NETS
IUSS Laboratory
for Neurolinguistics,
Computational
Linguistics &
Theoretical Syntax

IUSS
Scuola Universitaria Superiore Pavia

CRISTIANO CHESI

NETS @ IUSS PAVIA

What do "we need", in the end?

Intelligenza Linguistica, Sovranità AI
e Futuro Responsabile

Villanova.ai - Cagliari, 23 Settembre 2025

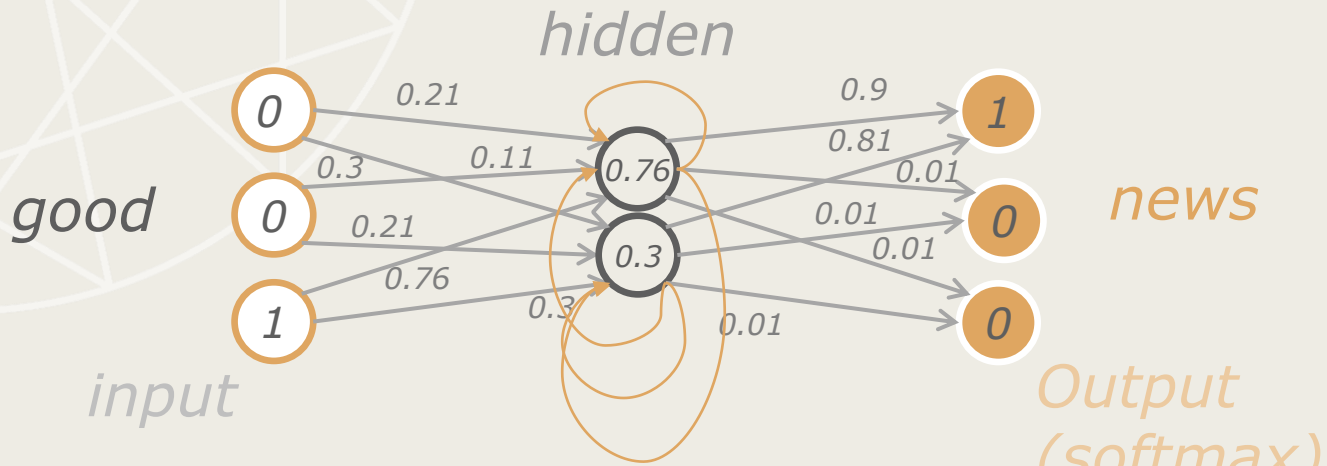
Plato's Problem

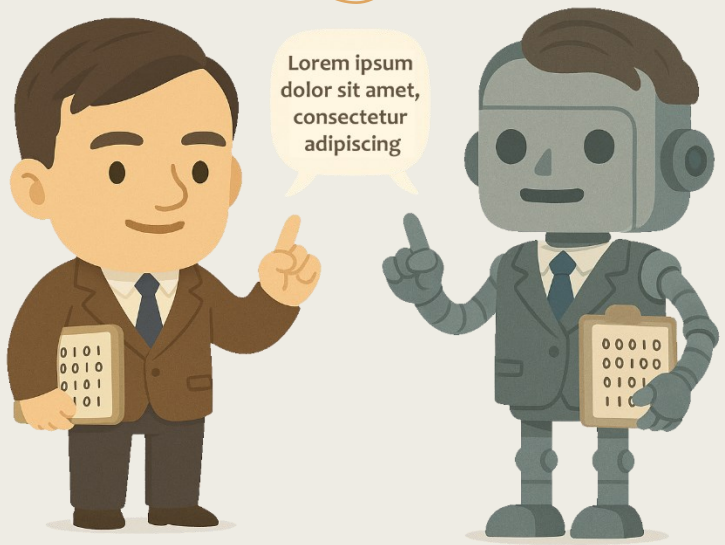
How is it possible that human beings, whose **contacts with the world** are **brief, personal,** and **limited**, are still able to know so much?

- A linguistic example:
  - **Which car** do you think the mechanic fixed _?
  - *__Which car__ do you think the mechanic fixed **the engine** ?
- A **logical conclusion**:
  We come with an **innate predisposition** to assign **a precise structural analysis** to the (linguistic) **input** we receive (and **not infinite other possible analyses**)

- … of **next word** (Elman 1990)

This is a good… **news**



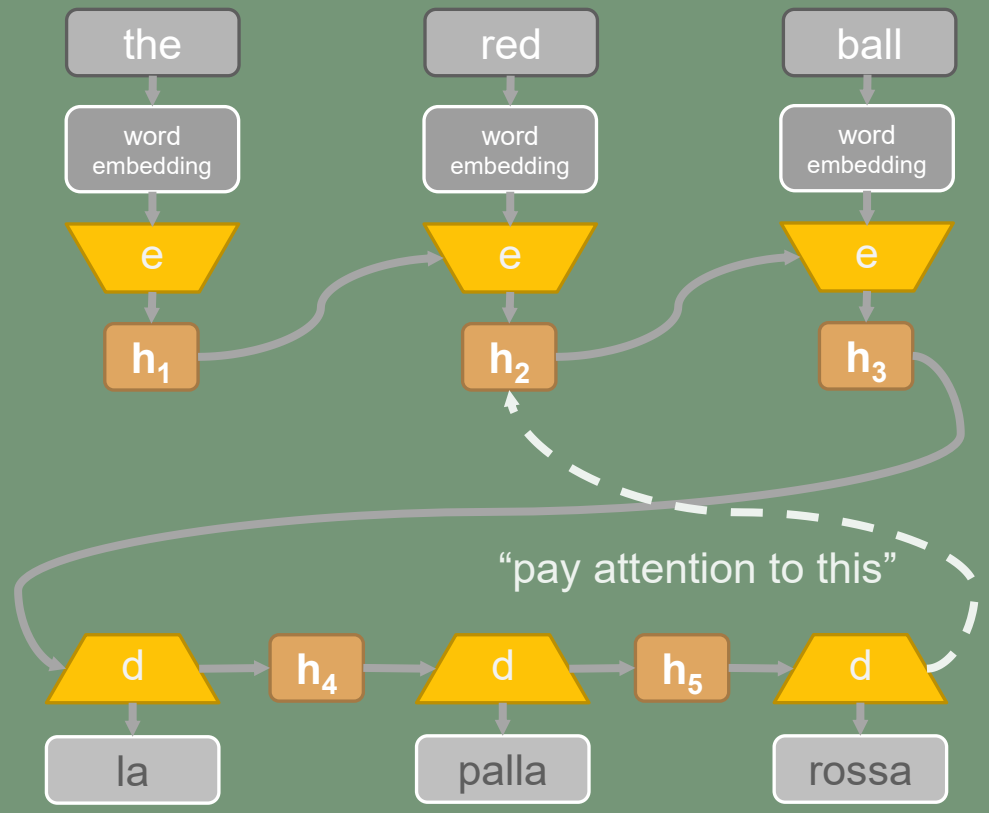


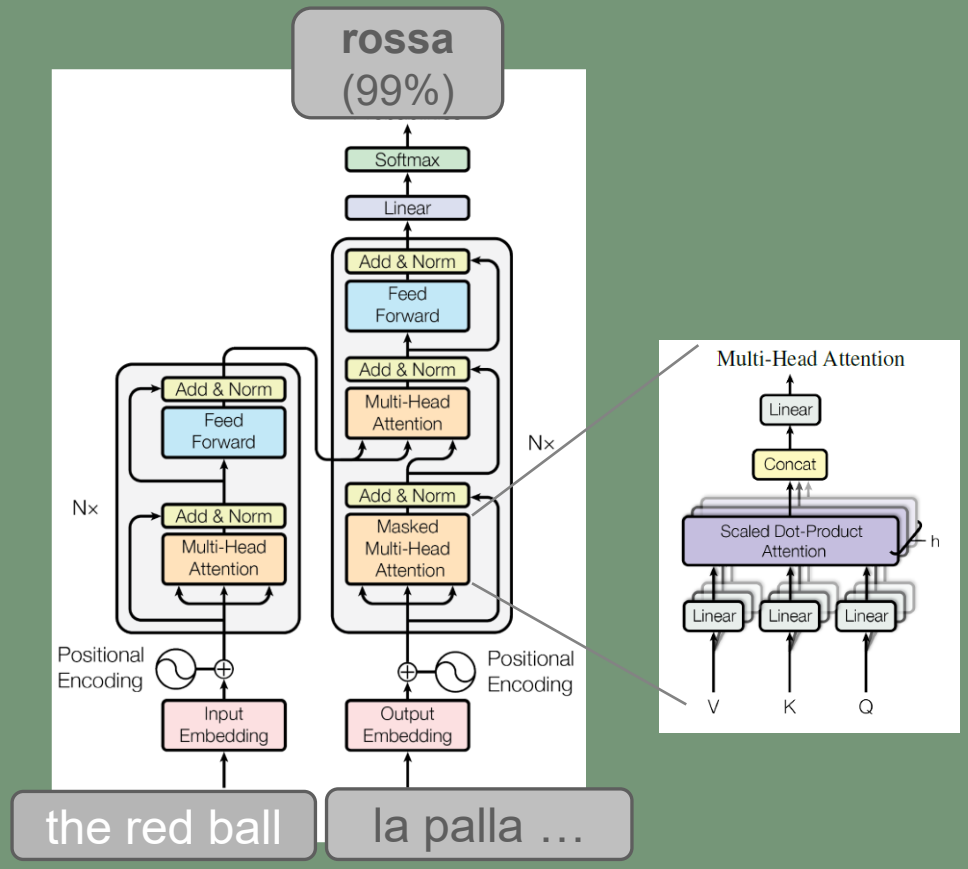
35 years of predictions…

(word embedding & Recurrent Networks)

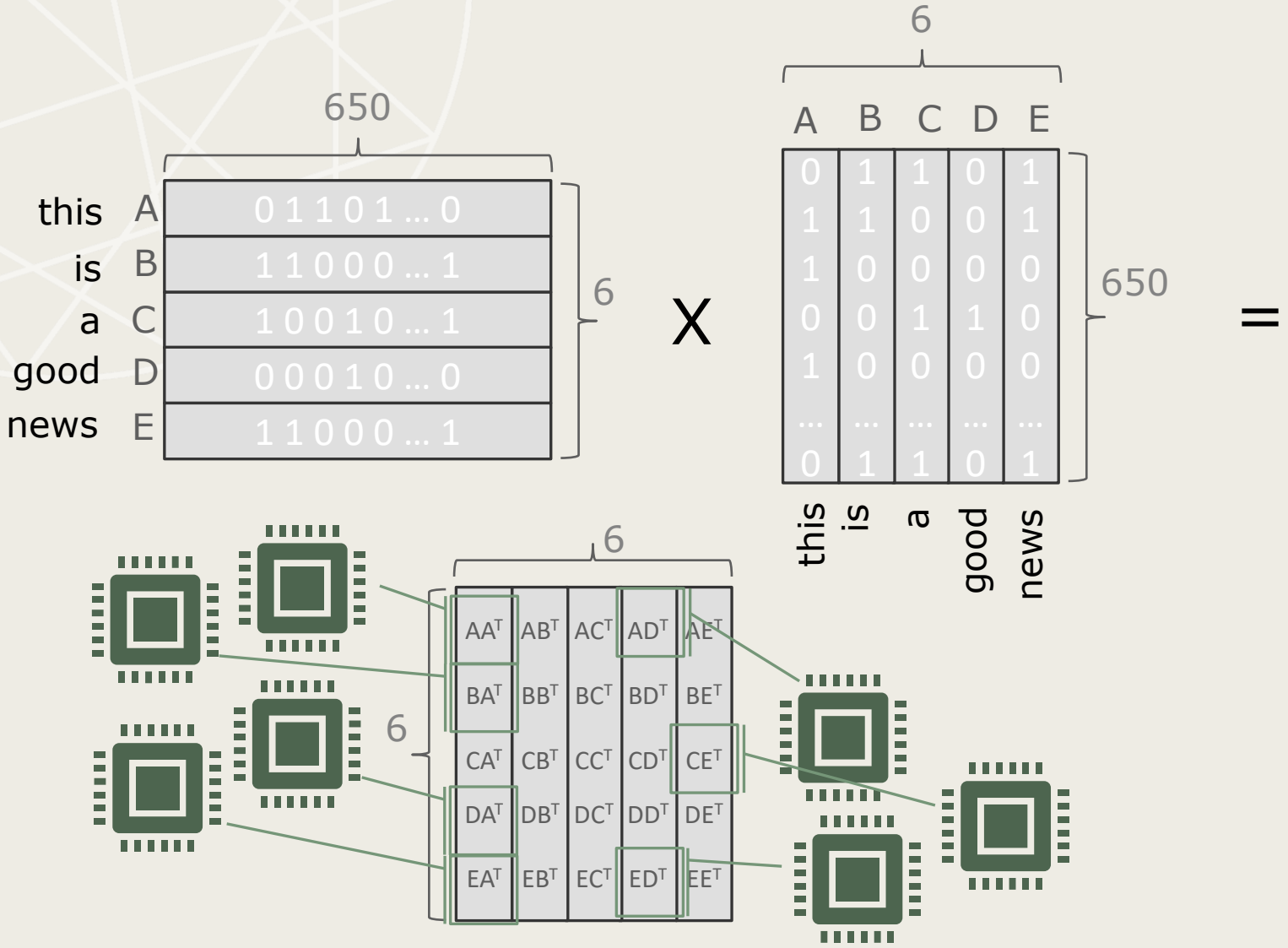
The age of Transformers:
"Attention is **all** you need" Vaswani et al. (2017)

A Machine Translation example

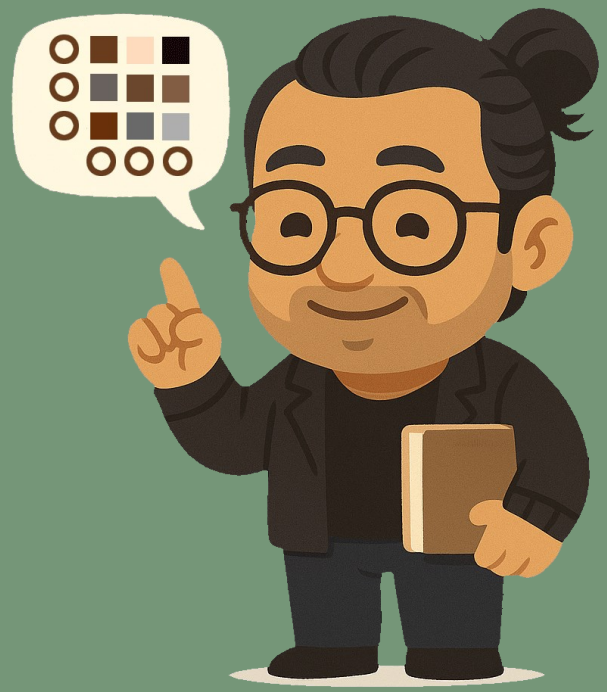




Multiplications, matrices, parallelism & word embedding:



Attention is ALL you need



BabyLM Challenge

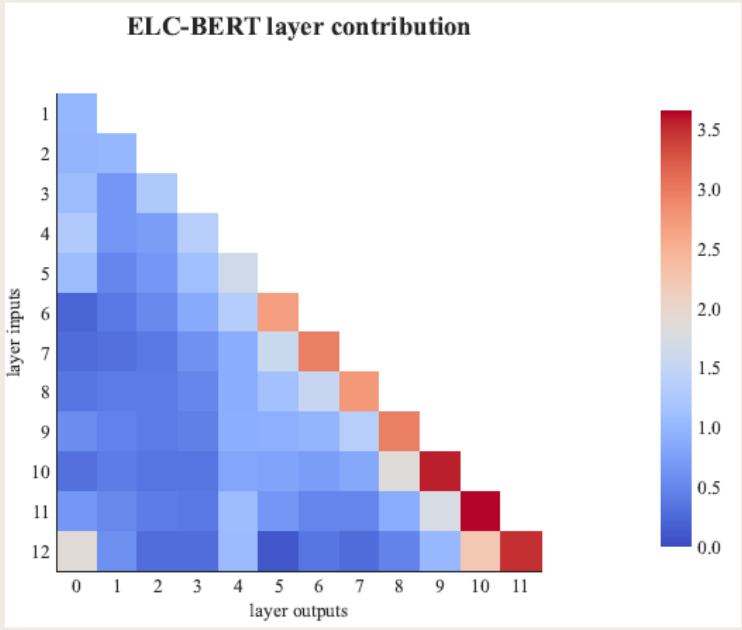
Sample-efficient pretraining on a developmentally plausible corpus

- **3 challenges**
  - **Small-strict**: 10M words corpus
  - **Small**: 100M words corpus
  - **Multimodale**: 50M words corpus + 50M words associated to pictures
- **Evaluation** (accessible platfotm freely available)
  - **BLiMP** (Benchmark for Linguistic Minimal Pairs)
    - E.g. thi sentence is correct vs. *this sentence are correct
  - **(Super) GLUE**
  - **…**

BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

- Winning model in 2024:
  Charpentier and Samuel (2024)
  Not all layers are equally as important: Every Layer Counts BERT



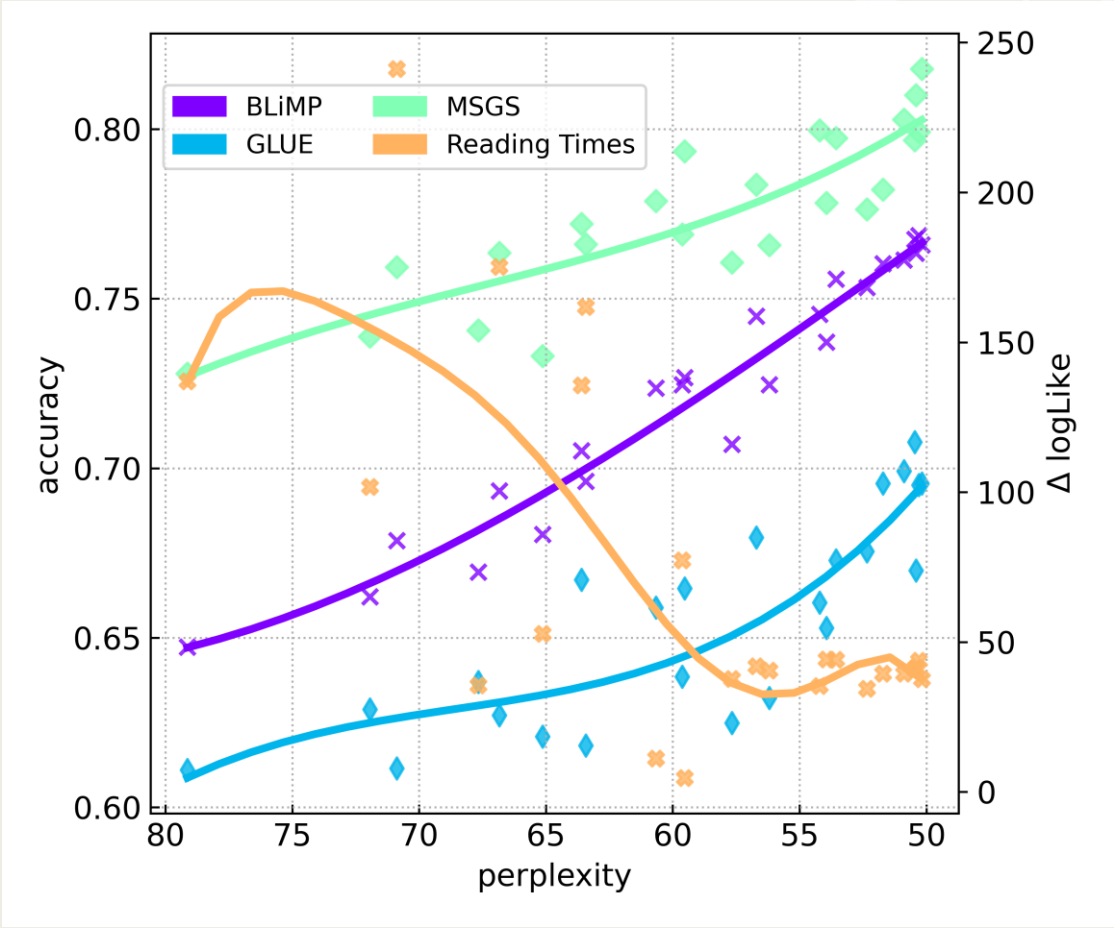
- Hod does it take to get the best coefficients?

A lot!
Restrictions to be used this year: **100M steps** (training sweep) **10 epochs** at most.

Steuer et al. (2023)

Large GPT-like Models are Bad Babies:
A Closer Look at the Relationship
between Linguistic Competence and
Psycholinguistic Measures

- II **best-paper award**:
  GPT-like Models are “Bad Babies”



Were **RNNs All** We Needed?

Feng, Tung, Ahmed, **Bengio (2024)**

- The **gating** mechanism in RNN can be simplified to enable parallel scan (**parallel prefix scan algorithm**, Blelloch, 1990).
- The output must be independent on the temporal scale (remove any link to previous hidden layer for which no prefix scan is available: e.g., *tanh* functions must be removed)

LSTM

$$
\begin{aligned}
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \\
\boldsymbol{o}_t &= \sigma(\text{Linear}_{d_h}([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}])) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \\
\boldsymbol{f}_t &= \sigma(\text{Linear}_{d_h}([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}])) \\
\boldsymbol{i}_t &= \sigma(\text{Linear}_{d_h}([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}])) \\
\tilde{\boldsymbol{c}}_t &= \tanh(\text{Linear}_{d_h}([\boldsymbol{x}_t, \boldsymbol{h}_{t-1}]))
\end{aligned}
$$

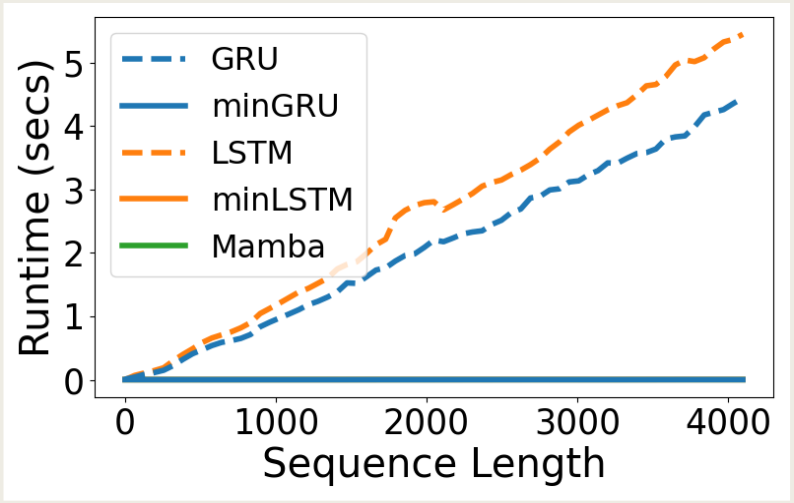
$\Rightarrow$

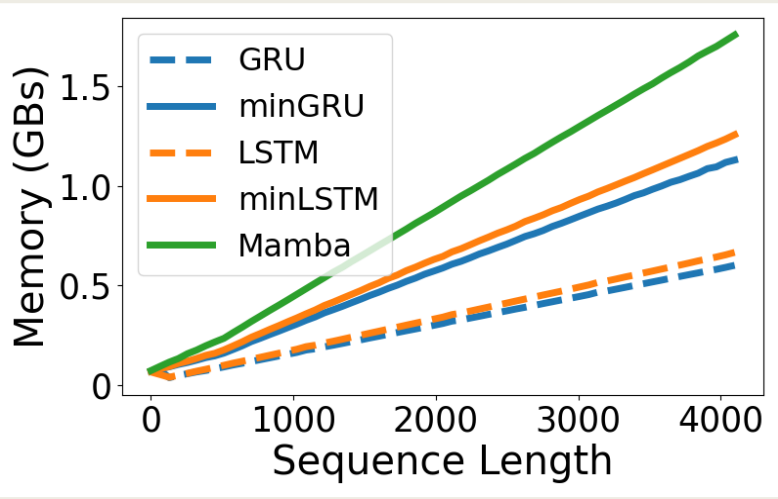
minLSTM

$$
\begin{aligned}
\boldsymbol{h}_t &= \boldsymbol{f}'_t \odot \boldsymbol{h}_{t-1} + \boldsymbol{i}'_t \odot \tilde{\boldsymbol{h}}_t \\
\boldsymbol{f}_t &= \sigma(\text{Linear}_{d_h}(\boldsymbol{x}_t)) \\
\boldsymbol{i}_t &= \sigma(\text{Linear}_{d_h}(\boldsymbol{x}_t)) \\
\tilde{\boldsymbol{h}}_t &= \text{Linear}_{d_h}(\boldsymbol{x}_t) \\
\boldsymbol{f}'_t, \boldsymbol{i}'_t &\leftarrow \frac{\boldsymbol{f}_t}{\boldsymbol{f}_t + \boldsymbol{i}_t}, \frac{\boldsymbol{i}_t}{\boldsymbol{f}_t + \boldsymbol{i}_t}
\end{aligned}
$$

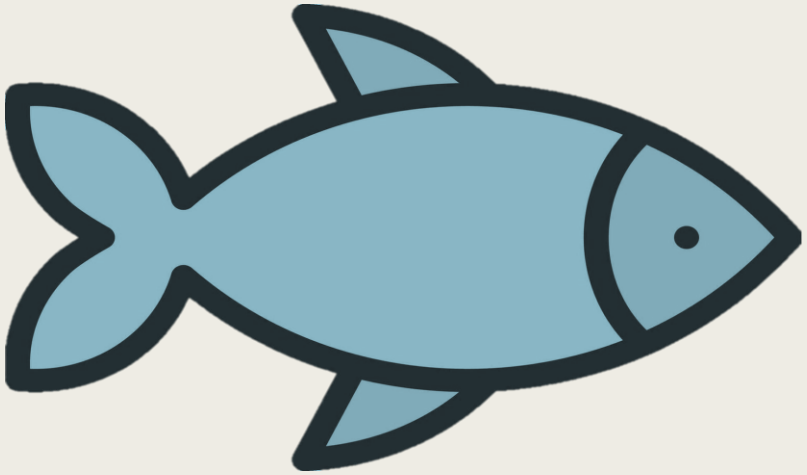
Were **RNNs All** We Needed?

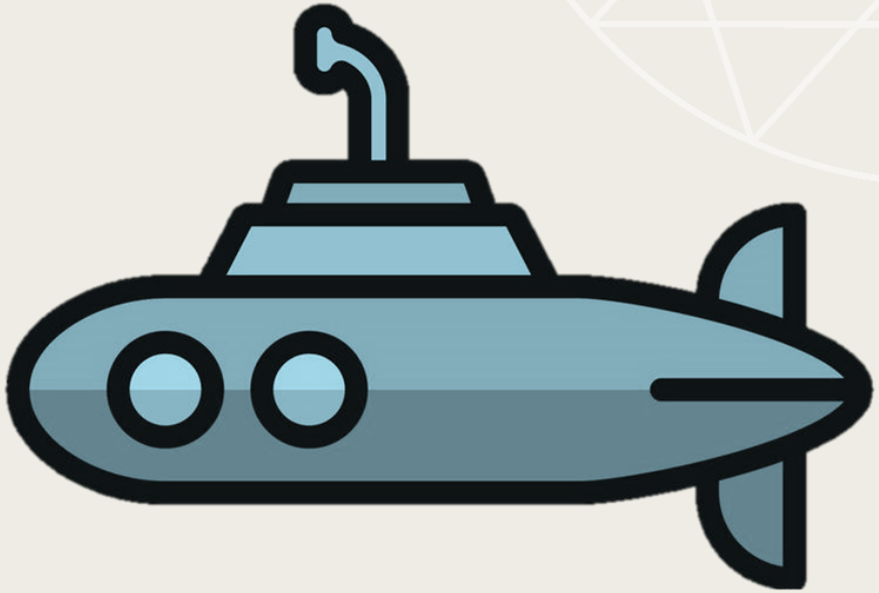
Feng, Tung, Ahmed, **Bengio (2024)**





The **real** and the **artificial**





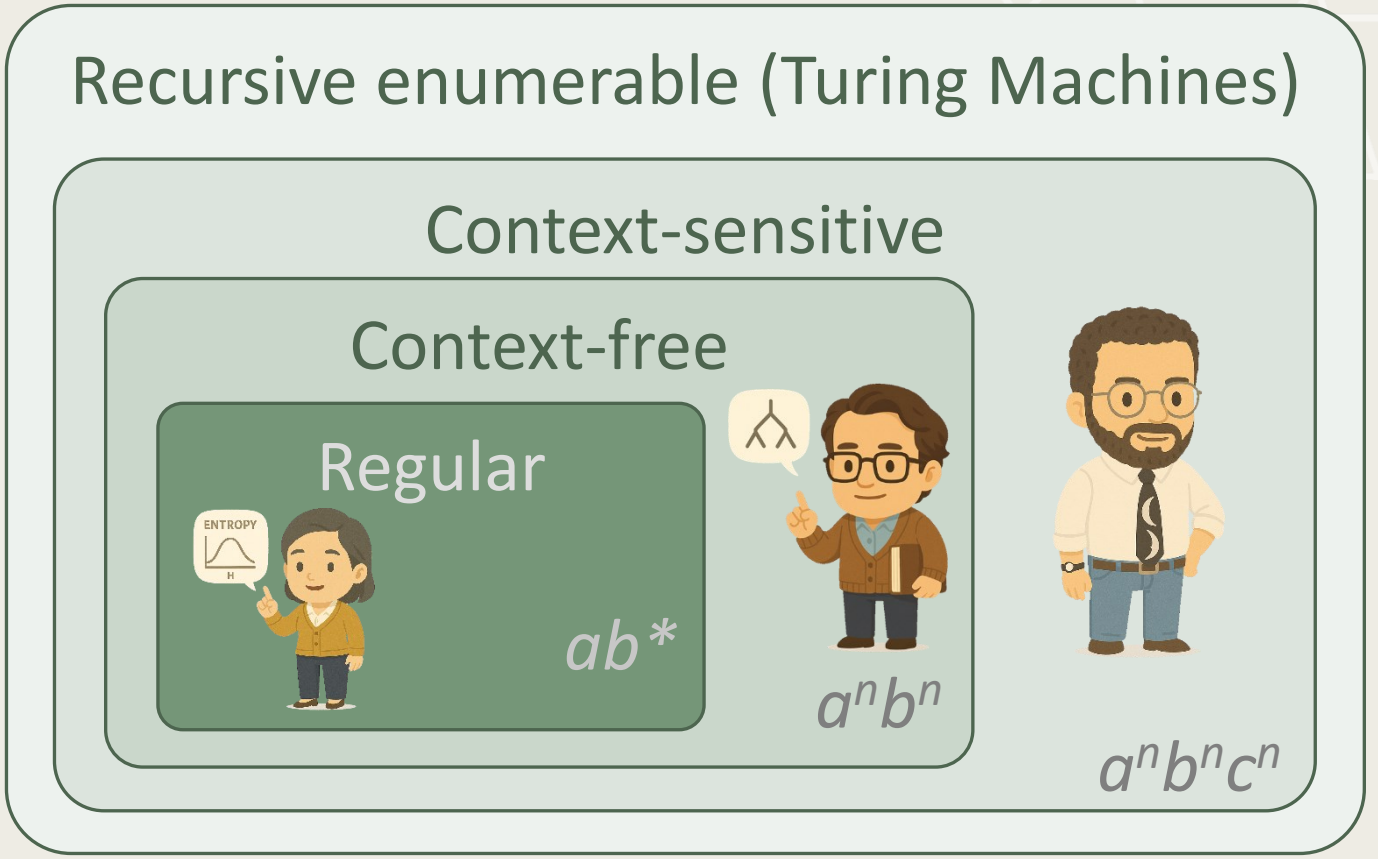
La **Gerarchia** di Chomsky

THE ALGEBRAIC THEORY OF CONTEXT-FREE LANGUAGES*

N. CHOMSKY
Massachusetts Institute of Technology

AND

M. P. SCHÜTZENBERGER
Harvard University



Cosa **sanno** i **bambini?**

Two fundamental «bias» (Rizzi 1990):

- **C-command** – once "boxed" can't enter any relation outside "the box" (i.e., the constituent)
  - [The friends [of grandpa]] **say*says** hy to everyone
- **Locality** – syntactic relations are **respectful** and **lazy**: they don't cross items that can enter the relation, when these items are closer (more local) and available:
  - **What**$_i$ do you think *he* ate _$_i$?
  - ***What**$_i$ do you think **who** ate _$_i$?
  - **Who**$_i$ do you think _$_i$ ate **what**?

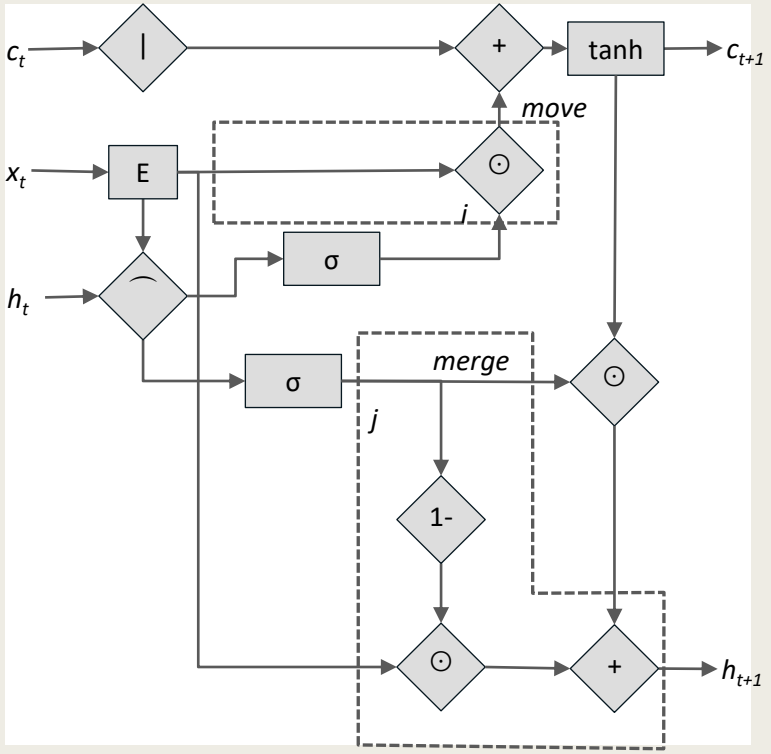
A “special” network

Fusco et al. 2024 @ Clic-IT

Our architecture

(expectation-based Minimalist Grammar Recursive Neural Network - **eMG-RNN**)

Architettura



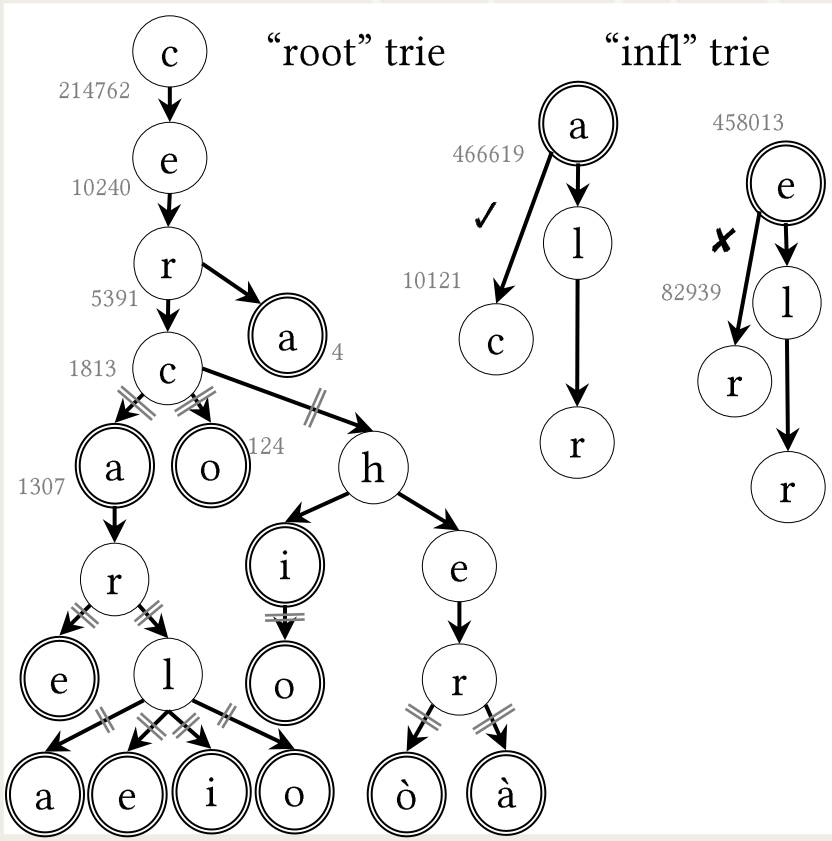
BLiMP performance

		eMG-RNN				
	LSTM	1	2	3	F-M	F-N
Ana. agr	0.67	0.82	0.76	0.77	0.88	0.81
Arg. str	0.56	0.65	0.64	0.63	0.64	0.66
Binding	0.54	0.69	0.66	0.63	0.57	0.65
Ctrl. / Rais.	0.59	0.58	0.59	0.60	0.58	0.60
D-N agr	0.57	0.67	0.63	0.67	0.68	0.68
Ellipsis	0.41	0.24	0.30	0.21	0.42	0.39
Filler. gap	0.55	0.64	0.60	0.47	0.48	0.65
Irregular	0.54	0.58	0.69	0.60	0.60	0.58
Island	0.54	0.58	0.54	0.53	0.50	0.62
Npi	0.45	0.33	0.50	0.55	0.32	0.31
Quantifiers	0.57	0.55	0.53	0.53	0.53	0.57
S-V agr	0.50	0.52	0.52	0.52	0.55	0.53
Overall	0.54	0.58	0.58	0.57	0.55	0.59

The “tokenization” step

eMG-RNN

- Byte Pair Encoding (**BPE**)
  - *low (5 times), newer (4 times), higher (9 times)*
  - e, g, h, i, l, n, o, r, w, _
  - *e r* (13) *r _* (13) *e w* (4) …
  - e, g, h, I, l, n, o, r, w, _, **er**
- MorPiece (**MoP**) based on Tolerance Principle (Yang 2016): a rule is productive if applied at least ***n/ln(n)*** times



NETS
IUSS Laboratory
for **Ne**urolinguistics,
Computational
Linguistics &
Theoretical **S**yntax

IUSS
Scuola Universitaria Superiore Pavia

CRISTIANO CHESI

NETS @ IUSS PAVIA

What do “we need”, in the end?


S. Rossi


M. Barbini


V. Bressan


A. Zanollo


G. Formichi


S. Neri


A. Fusco


M. L. Piccini Bianchessi


T. Sgrizzi