The omnipresence of the Arrow of Time

By Vassilis Papadopoulos

Based on ‘Arrows of Time for Large Language Models’ (2401.17505), VP, Jérémie Wenger, Clément Hongler

Costa through the e-mail lens

- PhD with Costa : 10/2019-10/2022
- Covid activity : 02/2020-05/2022 (?)
- Thesis writing : 06/2022-09/2022
- Most communication happened by e-mail… and Skype

Subject idea
To Me <vassilis.papadopoulos@phys.ens.fr>

Vassili,

exw tin exis idea: let us consider when x^\prime(\sigma_+) changes sign, as function of M_1 and M_2. This happens when the numerators in (5.6) vanish. On these curves, the solution should jump from a 2 center to a single-center solution (I count black holes as centers).
I dont worry for now about the relation between the Ms and the Ls.

One can solve the above equation easily to find M2/M1 = F(\lambda,\ell_1, \ell_2).
where F is a ratio of quadratic polynomials in \lambda^2.

Now the key question: Is the sign of F fixed, and if so in what ranges for \lambda?
Can you compute this with mathematica? if for example the sign is positive, this means that both Ms must be negative and we are in the vacuum sector. When the sign is negative we can have a transition from H1E1 to H1E2.

Subject Re: Skype today ?
To Me <vassilis.papadopoulos@phys.ens.fr>

vassili molis vlepo to email sou.

avrio exw thyo skype, tin triti to proi k

Costas

Le Vendredi, Octobre 09, 2020 16:40

Γειά σου Κώστα,

θέλεις να κάνουμε Skype σήμερα και

Θα πρεπει να μιλαμε πιο συχνα μεσω σκαιπ, μου φαινεται μεχρι να ηρεμησουν τα πραγματα

we should talk by Skype, at least until things calm down

καλο σαββατοκυριακο

Κωστας

An thes boroume na milisoume meta to seminario tou Maldacena, me skype,

Costas

to afisoume apo vdomada to grafeio.

milisoume avrio me skype (katalava kai ego kapoio pragma kalitera

Κώστα,

σε σχέση με αυτά που είπαμε με Skype. Αν κοιτάξεις τις εξισώσεις (6.

τίποτα δεν εμποδίζει να έχουμε transition E1->E2 smoothly a priori, η

eftheros?

Μπορούμε να κάνουμε ένα γρήγορο skype αύριο; Κοιτάζω το

From Costas Bachas <costas.bachas@phys.ens.fr>
Subject Re: News
To Me <vassilis.papadopoulos@phys.ens.fr>

skype stis 16h30

Subject Re: skype?
To Me <vassilis.papadopoulos@phys.ens.fr>

ok yia 14h00 ?

Subject Re: idea
To Me <vassilis.papadopoulos@phys.ens.fr>

oraia, alla kati den katalavenw

milisoume argotera simera me skype,

Subject Re: General AdS2 brane shape
To Me <vassilis.papadopoulos@phys.ens.fr>

Vassili na milisoume argotera simera me skype, ti ora

me sto skype an

Subject Re: skype?
To Me <vassilis.papadopoulos@phys.ens.fr>

ok yia 14h00 ?

Subject Re: Skype
To Me <vassilis.papadopoulos@phys.ens.fr>

I will call in 5 mins.

My free energies are given in t

Costas

Milame argotera me skype.

Costas

Great explanations

Subject **draft + skype**

Subject **sorry, akomi ena lathos**

Subject **geodesics**

Subject **small typo**

Subject **higher dims**

Subject **almost final draft**

Subject **nea ?**

Subject **bonjour**

Subject **Re: Νεα**

Subject **note**

Subject **notes**

Subject **14h45**

Subject **Re: kink**

Subject **coordinates**

Subject **lists**

Subject **Re: another question**

Subject **Re: skype**

Subject **Re: files**

Subject **google sheet**

Subject **BH**

Subject **one more**

Subject **printer**

Subject **regularity**

Subject **parler**

Subject **nea**

Subject **ok**

Great expectations

Ela Vassili 01.12.2020

o vanR evgale theftero paper. Prepei na to diavasoume prosektika kai i (1) an to overlap einai metrio synexizoume kanonika, i (2) an einai megalo, prepei to paper mas na vgei to grigorotero, kai owi se perisotero apo 10 meres. our paper must come out very soon, no more than 10 days from now

Ta leme argotera simera, pes mou ti ora se volevei to apogevma

Costas

said paper

Phases of Holographic Interfaces

Authors: Constantin Bachas, Vassilis Papadopoulos

Abstract: We compute the phase diagram of the simplest holographic bottom-up m
between three-dimensional Anti-de Sitter (AdS) vacua, anchored on a boundary circ
the intersection of its horizon with the wall, and the fate of inertial observers. We sh

Submitted 1 April, 2021; **v1** submitted 29 January, 2021; **originally announced** January 2021.

Comments: 57 pages, 14 figures. Minor changes

MSC Class: 81T35

Dear Vassilis and Zhongwu,

I think I have completed the writeup of the horizon story to my satisfaction, following the exchanges with Vassilis. It is subtle and very interesting.

11/05/2021

I am attaching the last files (including figures).

With some effort, the paper could be ready by the end of the month. Best, Costas

NB: Interesting seminar at 15h00 today

said paper

Steady States of Holographic Interfaces

Authors: Constantin Bachas, Zhongwu Chen, Vassilis Papadopoulos

Abstract: We find stationary thin-brane geometries that are dual to far-fror
heat at the boundary agrees with the result of CFT and the known energy-t
outgoing excitations the interface produces coarse-grained entropy at a ma

Submitted 21 July, 2021; **v1** submitted 2 July, 2021; **originally announced** July 2021.

Comments: 40 pages, 8 figures Added few discussion paragraphs

Arrow of time

- With no Arrow of Time, we would not be here today celebrating Costa’s career
- I will present recent work, where using Large Language Models, we uncover an universal ‘Arrow of Time’ for languages

Autoregressive models

- We will consider autoregressive (language) models
  - Input is a sequence of discrete tokens $\overrightarrow{X}_n = (x_0 \cdots x_n)$ in V^n, where V is a finite vocabulary set
  - Output is a probability distribution on V, namely the model yields probabilities:
    - $p_i^{\rightarrow}(x) = \mathbb{P}\left(X_i = x \,|\, (x_{i-1}, \cdots, x_0)\right)$
- We can see $\overrightarrow{X}_n$ as a random variable with probability distribution $\mathbb{P}_n$
  - Model is learning probability distribution $\mathbb{P}_n$ decomposed into $\prod_i p_i^{\rightarrow}$

Model training and loss function

- To train such a model one defines a loss function, which the model will attempt to minimise.

- The usual choice (and the best one, see (Hanson, 2012)) is the *cross-entropy loss :*

$$\ell_i^{\rightarrow} = \ell(p_i^{\rightarrow}, x_i) = -\ln p_i^{\rightarrow}(x_i)$$

- Model's prediction is optimised on sequences of fixed length n :

$$\ell_n^{\rightarrow} = \sum_{i=0}^{n} \ell_i^{\rightarrow}$$

Switching it up

Backward model

- The decomposition used by current Language Models (predicting the *next* token) is the most natural, especially if we are making a chatbot
- Still, what about other prediction orders (such as backward)? It is worse, better, or the same ?
- Call *backward* models those which are trained to predict the previous token
  - $p_i^{\leftarrow}(x) = P(X_i = x \,|\, (x_{i+1}, \ldots, x_n))$
  - Token i loss : $\ell_i^{\leftarrow} = -\ln p_i^{\leftarrow}(x_i)$

Switching it up

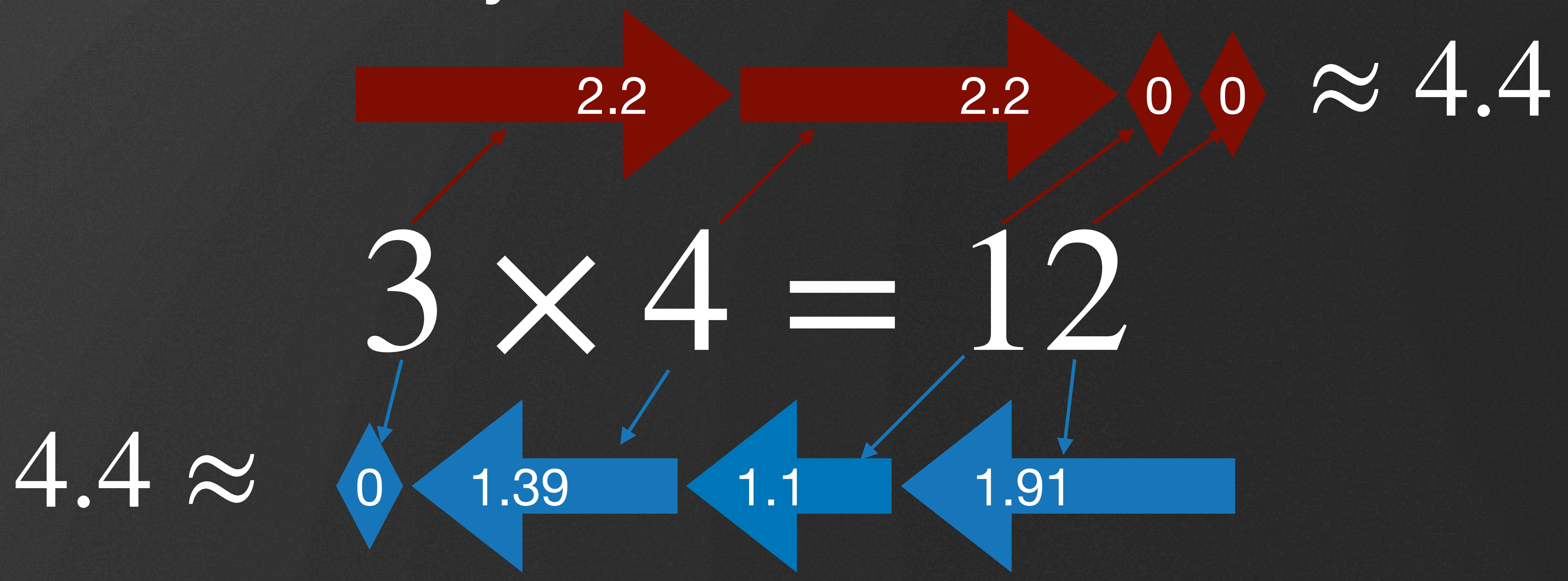
Information content perspective

- Start by comparing the cross-entropy loss of the FW vs BW model on a sequence
- $$\ell_C^{\leftrightarrows} = \sum_i^n \ell_i^{\rightleftarrows} = \sum_i^n - \ln p_i^{\leftrightarrows}(x_i) = - \ln \left(P_n^{\rightleftarrows}(x_1, \ldots, x_n) \right)$$
- Note : since we use the cross-entropy loss, the conditional probabilities ‘cancel out’.
- Because of this, both FW and BW models are trying to approximate the same distribution $P_n(x_1 \ldots, x_n)$, decomposed differently !

Switching it up

Example

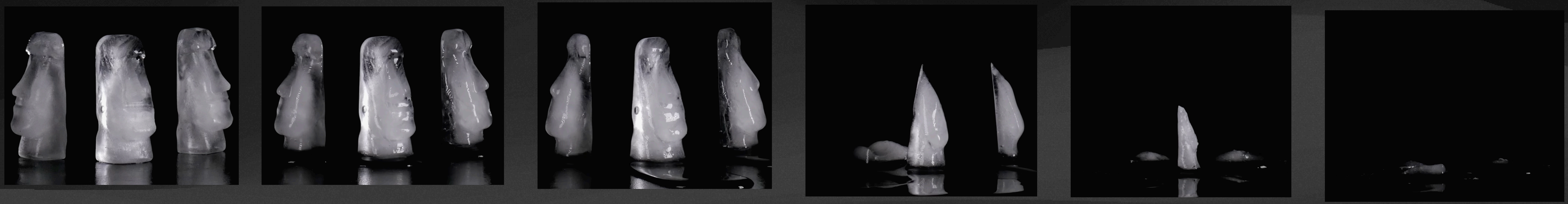
- Consider a dataset containing sentences of the form '$A \times B = CD$', with A, B uniformly sampled digits.
- It seems we are disadvantaged in the BW prediction; since CD can correspond to many $A \times B$ decompositions
- But things turn out to be okay :



Switching it up

Misleading example

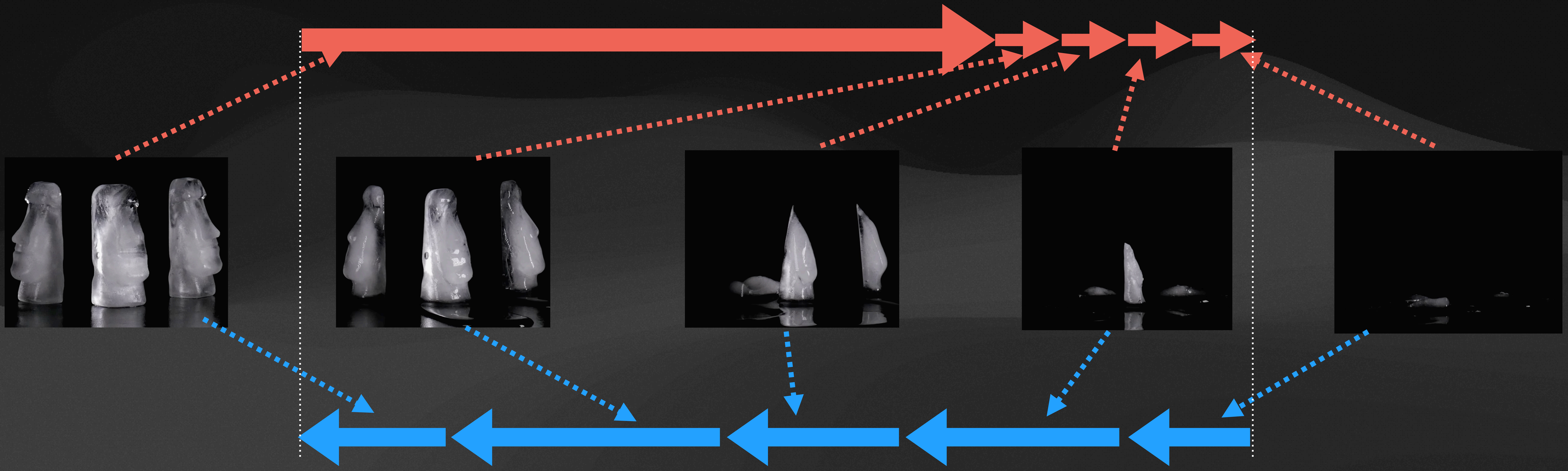
- Consider a dataset composed of snapshot of diverse glass sculptures over time.



- FW direction looks easier, BW looks hard because of entropy increase…

- This is resolved by remembering that the FW model must also predict the first frame.

FW : all loss concentrated on first token



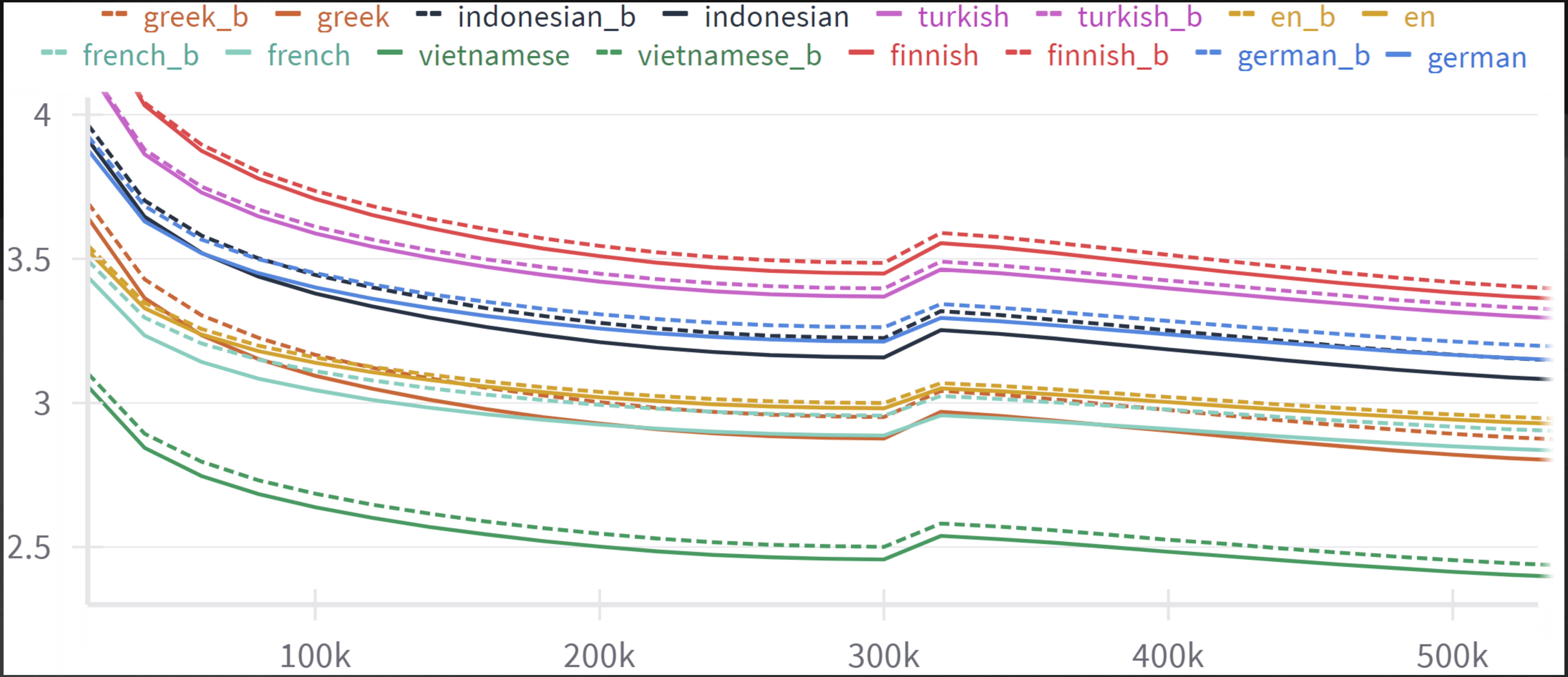
BW : loss more distributed over all tokens

An ‘Arrow of Time’

- We have seen that information-theory wise, FW or BW modelling are equivalent
- Potential differences in $\mathbb{P}_n^{\leftarrow}$ and $\mathbb{P}_n^{\rightarrow}$ thus tell us about an asymmetry of the dataset ($\mathbb{P}_n$), w.r.t. ‘how easy’ it is to learn/model.
- Whenever $\ell_n^{\rightarrow} - \ell_n^{\leftarrow}$ has a *consistent* sign across different experiments, we will say that the dataset has an **Arrow Of Time (AoT)**

Universal AoT for Languages

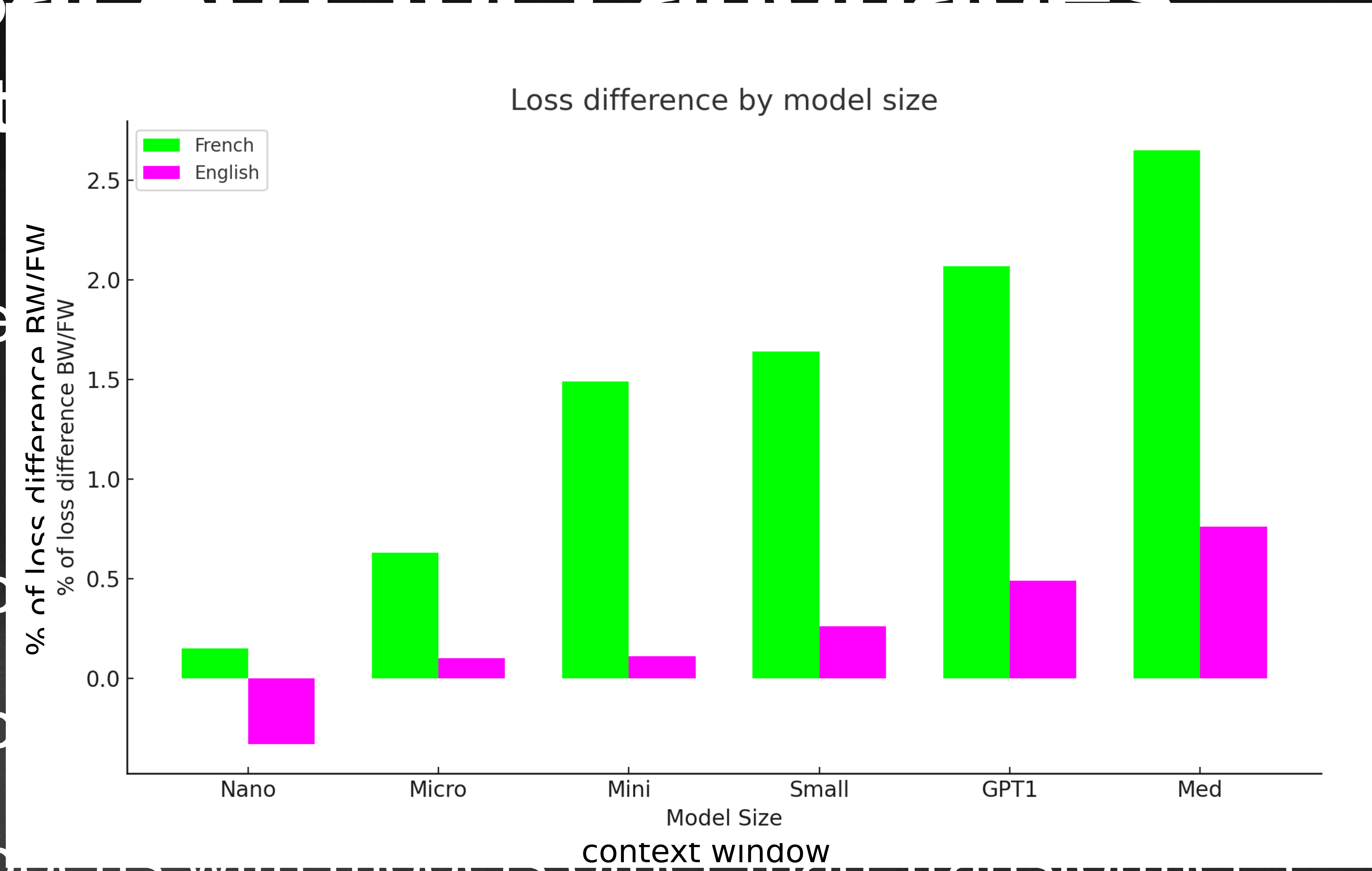
Experiments



Universal AoT for Languages

Experiment

- We observe … s we tested
- More data : …
  - What hap… can access ?
  - What hap… e train ?
  - What happens with models other than Transformer ?



Size	GRU S 4.92M	GRU M 13.7M	GRU L 22.0M	LSTM S 55.6M	LSTM M 162M	LSTM L 405M
Fr-FW	3.905	3.692	3.363	3.901	3.566	3.314
Fr-BW	+0.26%	+0.3%	+0.62%	+0.1%	+0.45%	+0.66%
En-FW	4.030	3.712	3.483	4.015	3.653	3.418
En-BW	-0.07%	+0.22%	+0.34%	-0.27%	+0.11%	+0.15%

Origin of the AoT

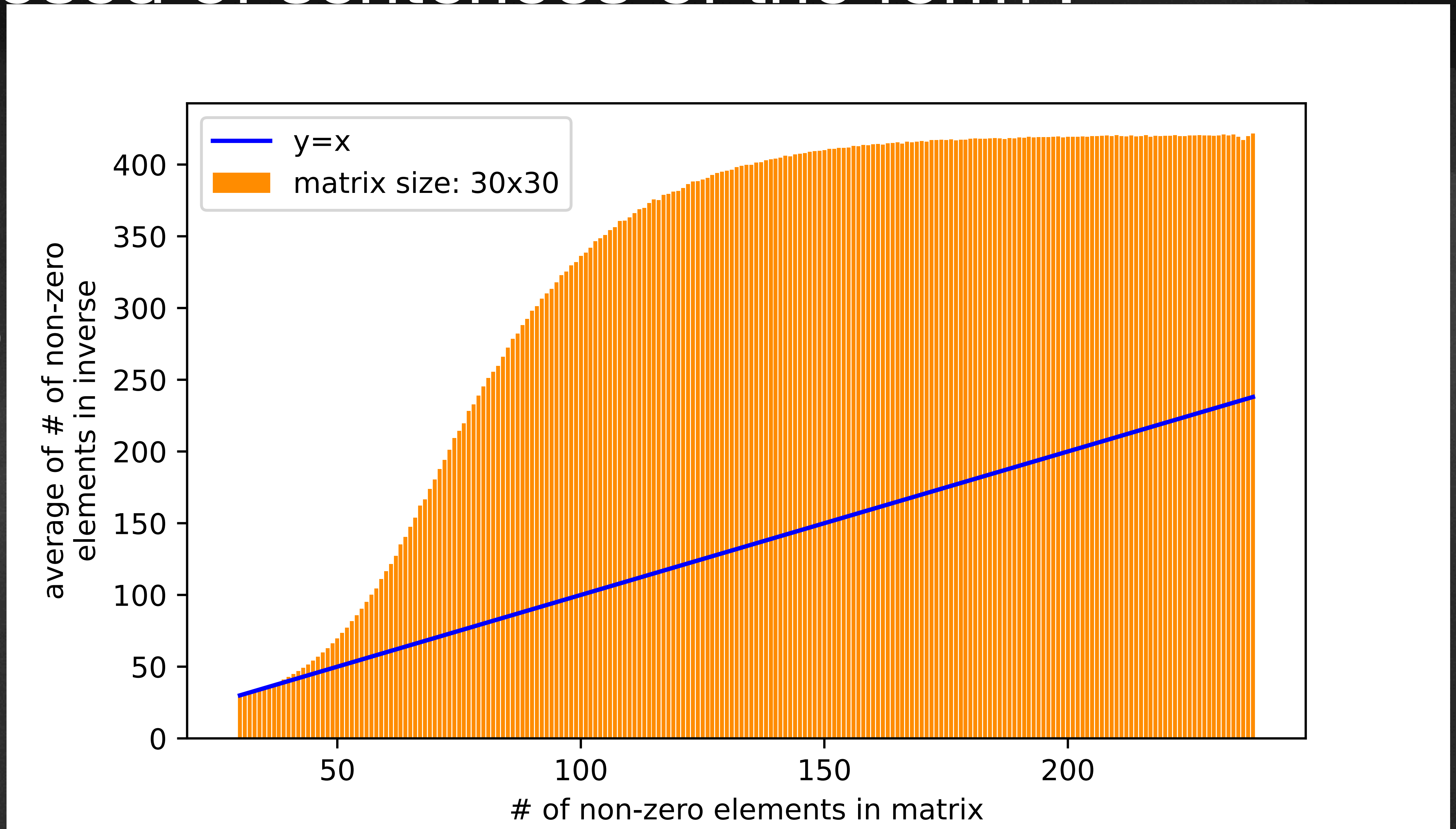
Representability, type 1

- Given the information-theoretic explanation, the AoT must arise due to an asymmetry in how easy are probabilities to learn BW vs FW
- A first asymmetry can come when one direction cannot be *represented* by the model being trained
- Typical example is a dataset of the form $p \times q = pq$, with p, q primes.
- Optimal loss FW : learn multiplication
- Optimal loss BW : learn prime factorisation $\rightarrow$ NP !

Origin of the AoT

Learnability, type 2

- Consider a ‘Linear Language’, composed of sentences of the form :
  - $x \leftrightarrow y,\, x, y \in \mathbb{Z}_2^n$, binary strings
  - $y = Mx,\;\; M \in M(\mathbb{Z}_2)_{n\times n}$, invertib



- To learn the language
  - FW model needs to learn M
  - BW model needs to learn M^{-1}
- If M is sparse, then the FW model’s task is easy !
- If M is sparse and generic, M^{-1} is generally much less sparse !

Origin of AoT

Why forward ?

- Explained existence of AoT, but not its consistent direction in language
- General (speculative) idea :
  - Say Alice wants to teach Bob something new she learned
  - The idea is that she will do this in ‘easy’ steps
    - She will send Bob information in a sparse (i.e. easily learnable) way
    - Given what we know, this makes it so the backwards direction is automatically harder

Possible future directions

- What is the relation with the ‘entropic’ AoT ?
    - E.g., train on the melting sculpture dataset. Can we somehow connect the entropy increase with the AoT ?
    - The token losses distribution in this case suggest a connection with diffusion models
- Can the AoT be a proxy for intelligent processing ?
    - Does code have an AoT (yes)
    - Does DNA have an AoT ?
- Given Costas’s explanations are very easy to understand, is there a higher than normal AoT on a collection of his papers ?