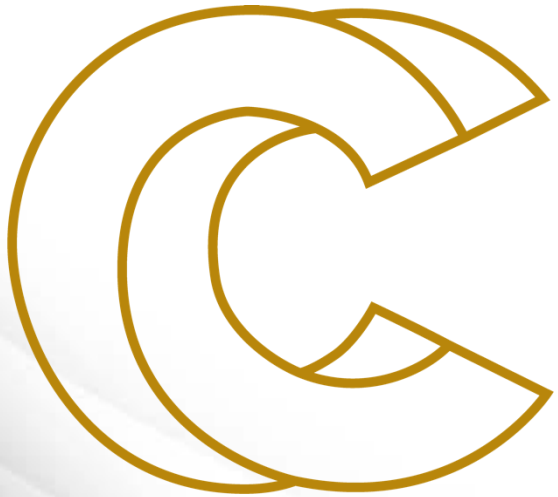


EURO^2

Presentation at University of Sofia, 10.10.2024

There is Still Place for Humans

Dr. Hristo Iliev, NCC-BG

EuroCC2 Workshop, University of Sofia, Bulgaria

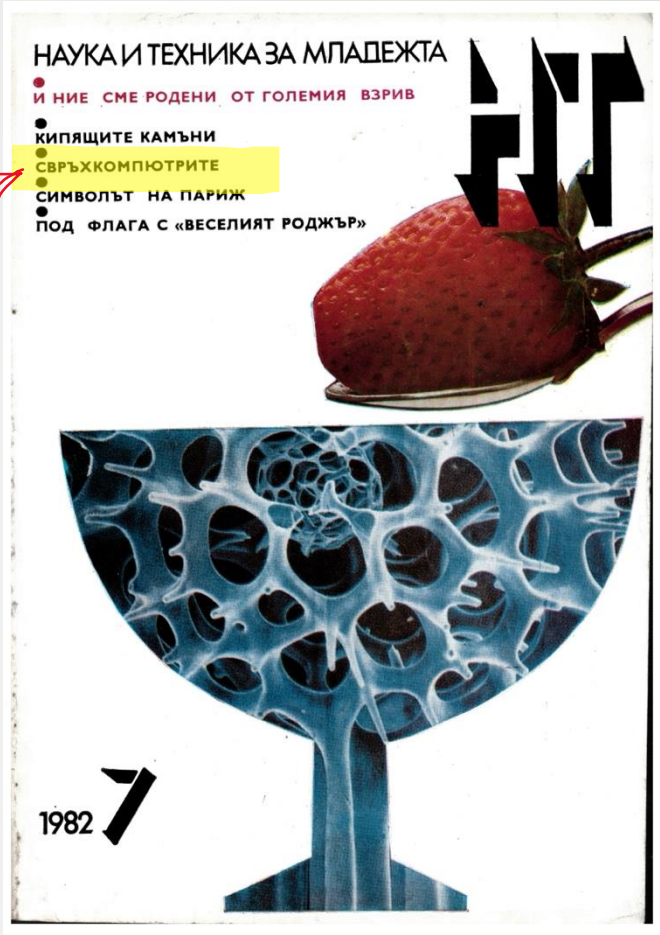
There is Still Place for Humans

Dr. Hristo Iliev, NCC-BG

EURO²

About me

- First encounter with a computer ca. 1986
- Fixation on HPC – ca. 1991
- First [legal] access to HPC – 2003 at EPCC
- Ph.D. in Physics (Uni Sofia) – 2010
- HPC Team @ RWTH (Aachen, DE) – 2012-2018
- Chief Data Scientist @ holler.live – 2018-2020
- Chief Data Scientist @ NOTO – 2020-



Language Models

What is a language model?

- Probabilistic model of a natural language
- Models prediction and/or correction of text
- Not necessarily a cognitive model

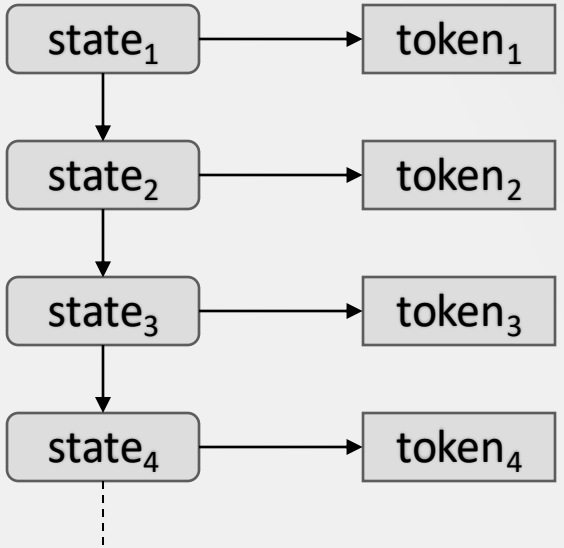
Language Models

Large Language Models

- Giant statistical models that learn to output human-like text
- Attributed almost magical properties
- “Glimpses of AGI” (according to Microsoft)

Language Models

Sequence models



- Map the current state to an output token
- Modify somehow the state and repeat
- Generation ends when a special end-of-sequence token is produced

Language Models

Remember IRC?

Language Models

Markov chain bots

- Popular in the 90's and early 00's
- Start with a corpus of IRC channel messages or books/articles
- Build an N-gram frequency table
  - Usually bigrams due to memory limitations
- Turn the table into $P(w_i|w_j)$ and use a Markov process to generate text
- Results were entertaining and sometimes uncannily human-like

Language Models

Markov chains

"In publishing and graphic design, Lorem ipsum is a placeholder text commonly used to demonstrate the visual form of a document or a typeface without relying on meaningful content."

	publishing				
in:1	1				

Language Models

Markov chains

“In publishing and graphic design, Lorem ipsum is a placeholder text commonly used to demonstrate the visual form of a document or a typeface without relying on meaningful content.”

	publishing	**and**			
in:1	1	0			
publishing:1	0	1			

Language Models

Markov chains

“In publishing and graphic design, Lorem ipsum is a placeholder text commonly used to demonstrate the visual form of a document or a typeface without relying on meaningful content.”

	publishing	**and**	**placeholder**		
in:1	1	0	0		
publishing:1	0	1	0		
a:1	0	0	1		

Language Models

Markov chains

“In publishing and graphic design, Lorem ipsum is a placeholder text commonly used to demonstrate the visual form of a document or a typeface without relying on meaningful content.”

	publishing	and	placeholder	document	
a:2	0	0	0,5	0,5	
in:1	1	0	0	0	
publishing:1	0	1	0	0	

Language Models

Markov chains

EURO 2

“In publishing and graphic design, Lorem ipsum is a placeholder text commonly used to demonstrate the visual form of a document or a typeface without relying on meaningful content.”

	publishing	and	placeholder	document	typeface
a:3	0	0	0,33	0,33	0,33
in:1	1	0	0	0	0
publishing:1	0	1	0	0	0

Language Models

Markov chains

Lorem ipsum is a

	publishing	and	placeholder	document	typeface
a:3	0	0	0,33	0,33	0,33

Lorem ipsum is a placeholder

Lorem ipsum is a document

Lorem ipsum is a typeface

Language Models

Markov chains

- Bigrams hardly capture any context
- Generated text is incoherent gibberish
- More context can be captured with higher N-grams
  - Index size limitation (database of bigrams from Wikipedia is 50 GB)
  - Still very “rigid” – context in human languages is quite fluid

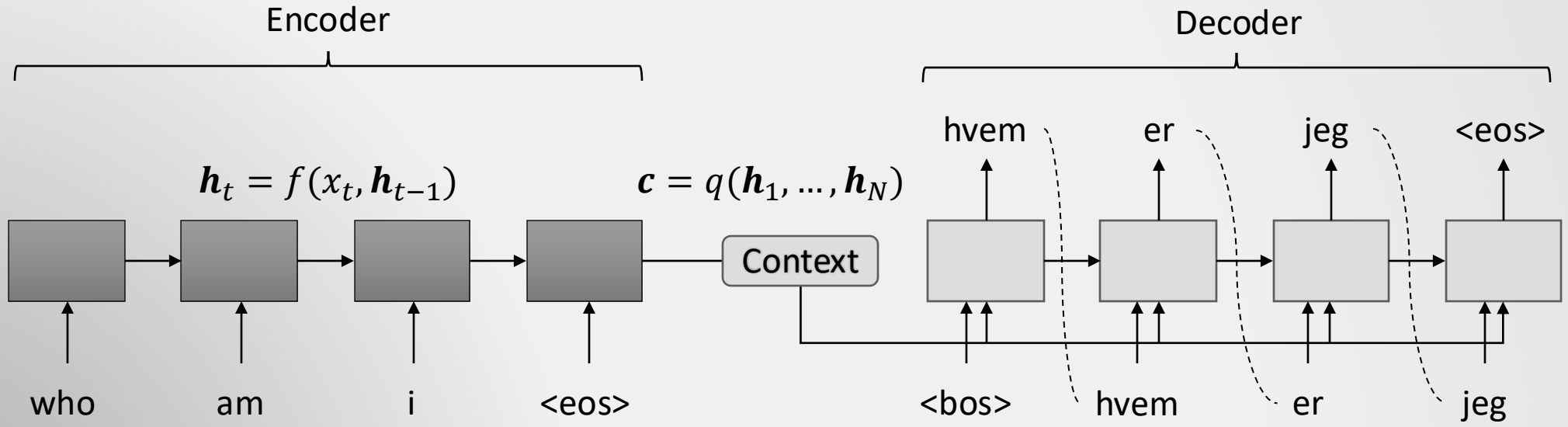
Language Models

Encoder-Decoder Sequence to Sequence (seq2seq)

- Two stages of Recurrent Neural Networks (RNNs)
- Encoder – takes in an input sequence and produces a latent context
- Decoder – takes in the context and produces an output sequence

Language Models

Encoder-Decoder Sequence to Sequence (seq2seq)



$$\boldsymbol{s}_{t'} = g(y_{t'-1}, \boldsymbol{c}, \boldsymbol{s}_{t'-1}) \xrightarrow{\text{softmax}} P(y_{t'}|y_1, \ldots, y_{t'-1}, \boldsymbol{c})$$

Language Models

Encoder-Decoder Sequence to Sequence (seq2seq)

- Trained on pairs of sequences: input and expected output
- Fitness metric (e.g. BLEU) measures output precision and alignment
- Limited input and output sequence lengths
  - Information capacity of the context vector
  - The decoder has access to the context only and not to the input sequence

Language Models

Attention mechanism

- Given: database of *m* key-value pairs D = {($\mathbf{k}_i$, $\mathbf{v}_i$)} and input query **q**
- Attention over D is the linear combination

$$Attention(\boldsymbol{q}, D) = \sum_{i=1}^{m} \alpha(\boldsymbol{q}, \boldsymbol{k}_i)\boldsymbol{v}_i$$

- $\alpha(\mathbf{q}, \mathbf{k})$ – attention weight derived from the similarity of **q** to **k**
- Often referred to as attention pooling

Language Models

Bahdanau architecture

- Replaces the static context with attention pooling over encoder states

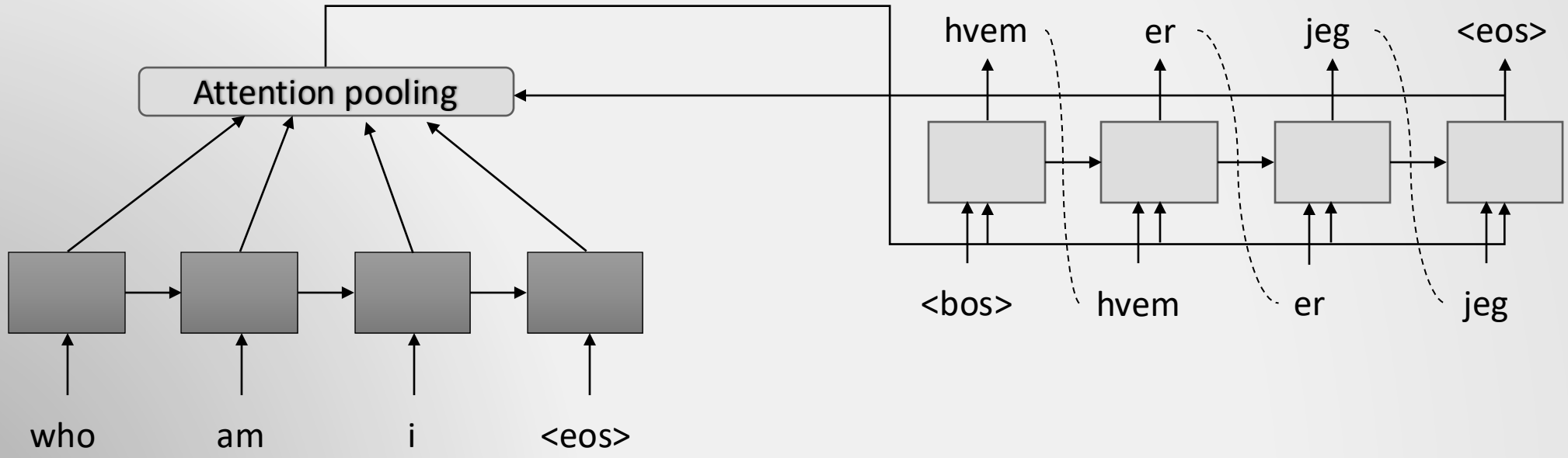
$$\boldsymbol{c}_{t'} = \sum_{i=1}^{T} \alpha(\boldsymbol{s}_{t'-1}, \boldsymbol{h}_i)\boldsymbol{h}_i$$

- The query is the hidden state of the previous decoder stage
- $\alpha(.)$ are derived from an alignment model which is jointly trained with the rest of the network

D. Bahdanau et al. “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR, 2015

Language Models

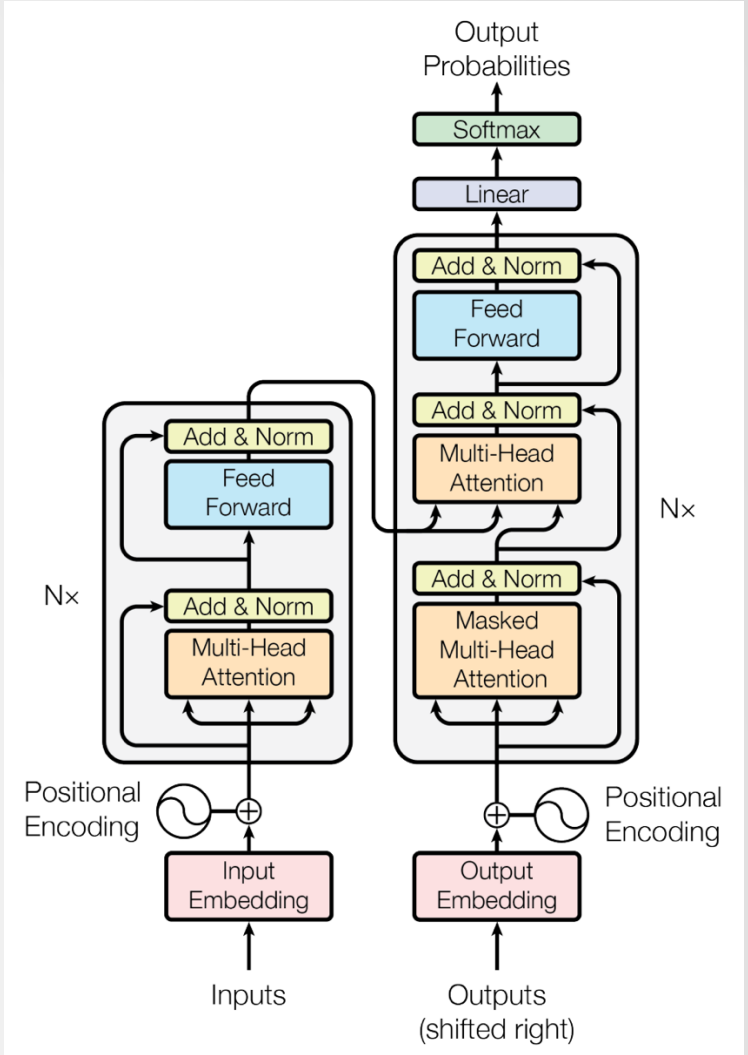
Bahdanau architecture



Language Models

Transformer models

- Advanced seq2seq model
- Attention and self-attention for context tracking
- Developed initially for automatic language translation



A. Vaswani et al. “Attention Is All You Need”, NIPS, 2017

Language Models

EURO^2

Self-Attention

- Attention over a sequence with each token as query

$$SelfAttention(\boldsymbol{x}_i, \{(\boldsymbol{x}_j, \boldsymbol{x}_j)\}) = \sum_{j=1}^{m} \alpha(\boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{x}_j$$

- Learning a similarity function allows for learning tokens in context
- Scaled dot-product as similarity

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Language Models

Multiheaded (self-)attention

- Project **q**, **k**, **v** with projection matrices W^Q, W^K, W^V
- Attention(QW^Q, KW^K, VW^V) is called an attention head
- Different sets of projection matrices define several attention heads
- Each head learns to see (transformed) tokens in a different context
- All heads are concatenated together – the model works with all contexts simultaneously

Language Models

Generative models

- Trained to predict the next token(s) in a sequence

  ”Lorem ipsum is a ...” -> “placeholder”

  - Single corpus, no pairs
- The output is a *max likelihood* continuation of the input

Large Language Models

GPT

- **G**enerative **P**re-trained **T**ransformer
- The most well-known LLM by OpenAI
- From GPT-1 (117m params, 2018) to GPT-3.5 (175b params, 2022)
- GPT-4 (2023) – details are not public, but speculation is it has ~1,8t params and/or is a combination of several smaller models (mixture of experts)

Large Language Models

ChatGPT

- Conversational interface to GPT
- Model output is a max likelihood sequence of tokens in response to a user prompt embedded in a system prompt
- Very likely syntactically correct – syntax rules are easy to derive
- Factually – not so much
- Hallucinations

Large Language Models

“ChatGPT is a blurry JPEG of the web”

Ted Chiang, The New Yorker
https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

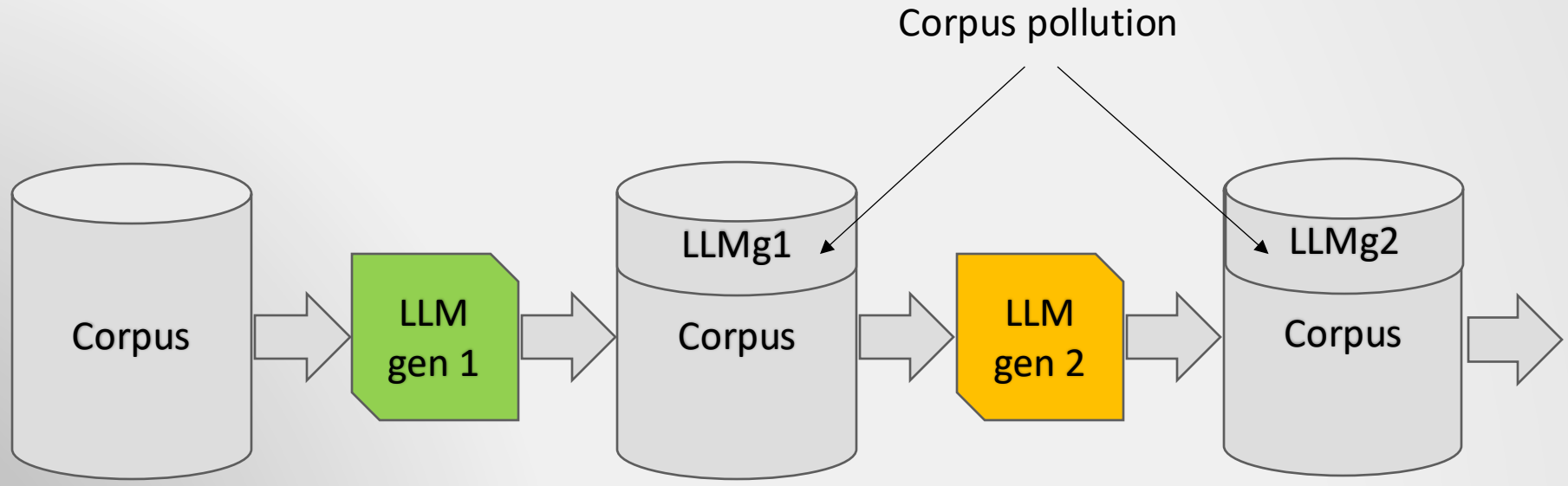
Model Collapse

Recursive training

- LLMs are progressively used to compose vast amounts of text that end up on the Internet
  - Legitimate services, content farms, fake news sites, etc.
  - No clear indication that this is not human generated
- Next generations of LLMs will inevitably include output from previous generations in their training data

Model Collapse

Recursive training



Model Collapse

Definition

- "Model collapse is a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality."

I. Shumailov et al, "AI models collapse when trained on recursively generated data" Nature **631**, pp. 755-759 (2024)

Model Collapse

Driving factors

- **Statistical approximation error**
  - Information loss due to finite sample size
- Functional expressivity error
  - Imperfect approximation due to finite neural network size
- Functional approximation error
  - Limitations of the learning procedure

Model Collapse

Statistical approximation error

- Major driver of the model collapse
- Finite sample size means tokens from the tail of the distribution may never end up in the output
- Inevitable loss of rare tokens
- Shrinking distribution support
- Disappears in the limit of infinitely large samples

Model Collapse

Functional expressivity error

- Secondary driver of the model collapse
- Finite-size neural networks have limited expressivity
  - Some parts of the distribution may not be represented and disappear
  - … while likelihood may get introduced outside of the original support
- Higher expressivity may lead to overfitting the noise in training data
- Introduces errors even in the limit of infinitely large samples

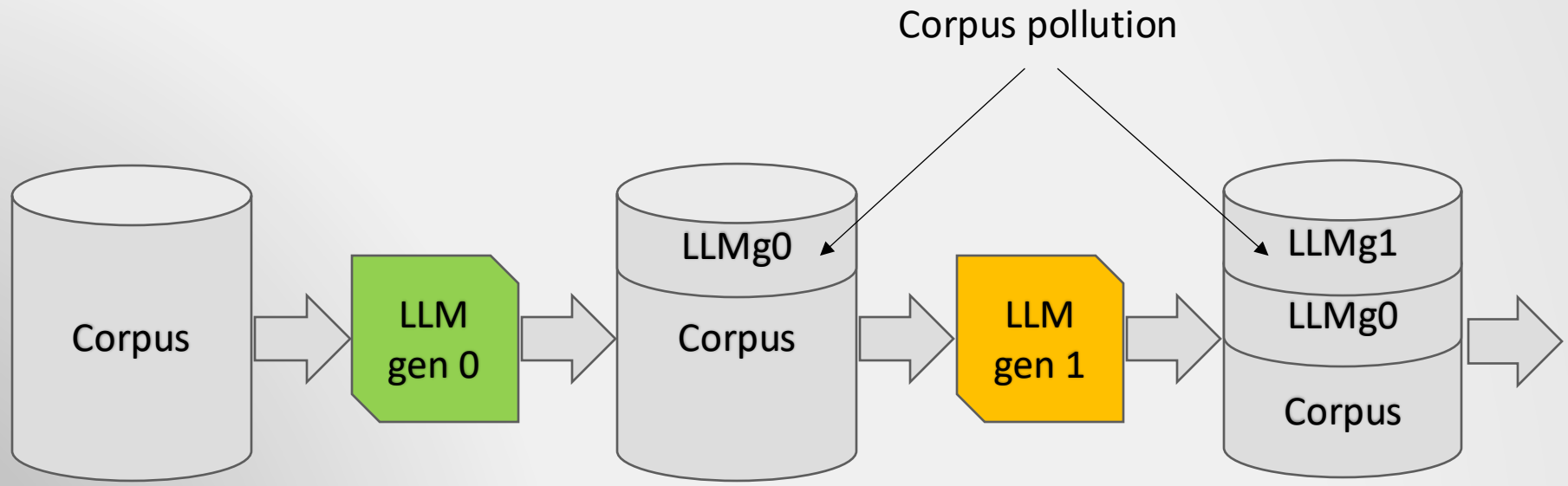
Model Collapse

Functional approximation error

- Secondary driver of the model collapse
- Limitations of the learning procedure
  - Wrong choice of objective function
  - Structural bias of SGD
- Introduces errors even in the limit of perfect expressivity and infinitely large samples

Model Collapse

Recursive training



p_0

$p_1 = \alpha_1 \mathcal{F}_0(p_0) + \gamma_1 p_0$

$p_2 = \alpha_2 \mathcal{F}_1(p_1) + \beta_2 p_1 + \gamma_2 p_0$

Model Collapse

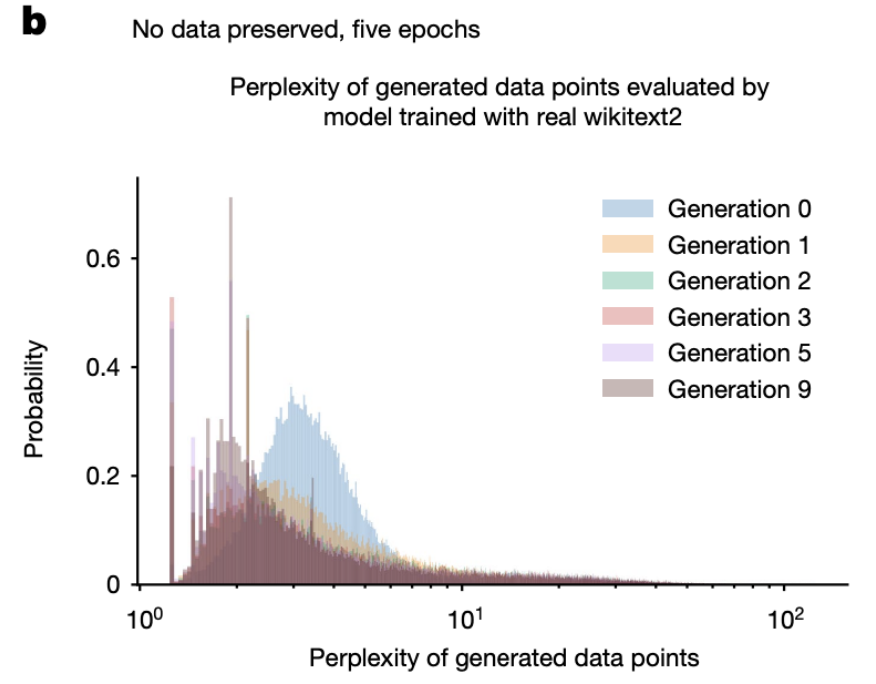
Distribution collapse

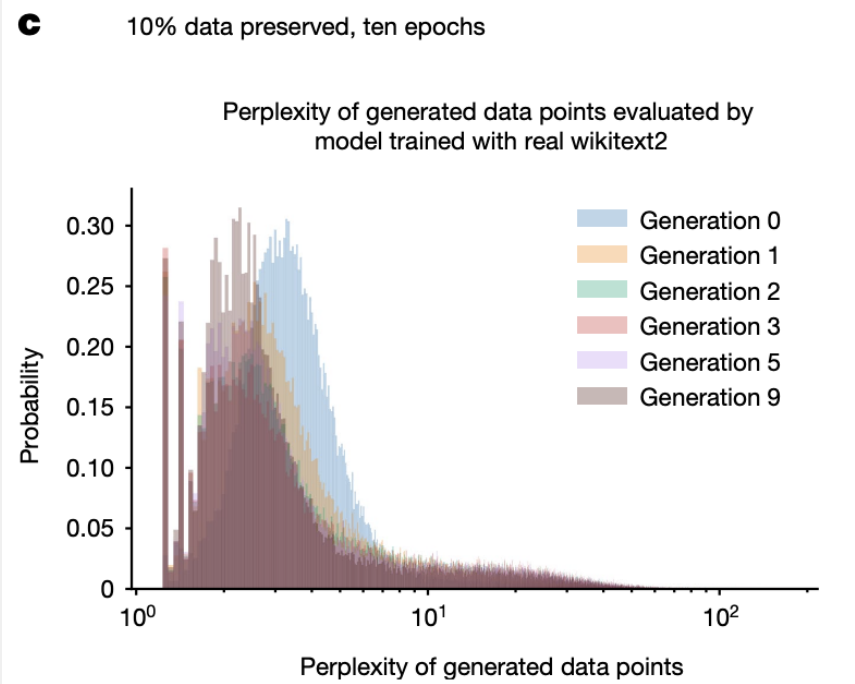
- Various toy models result in distribution collapse
- Discrete distribution with exact approximation sampled as a Markov chain converges to an absorbing state
- Multidimensional Gaussian model collapses to a delta function

Model Collapse

EURO²

Collapse during fine tuning OPT-125m with wikitext2





I. Shumailov et al, "AI models collapse when trained on recursively generated data", Nature **631**, pp. 755-759 (2024)

Model Collapse

Why is it important?

- Low-probability outputs are important for:
  - Fairness
  - Cultural and knowledge preservation
  - Marginalised groups
  - Fighting stereotypes
  - Understanding complex systems

Model Collapse

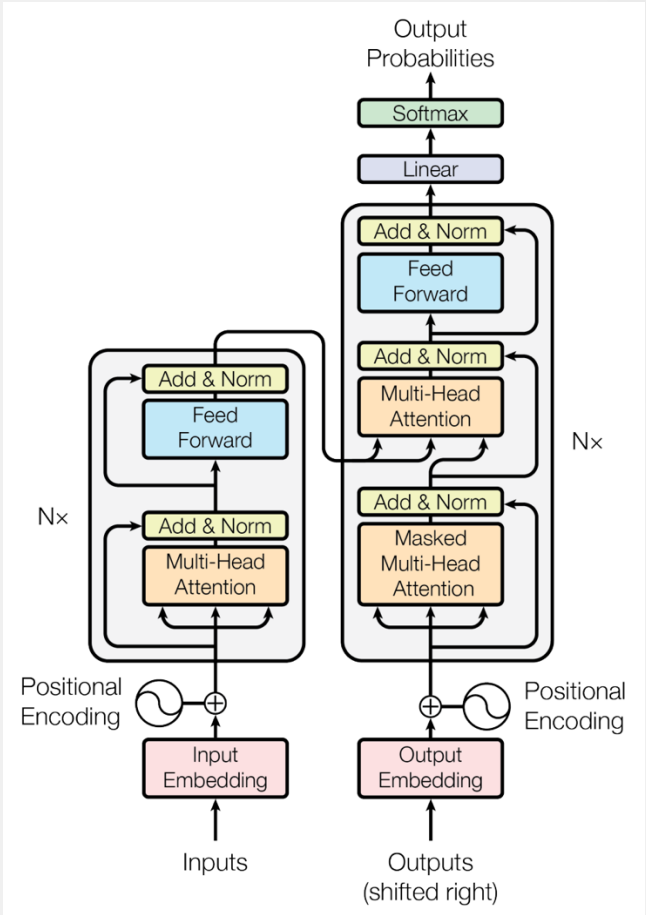
First mover advantage

- Scraping the Internet tomorrow will include much more generated text
  - Distribution shift, then model collapse
- Preservation of the original data over time
- Data provenance – what is LLM-generated and what is genuine
- Community-wide coordination

Systems Thinking

Always look at the big picture

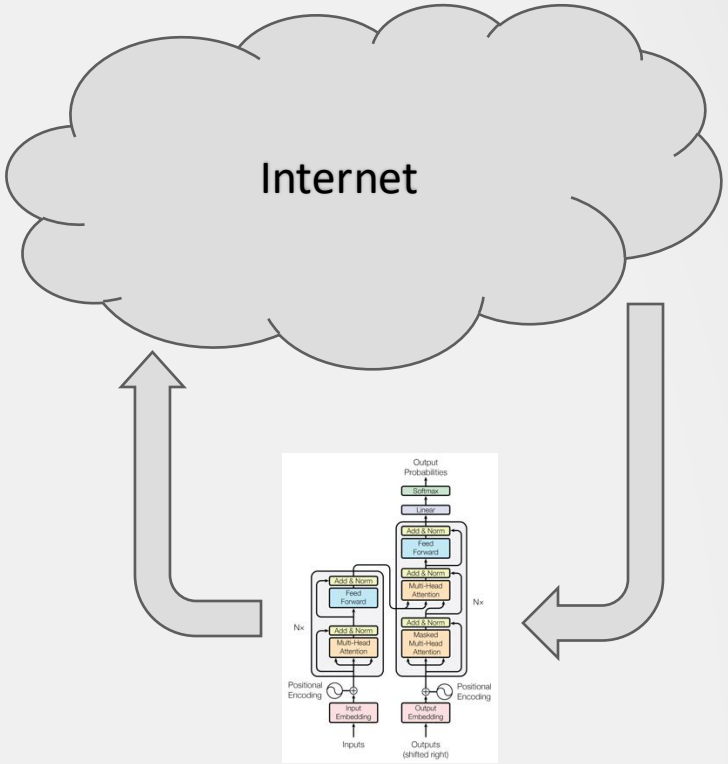
- It’s too easy to focus on this



Systems Thinking

Always look at the big picture

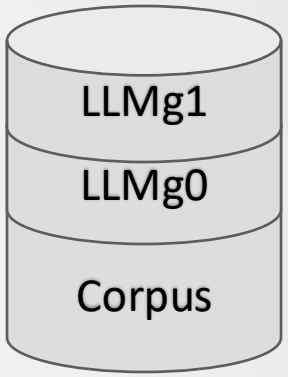
- … and miss that



Conclusion

There is still place for humans

- Human creativity fuels all AI models
- Human curiosity and cultural systems preserve the long tails of the distribution
- Up to us to not let blurry JPEGs pollute the Internet



Thanks!

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 951732. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, United Kingdom, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Switzerland, Turkey, Republic of North Macedonia, Iceland, Montenegro