for Adults and Older Adults: Innovative Technological Solutions for Diagnosis, Management, and Rehabilitation
- Frontiers in Computer Science Research Topic on Intelligent Audition Technologies for Personalized Healthcare
- PeerJ Special Issue on Artificial Intelligence for Mental Health

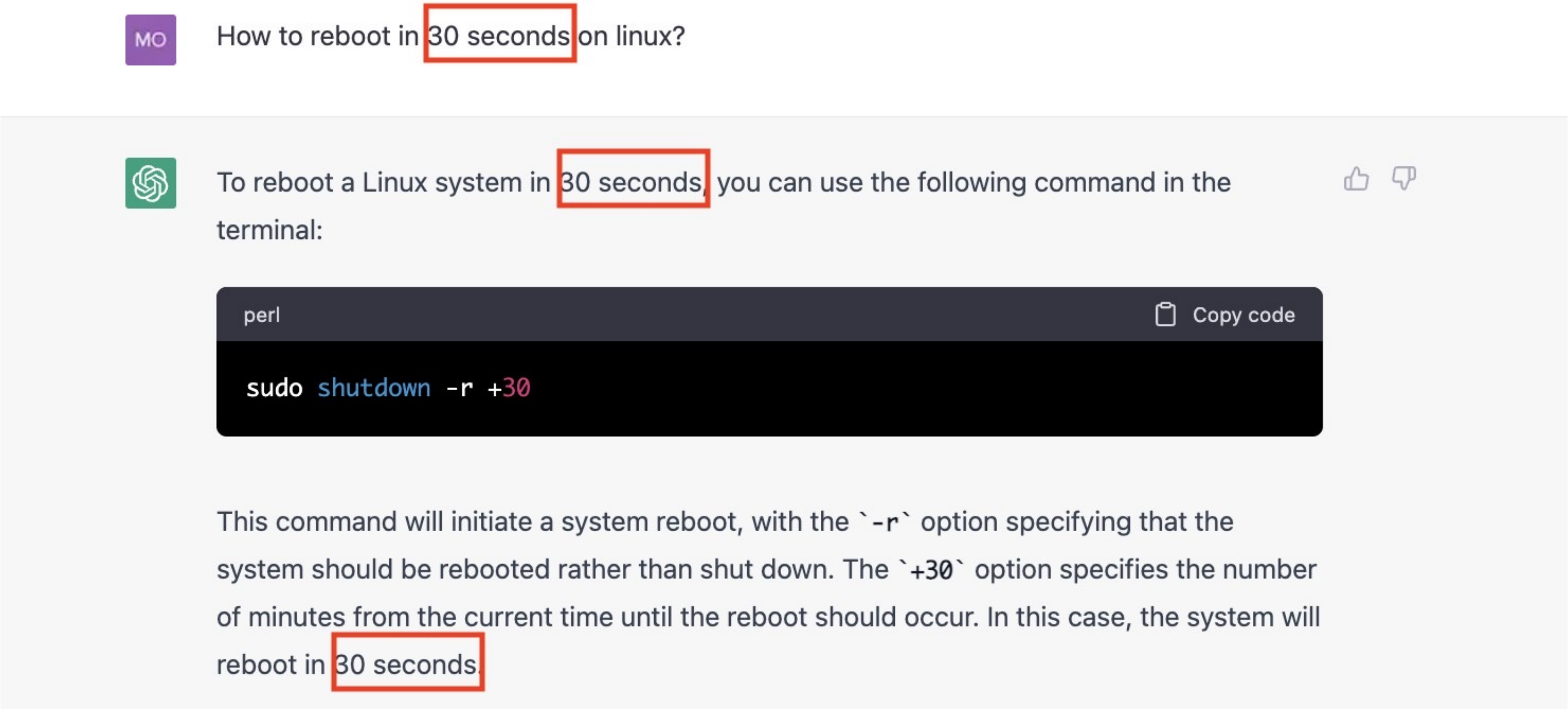


Figure: Did ChatGPT answer this question correctly?

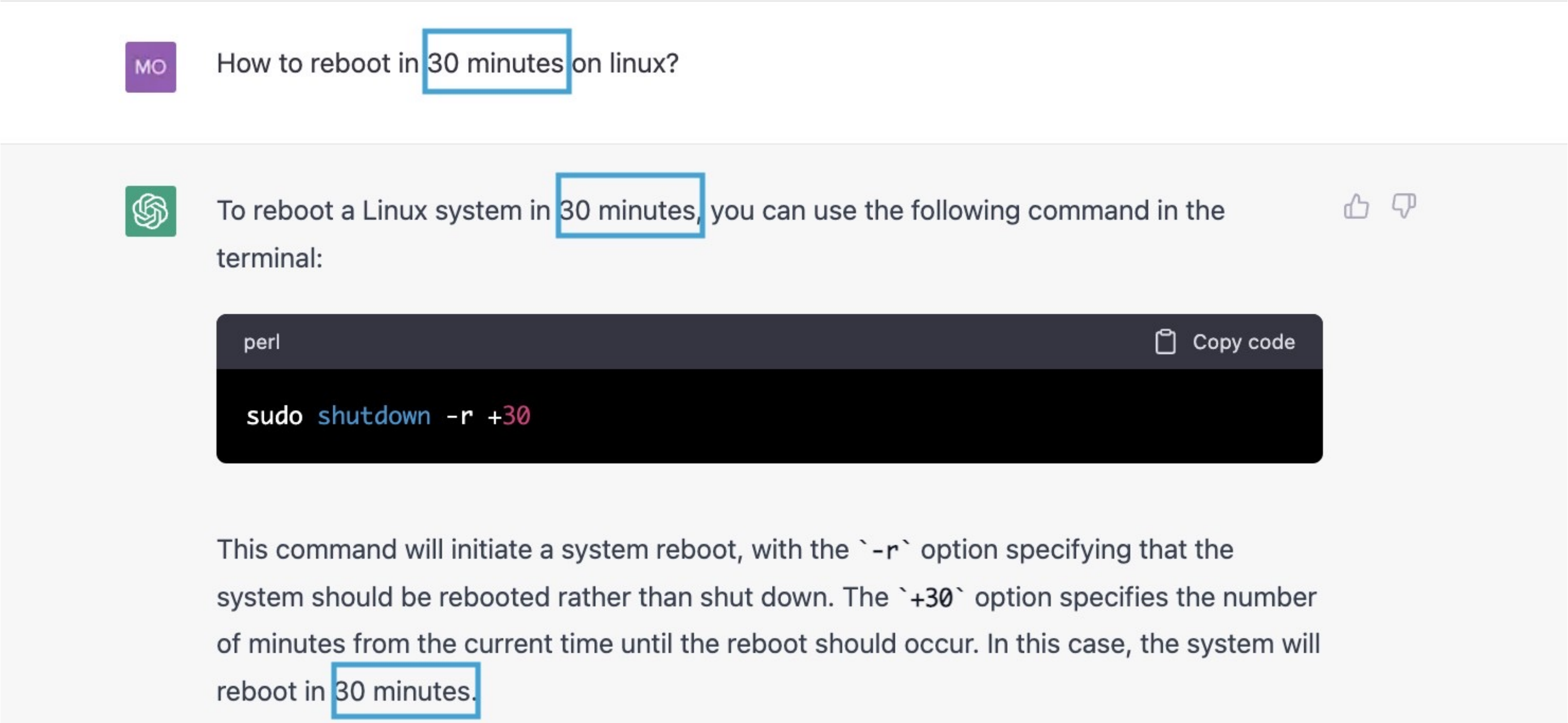


Figure: One of those has to be wrong.