Arrow of Time for LLMs

ICML 2024

Vassilis Papadopoulos, Jérémie Wenger, Clément Hongler

Language Modeling

Next-Token Prediction (Forward model)

- Estimate $\mathbb{P}\left(X_k = x \mid x_1 \cdots x_{k-1}\right)$ as $\mathbb{P}^{\rightarrow}\left(X_k = x \mid x_1, \cdots, x_{k-1}\right) = p_k^{\rightarrow}(x)$
- Factor sequence probability as $\mathbb{P}^{\rightarrow}\left(x_1, \cdots, x_n\right) = \prod_{k=1}^{n} p_k\left(x_k\right)$

Previous-Token Prediction (Backward model)

- Estimate $\mathbb{P}\left(X_k = x \mid x_{k+1} \cdots x_n\right)$ as $\mathbb{P}^{\leftarrow}\left(X_k = x \mid x_1, \cdots, x_{k-1}\right) = p_k^{\leftarrow}(x)$
- Factor sequence probability as $\mathbb{P}^{\leftarrow}\left(x_1, \cdots, x_n\right) = \prod_{k=1}^{n} p_k^{\leftarrow}\left(x_k\right)$

$$\mathbb{P}\left(X_1 = x_1\right) \times \mathbb{P}\left(X_2 = x_2 \mid x_1\right) \times \cdots \times \mathbb{P}\left(X_n = x_n \mid x_1 \cdots x_{n-1}\right) = \mathbb{P}\left(x_1, \cdots, x_n\right) = \mathbb{P}\left(X_n = x_n\right) \times \mathbb{P}\left(X_{n-1} = x_{n-1} \mid x_n\right) \times \cdots \times \mathbb{P}\left(X_1 = x_1 \mid x_n \cdots x_2\right)$$

Do we have $\mathbb{P}^{\rightarrow} = \mathbb{P}^{\leftarrow}$?

Which model is better?

in estimating the joint distribution $\mathbb{P}$ of a corpus

A. Forward model (FW)

B. Backward model (BW)

C. On a par

D. Different for different languages

Cross-Entropy Loss

- Models are trained with Cross-Entropy loss ℓ_{CE}
- We denote
  - $\ell_{CE}^{\rightarrow} = \sum_{k=1}^{n} - \ln \mathbb{P}^{\rightarrow}\left(X_k = x_k \mid x_1, \cdots, x_{k-1}\right) = - \ln \mathbb{P}^{\rightarrow}\left(X_1 = x_1, \cdots, X_n = x_n\right)$
  - $\ell_{CE}^{\leftarrow} = \sum_{k=1}^{n} - \ln \mathbb{P}^{\leftarrow}\left(X_k = x_k \mid x_n, \cdots, x_{k-1}\right) = - \ln \mathbb{P}^{\leftarrow}\left(X_1 = x_1, \cdots, X_n = x_n\right)$
- In theory we have

$$\partial_{CE}^{\leftrightarrow} = \frac{\mathbb{E}\left[\ell_{CE}^{\leftarrow} - \ell_{CE}^{\rightarrow}\right]}{\frac{1}{2}\left(\mathbb{E}\left[\ell_{\overleftarrow{CE}} + \ell_{\overrightarrow{CE}}\right]\right)}$$

$\partial_{CE}^{\leftrightarrow} > 0 \Leftrightarrow$ FW better

$\partial_{CE}^{\leftrightarrow} < 0 \Leftrightarrow$ BW better

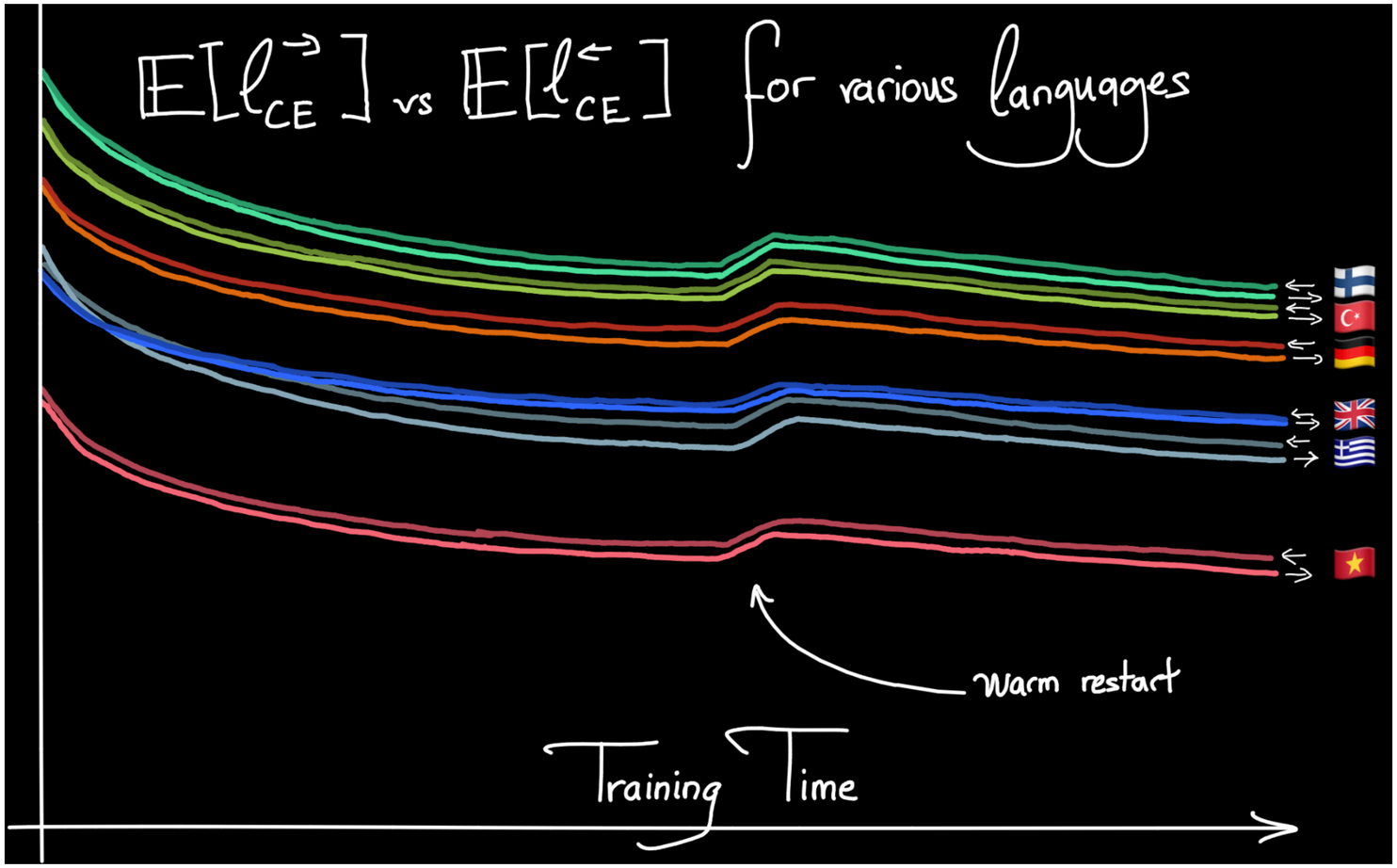
Arrow of Time for languages

Natural Language Experiments

- Dataset: CC100 (>30Gb of text per language)
- Tokenization: Byte-Pair Encoding (BPE), recomputed for each language
- Model: GPT2-Medium (~350M params), 256-token context length

Arrow of Time for languages

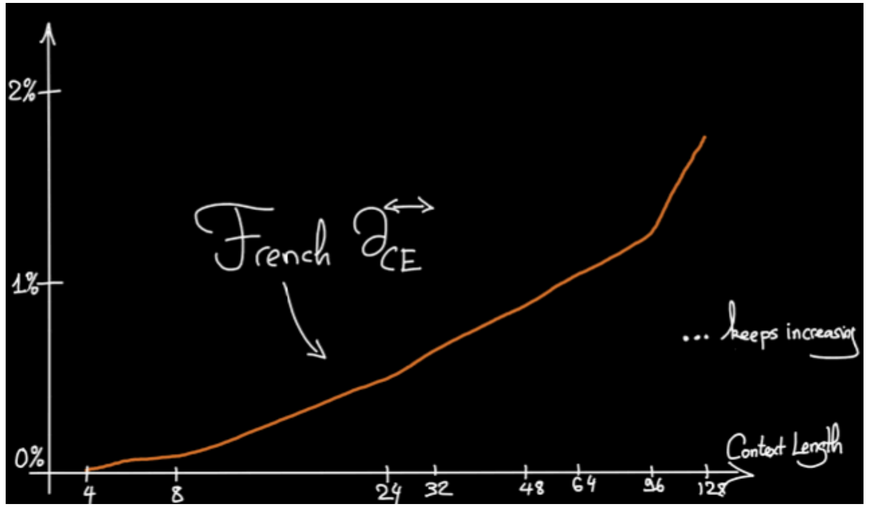
Natural Language Experiments

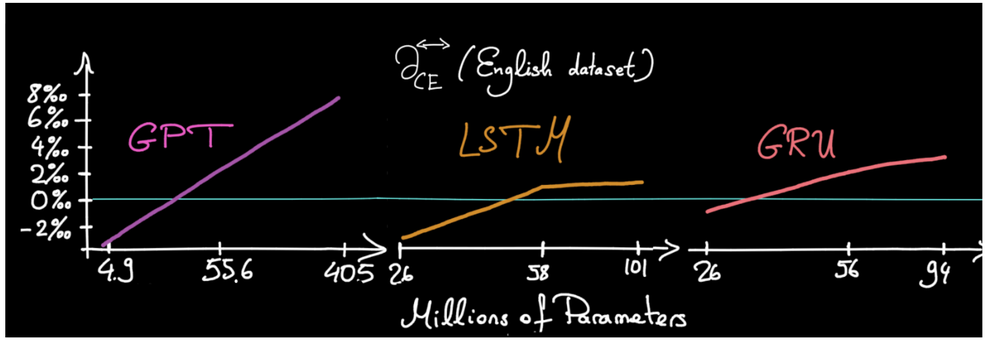


AoT for Languages

Key takeaways (from 100+ experiments)

- FW AoT universality across languages (11 languages)
- AoT $\partial^{\leftrightarrow}_{CE}$ increases with context length
  - long-range correlations essential
- AoT $\partial^{\leftrightarrow}_{CE}$ increases with model size
  - AoT origin is semantic, rather than grammatical
- AoT universal across architectures (LSTMs, GRUs, GPTs)
  - As models get stronger, AoT increases





Origin of AoT

via Computational Hardness

Examples:

$151 \times 353 = 053303$

$367 \times 593 = 217631$

$463 \times 997 = 461611$

- Consider a language of the form $p_1 \times p_2 = n$ with $p_1 < p_2$ primes
- Theoretical FW and BW losses match, as they should:
  - For FW, LHS determines RHS, for BW, RHS determines LHS (**Bijective**)
- For the FW model to do well, it needs to learn the multiplication
- For the BW model to do well, it needs to learn to factor n

	p	q	pq	total
FW	8.98	8.67	4.55	22.20
BW	0.02	8.41	21.56	29.99

Perplexity

Emergence of AoT
via learnability Asymmetry

Examples:

010101 : 101110

011001 : 110101

100010 : 011101

- Example: Linear Languages
  - Dataset $x{:}y$, with x and y both random m-bit strings
  - x,y related by invertible matrices $A^{\leftrightarrows}$ over the field $\mathbb{F}_2$
    - $y = A^{\rightarrow}x$, $x = A^{\leftarrow}y$, and $A^{\leftarrow} = (A^{\rightarrow})^{-1}$
  - FW model learns $A^{\rightarrow}$, BW model learn $A^{\leftarrow}$
    - Sparser matrices are easier to learn
  - Symmetry breaking: $A^{\rightarrow}$ sparse $\Rightarrow$ $A^{\leftarrow}$ typically less sparse

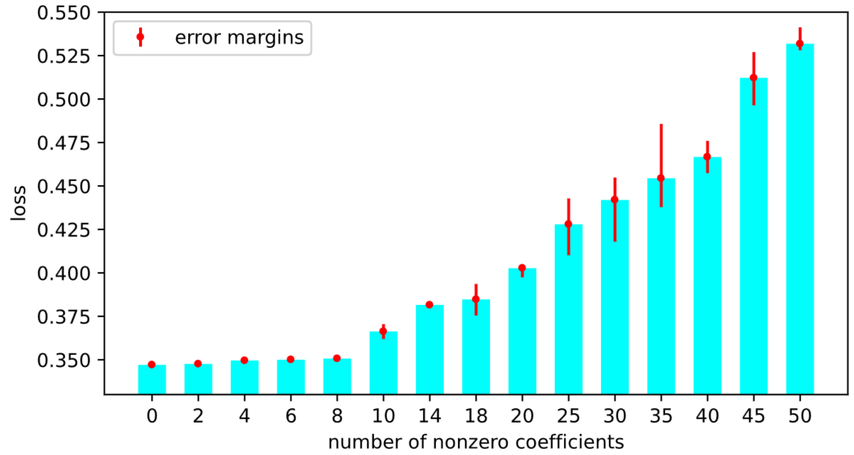


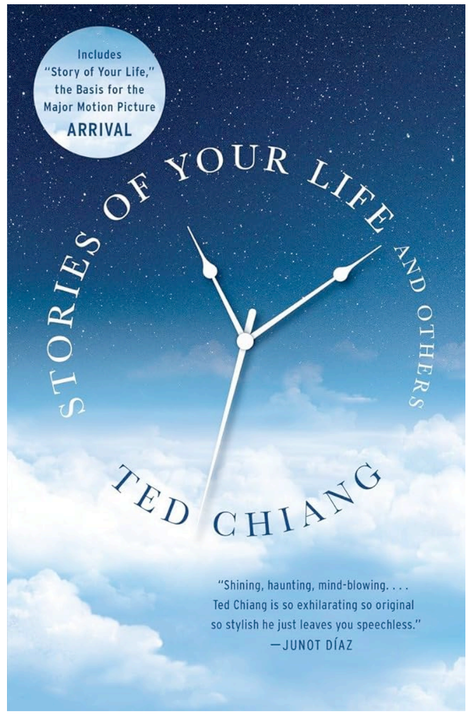
Figure 4: Models loss at the end of training vs $f^{\rightarrow}$ sparsity.

Theory of AoT

Time Symmetry Breaking in Language

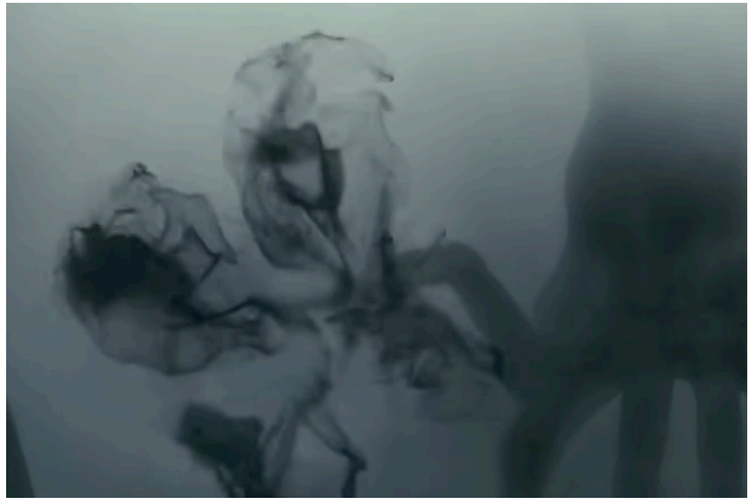
- Alice (human), Bob (human), and Carol (alien) all share a common language
  - Alice and Bob both speak (and think) forwards
  - Carol speaks backwards
- Alice will only share forward-sparse modifications of the language to Bob: that's all he can learn easily
- For Carol, things are harder: the update is not backward-sparse
- AoT emerges from the selection process
  - Alice only communicates sparse-forward updates (because that's what is easy for Bob); typically Caroal struggles more.

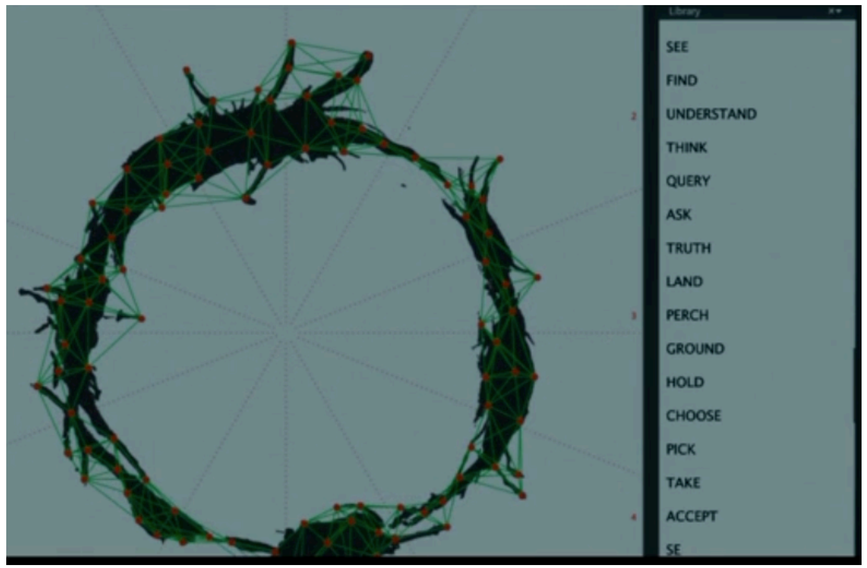
The Language of Heptapod



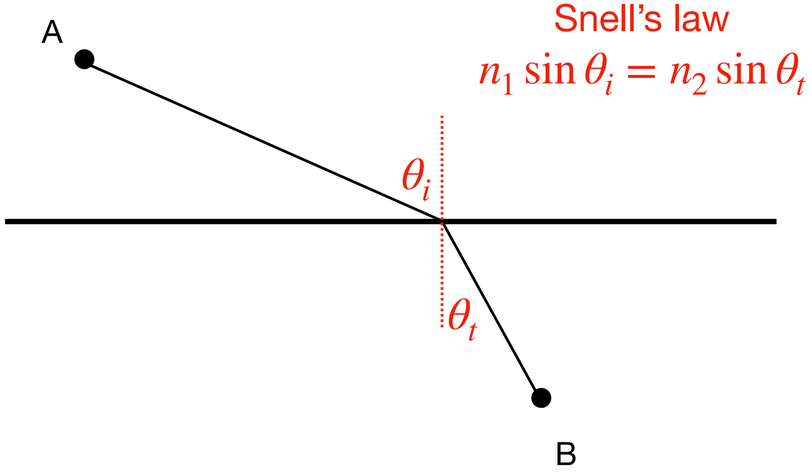
Story of Your Life, 1998

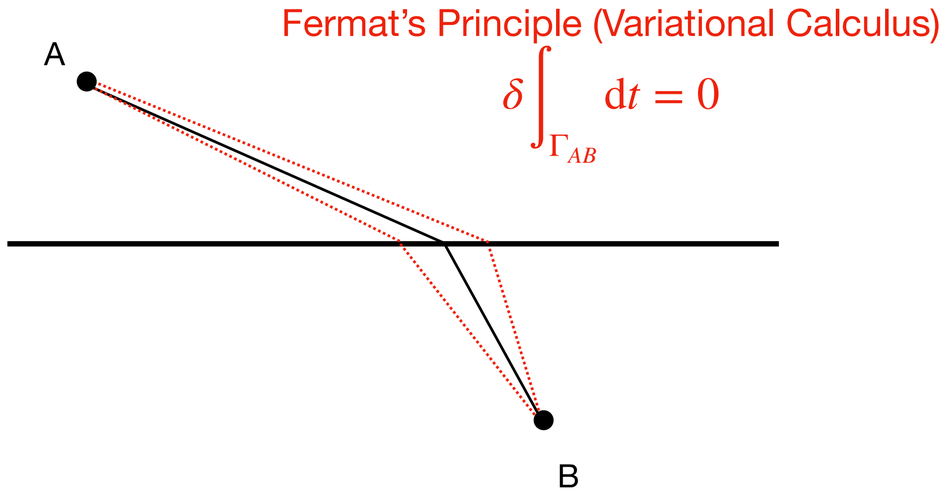
Arrival, 2016





The Language of Heptapod





Comparing that initial stroke with the completed sentence, I realized that the stroke participated in several different clauses of the message. It began in the semagram for "oxygen," as the determinant that distinguished it from certain other elements; then it slid down to become the morpheme of comparison in the description of the two moons' sizes; and lastly it flared out as the arched backbone of the semagram for "ocean." Yet this stroke was a single continuous line, and it was the first one that *Flapper* wrote. That meant the *heptapod* had to know how the entire sentence would be laid out before it could write the very first stroke.

—Ted Chiang, "Story of your life"

Future Research Directions

- Is $\partial^{\leftrightarrow}_{CE} > 0$ linked with intelligence, life?
- AoT on other types of data (code, binaries, DNA, animal sounds)?
- Relation with AoT in thermodynamics
- AoT link with causality
- Are there less data-intensive ways to detect an AoT?
- Scaling laws for $\partial^{\leftrightarrow}_{CE}$?