Language Models

Analogy

Guess the next word: I enjoy drinking coffee, because I like…

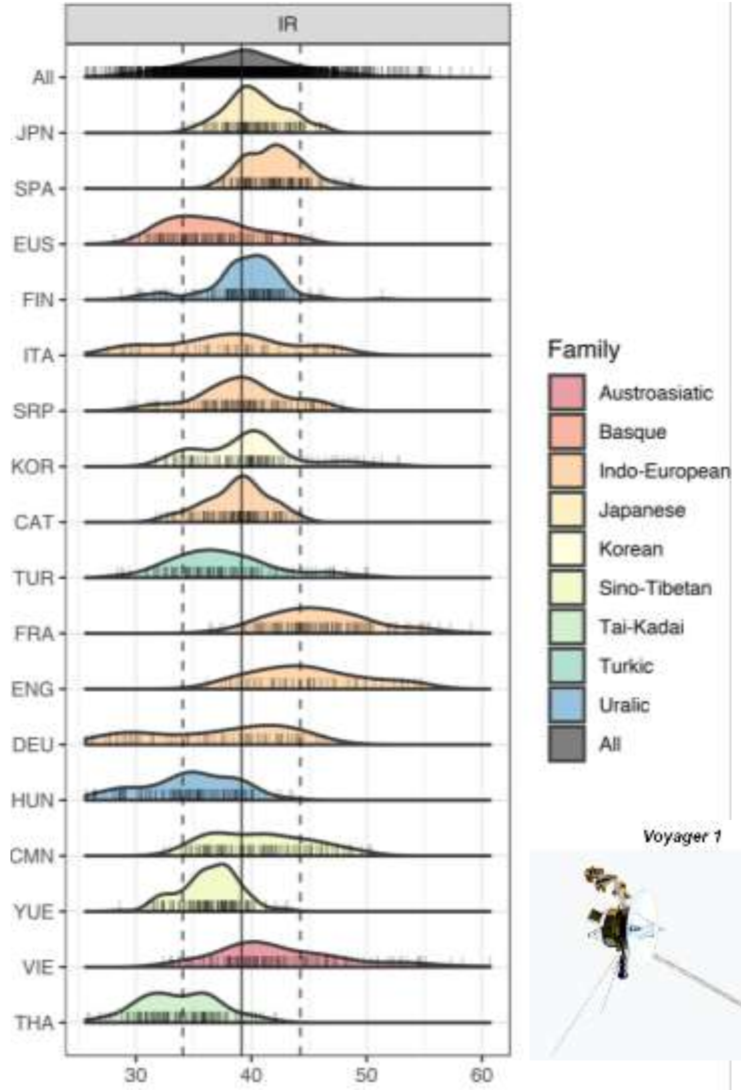
- Language or code has low information density
- Plenty of equivalent formulations / phrasings
- Simplistic model: Markov chain n-grams

1-gram:
P(next word | ”like”)

2-gram:
P(next word | ”I like”)

3-gram:
P(next word | “because I like”)

… no coffee!



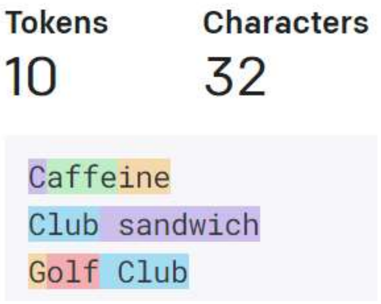
Bits/s

C Coupé, et al., *Sci. Adv.* 2019.

Embedding

Tokenisation

- Split text into a finite set of substrings
- May already include some context



Tokens 10

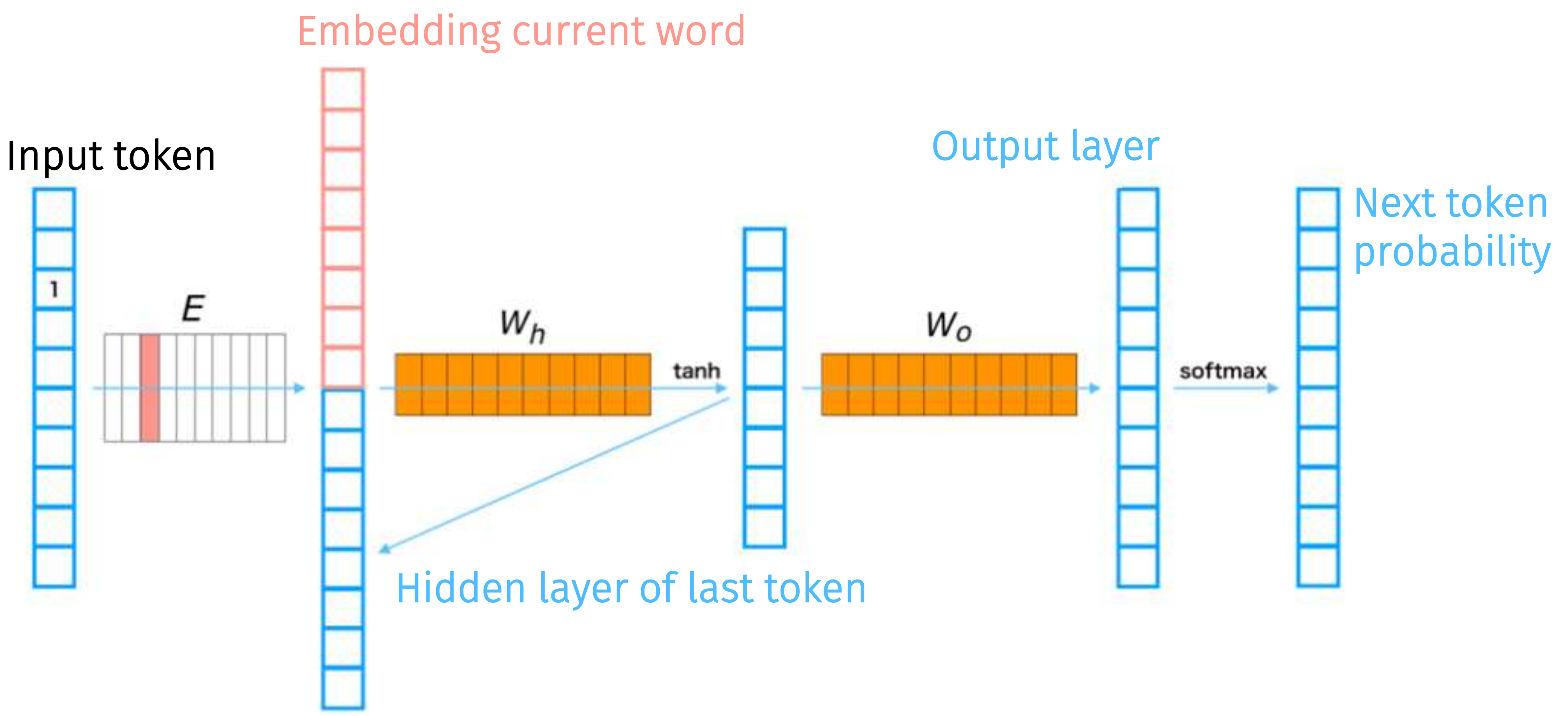
Characters 32

```
Caffeine
Club sandwich
Golf Club
```

Embedding

- Starts from one vector for each token
- Typically fed through neural network to include context

Recurrent Neural Network



Embedding current word

Input token

Output layer

Next token probability

Hidden layer of last token

Transformer

Self-attention

- Output y is weighted sum of all embeddings x so far
- Weights calculated, not trained
- Parameters introduced to tailor to three roles
  - Query, Key, Value

$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{x}_j$$

$$w'_{ij} = \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j \qquad w_{ij} = \frac{\exp w'_{ij}}{\sum_j \exp w'_{ij}}$$

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i \qquad \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i \qquad \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i$$

$$w'_{ij} = \mathbf{q}_i^{\mathsf{T}} \mathbf{k}_j$$

$$w_{ij} = \text{softmax}(w'_{ij})$$

$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{v}_j \,.$$

Architecture: free to choose, as long as it includes self-attention

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

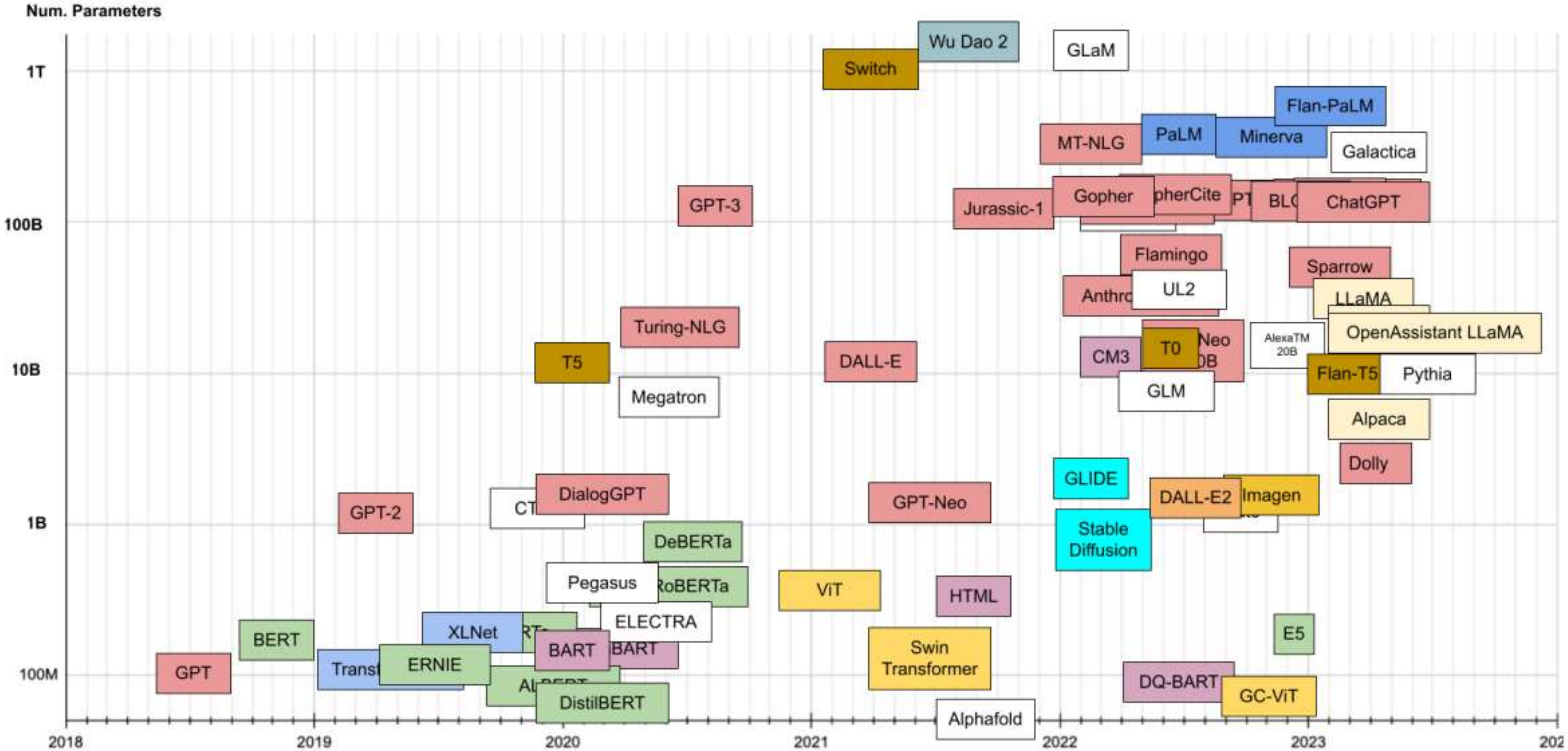
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

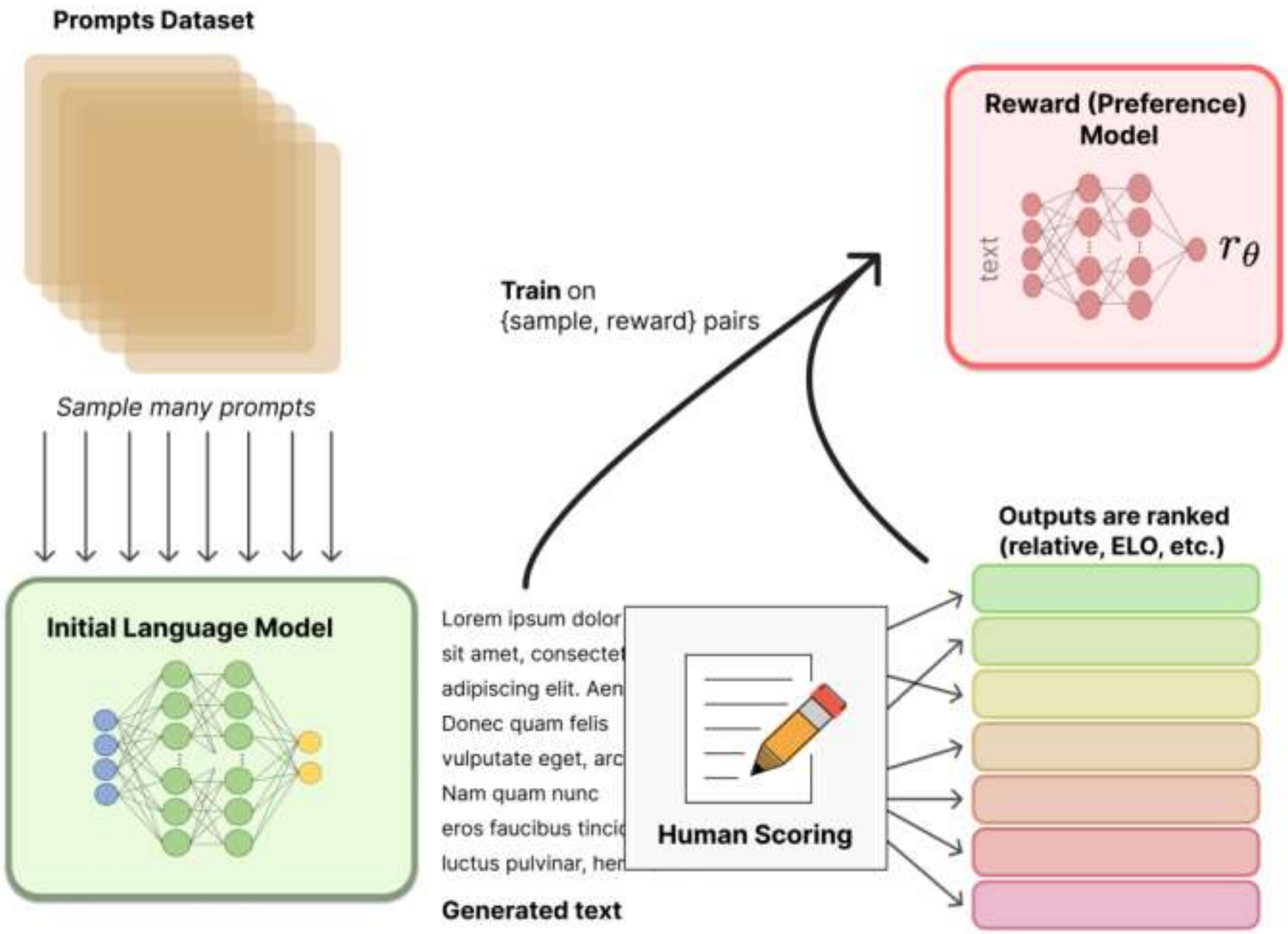
A Vaswani, et al., *NIPS*. 2017.

Transformer



X. Amatriain, et al., *arXiv*, 2302.07730, **2023**.

Reinforcement Learning


Prompts Dataset
Sample many prompts
Initial Language Model
Train on {sample, reward} pairs
Reward (Preference) Model
text
r_θ
Lorem ipsum dolor sit amet, consectet adipiscing elit. Aen Donec quam felis vulputate eget, arc Nam quam nunc eros faucibus tincid luctus pulvinar, her
Generated text
Human Scoring
Outputs are ranked (relative, ELO, etc.)

https://huggingface.co/blog/rlhf

Limitations

Reasoning

- Statistical model, not a factual one
- Sounds trustworthy, since human-like

Training data

- (Data) privacy
- Biases

Transparency and Governance

- Intransparent model
- Intransparent access

Reliability

- Code generation: edge cases, API changes
- Statistical output

Summary Large Language Models

- Trying to build long-range context in sequential data
- Expensive in training
- Context-aware predictions
- Helps in data mining, code generation, interpretation, summaries