EACL 2024

Tutorial on *Transformer-specific Interpretability,* Part 2:

Measures of Context Mixing

Hosein Mohebbi, Jaap Jumelet, Michael Hanna,
Afra Alishahi, and Willem Zuidema

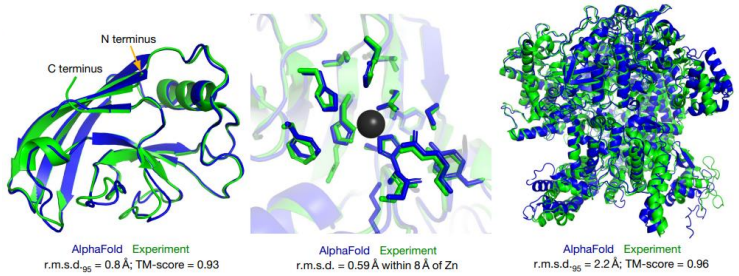
March 21, 2024

Malta

Universiteit van Amsterdam

Waltz of Transformers

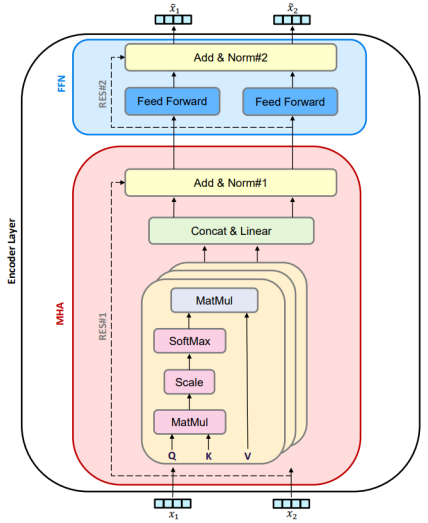
(AlphaFold)



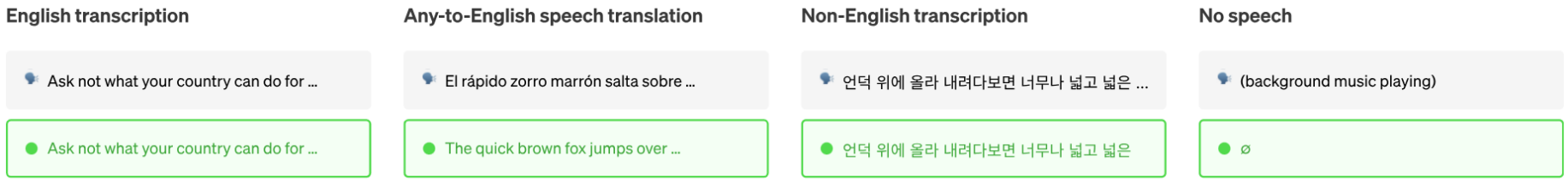
(ChatGPT)

(Stable Diffusion)

(Seamless)



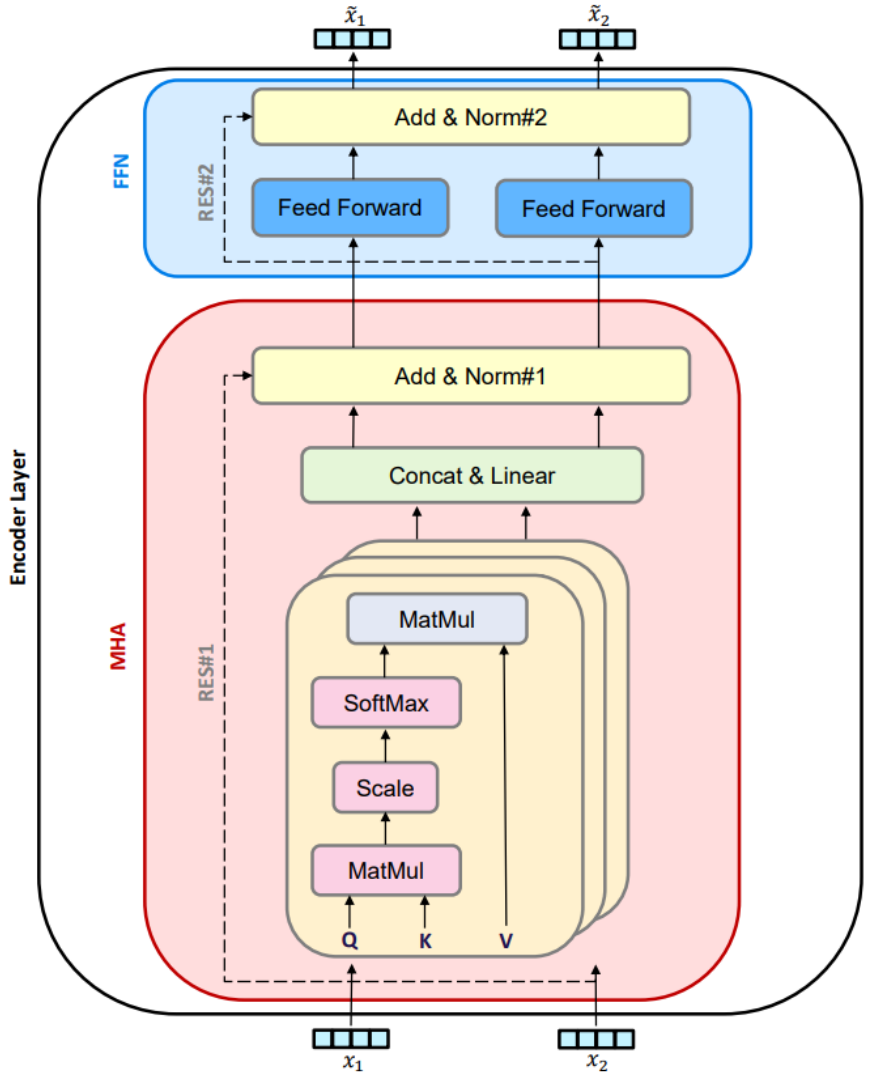
(Sora)



(Whisper)

Mathematics in Transformer

$$(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$$



(Vaswani et al., 2017)

Mathematics in Transformer

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

MHA

$$\left.\begin{aligned} \boldsymbol{q}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{h}} + \boldsymbol{b}_Q^h \\ \boldsymbol{k}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{h}} + \boldsymbol{b}_K^h \end{aligned}\right\} \quad \alpha_{i,j} = \underset{\boldsymbol{x}_j \in \mathcal{X}}{\text{softmax}}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d}}\right) \in \mathbb{R}$$

$$\boldsymbol{v}_i^h = \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{V}}^{\boldsymbol{h}} + \boldsymbol{b}_V^h$$

$$\boldsymbol{z}_i^h = \sum_{j=1}^{n} \alpha_{i,j}^h \boldsymbol{v}_j^h$$

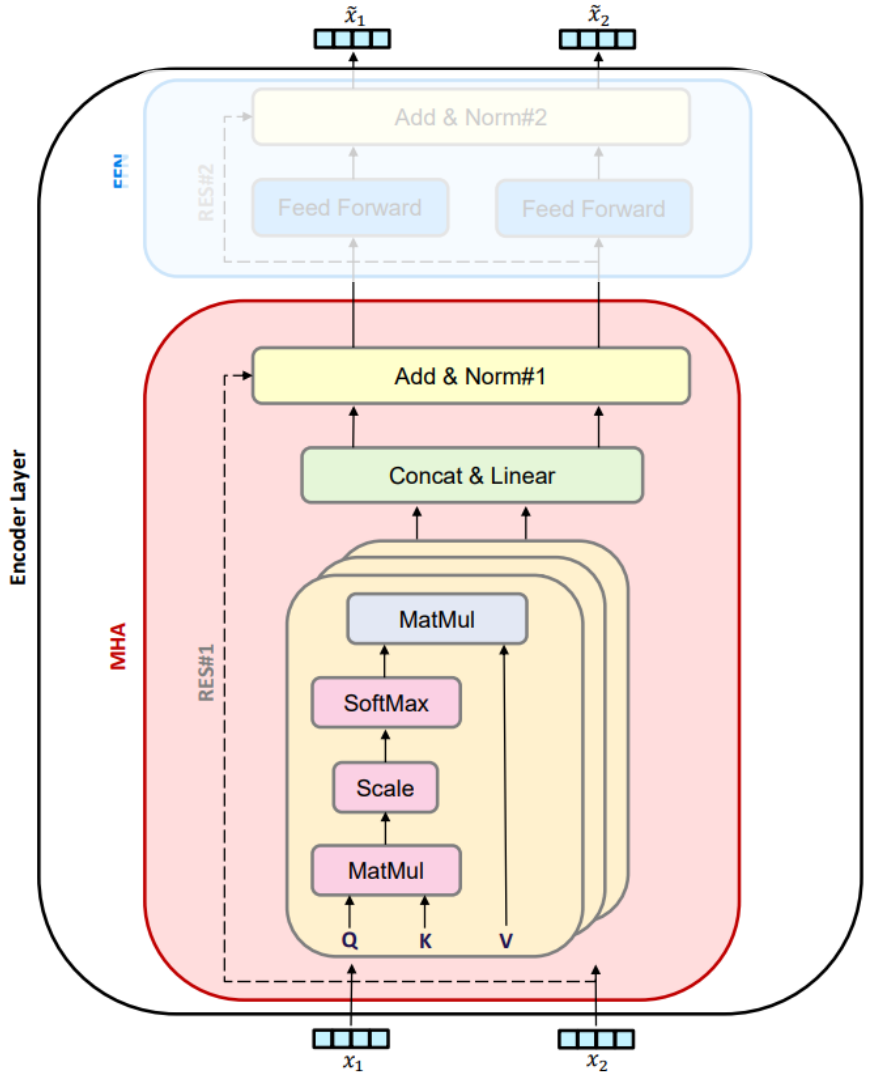
$$\boldsymbol{z}_i = \text{CONCAT}(\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H)\boldsymbol{W}_O$$

$$\boldsymbol{z}_i = \text{LN}_{\text{MHA}}(\boldsymbol{z}_i + \boldsymbol{x}_i)$$

FFN

$$\tilde{\boldsymbol{x}}_i = \max(0, \boldsymbol{z}_i \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$\tilde{\boldsymbol{x}}_i = \text{LN}_{\text{FFN}}(\tilde{\boldsymbol{x}}_i + \boldsymbol{z}_i)$$



Mathematics in Transformer

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

MHA

$$\left.\begin{aligned} \boldsymbol{q}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}}^h + \boldsymbol{b}_Q^h \\ \boldsymbol{k}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{K}}^h + \boldsymbol{b}_K^h \end{aligned}\right\} \alpha_{i,j} = \underset{\boldsymbol{x}_j \in \mathcal{X}}{\text{softmax}}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d}}\right) \in \mathbb{R}$$

$$\boldsymbol{v}_i^h = \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{V}}^h + \boldsymbol{b}_V^h$$

$$\boldsymbol{z}_i^h = \sum_{j=1}^{n} \alpha_{i,j}^h \boldsymbol{v}_j^h$$

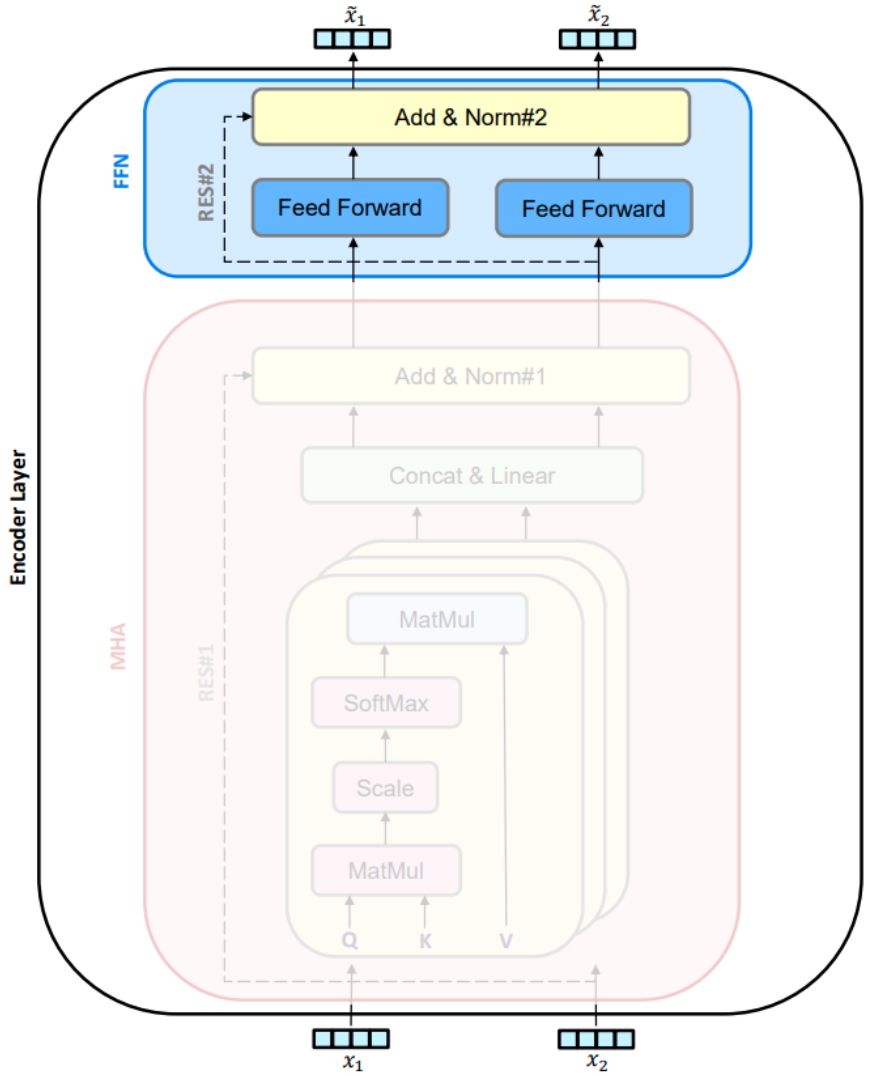
$$\boldsymbol{z}_i = \text{CONCAT}(\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H)\boldsymbol{W}_O$$

$$\boldsymbol{z}_i = \text{LN}_{\text{MHA}}(\boldsymbol{z}_i + \boldsymbol{x}_i)$$

FFN

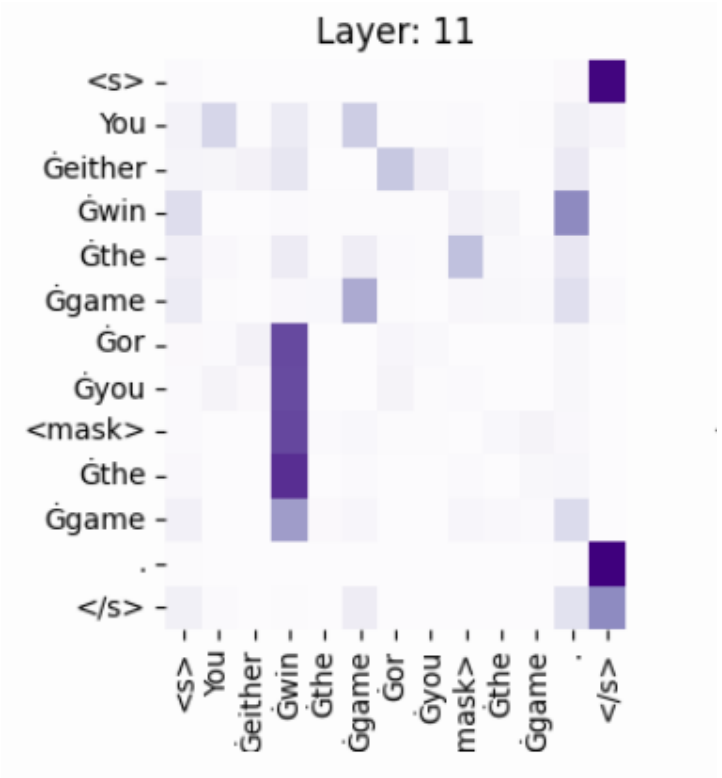
$$\tilde{\boldsymbol{x}}_i = \max(0, \boldsymbol{z}_i \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$\tilde{\boldsymbol{x}}_i = \text{LN}_{\text{FFN}}(\tilde{\boldsymbol{x}}_i + \boldsymbol{z}_i)$$

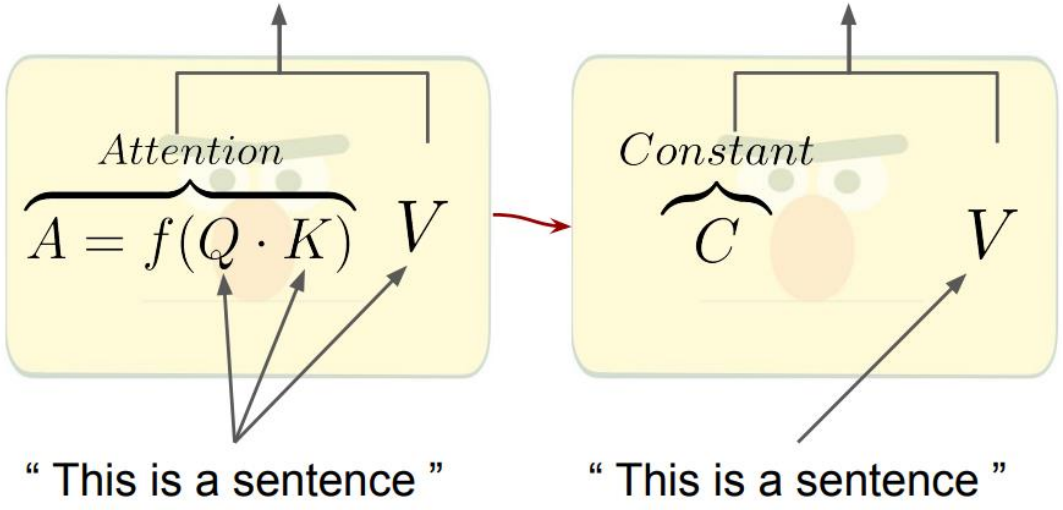


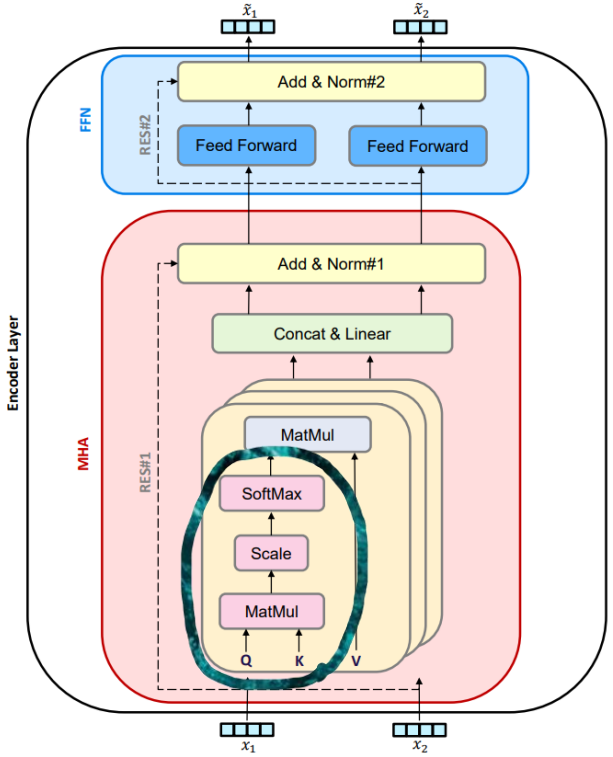
What is Context Mixing?

Either you win the game or you <mask> the game.



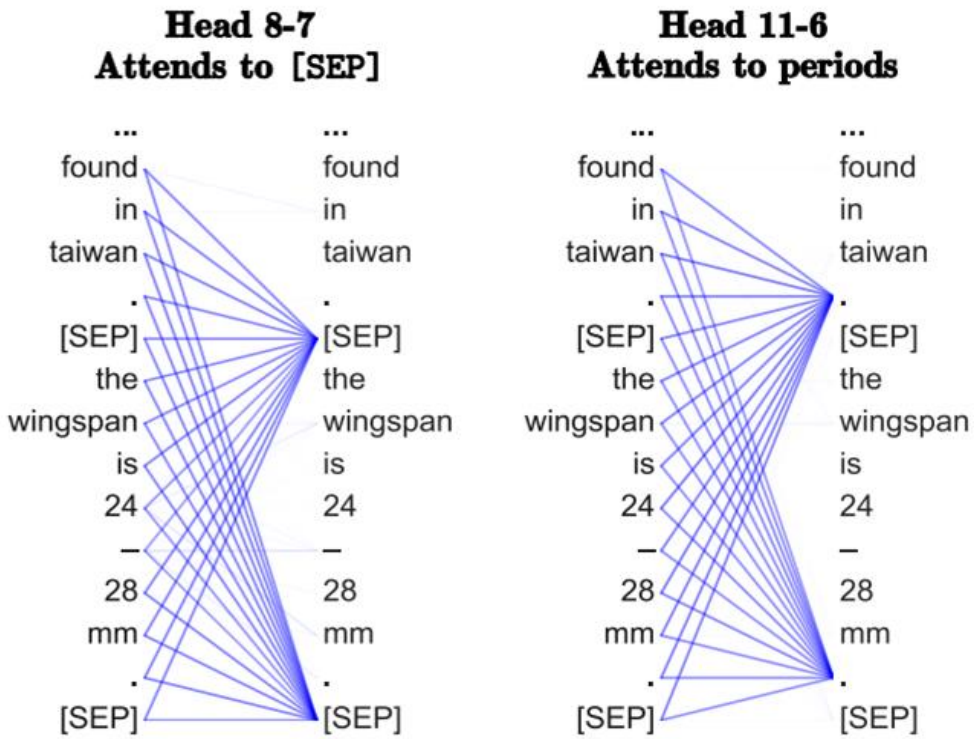
Self-Attention





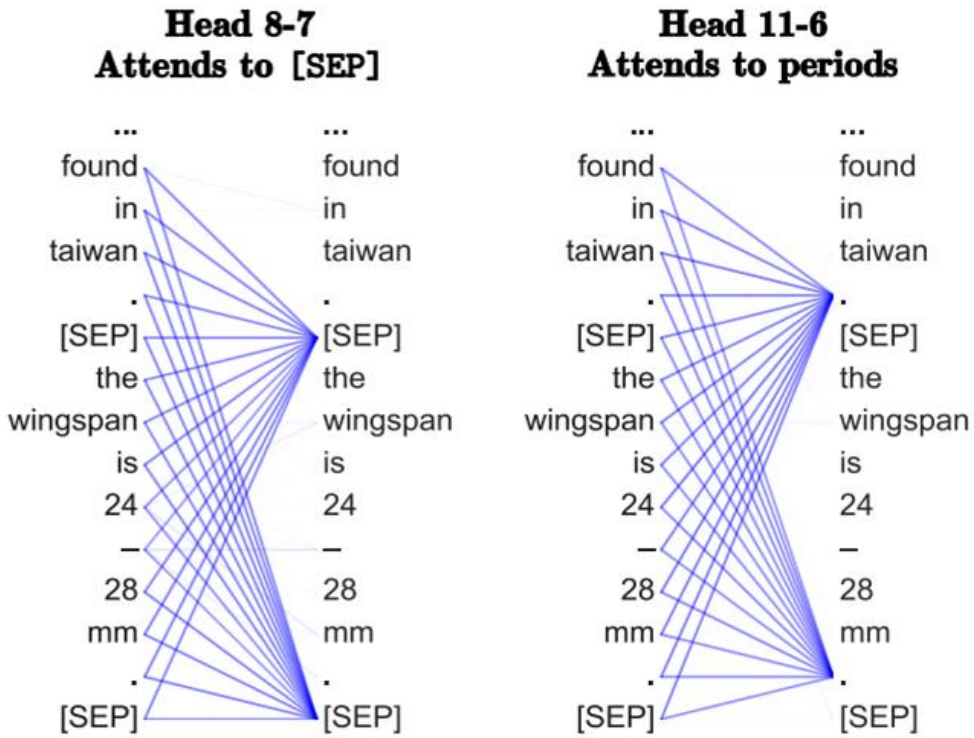
(Hassid et al., 2020)

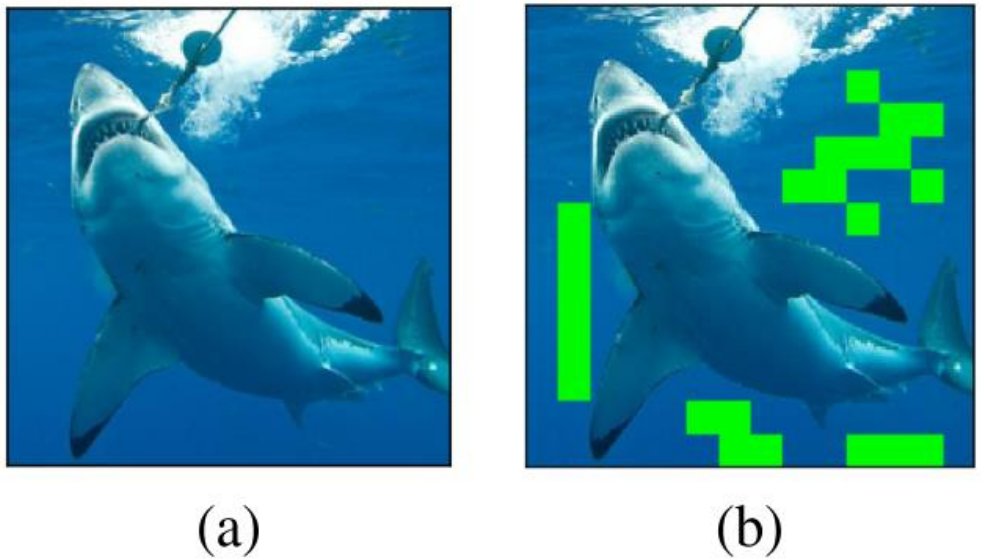
Self-Attention



(Clark et al., 2019)

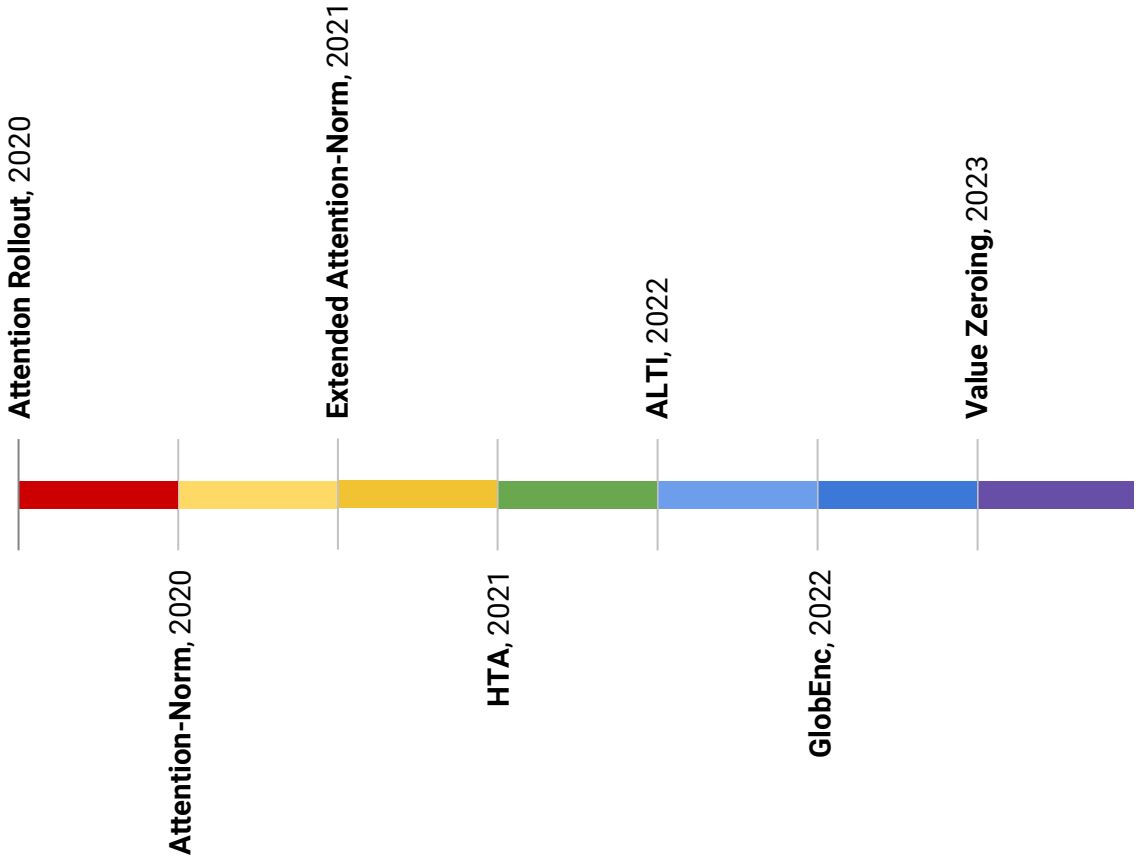
Self-Attention

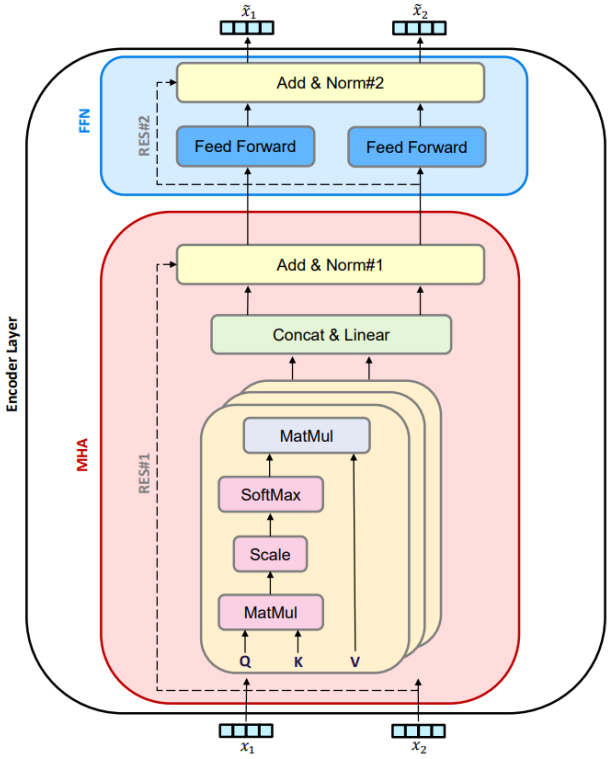




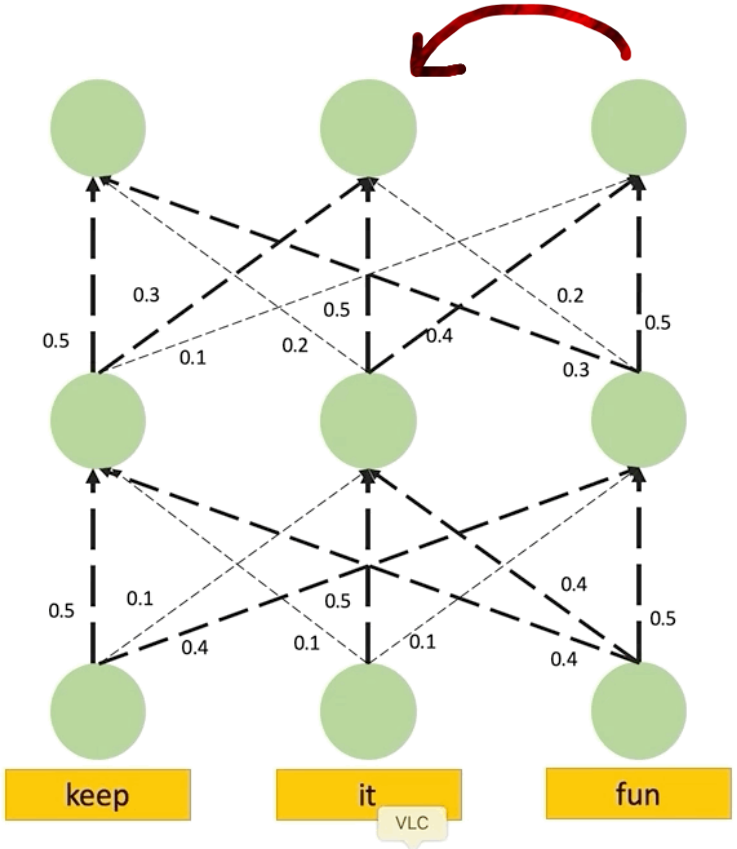
(Clark et al., 2019; Bondarenko et al., 2023)

Measures of Context Mixing



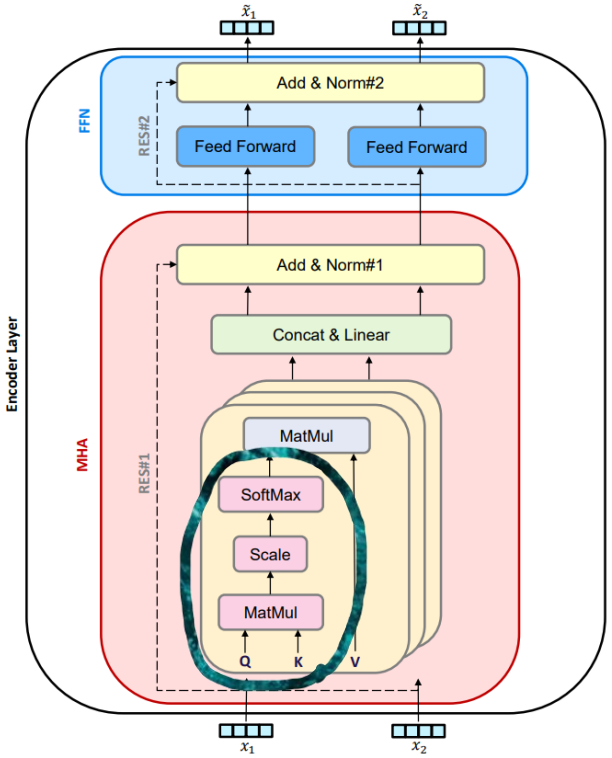


Attention-Rollout



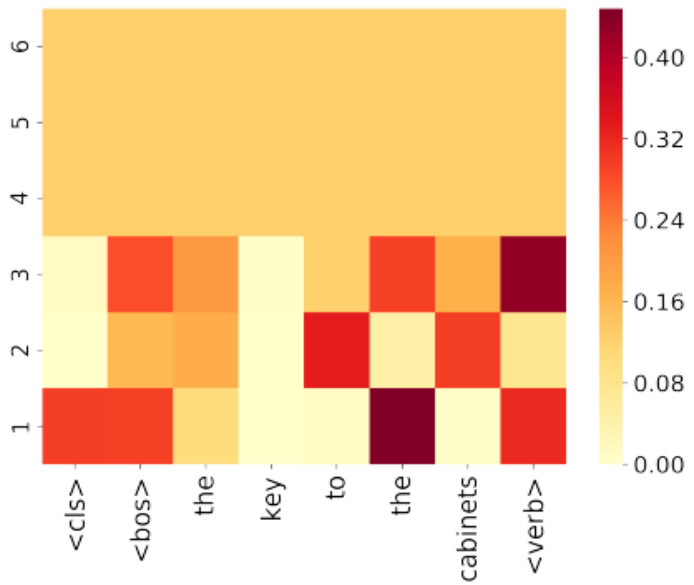
$$\tilde{\boldsymbol{A}}_\ell = \begin{cases} \hat{\boldsymbol{A}}_\ell \tilde{\boldsymbol{A}}_{\ell-1} & \ell > 1 \\ \hat{\boldsymbol{A}}_\ell & \ell = 1 \end{cases}$$

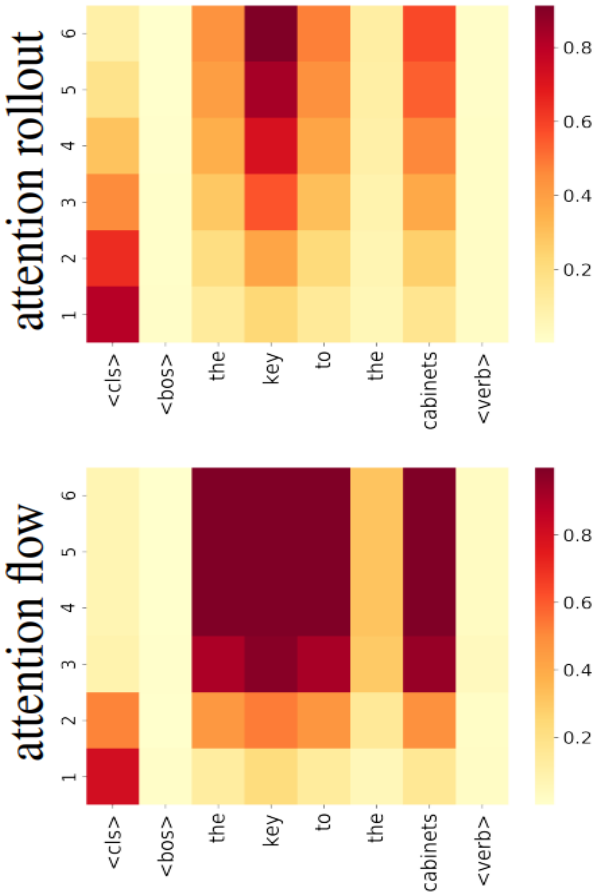
$$\hat{\boldsymbol{A}}_\ell = 0.5\bar{\boldsymbol{A}}_\ell + 0.5\boldsymbol{I}$$

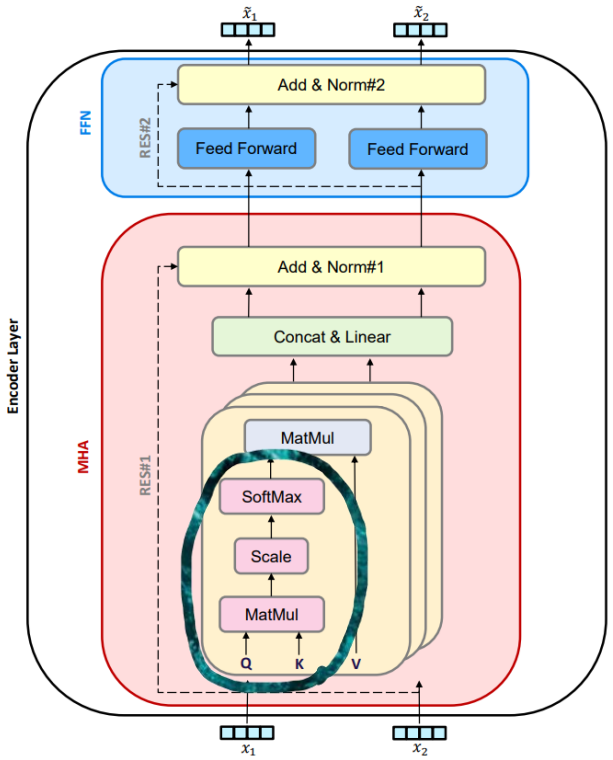


(Abnar & Zuidema, 2020)

Attention-Rollout

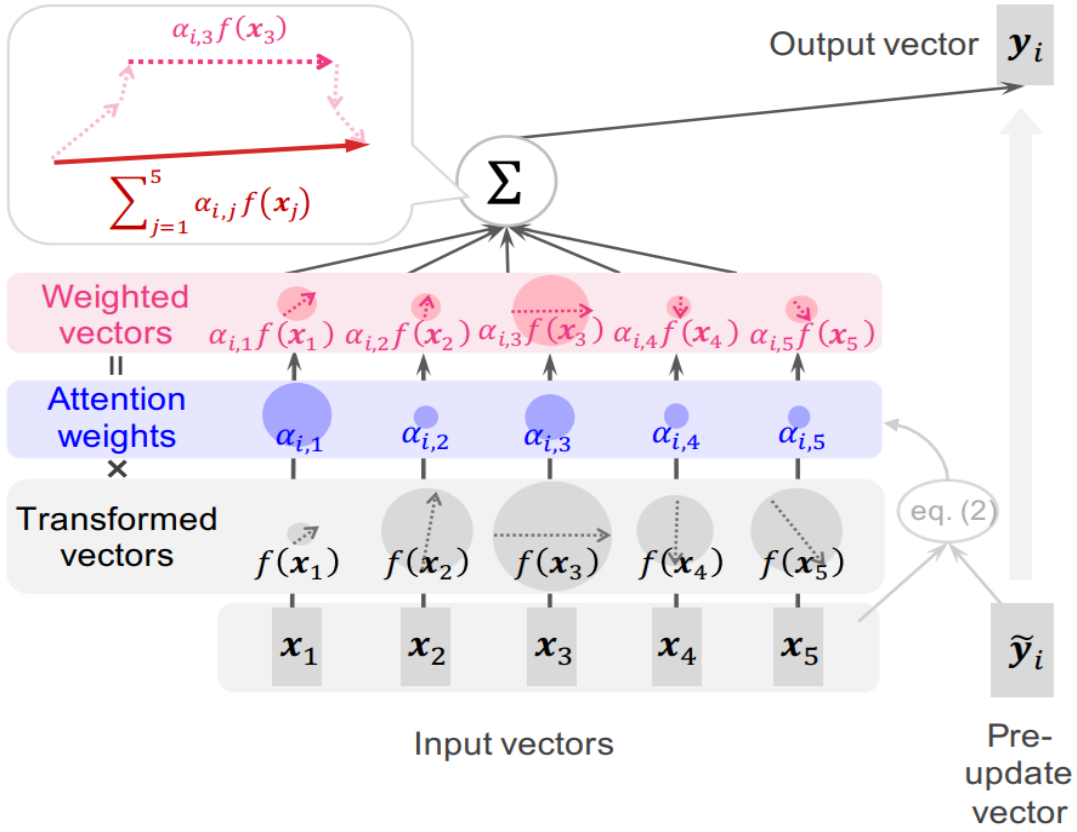


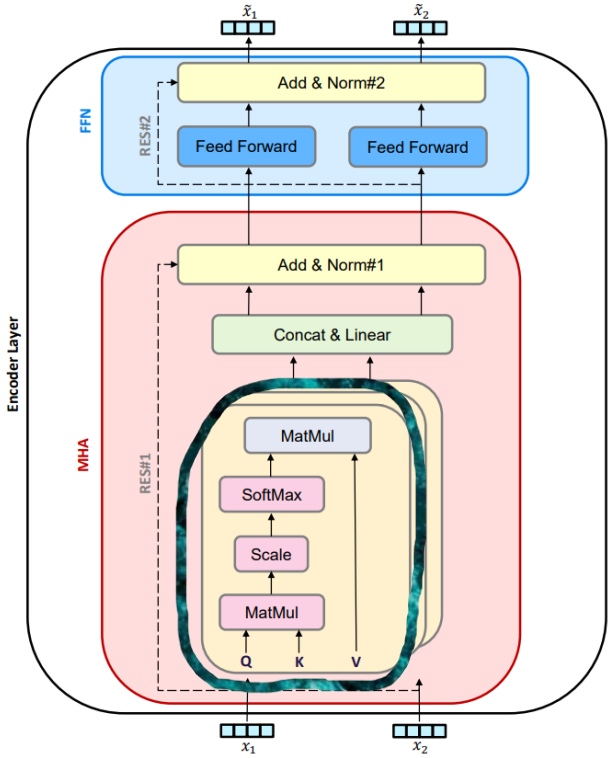




(Abnar & Zuidema, 2020)

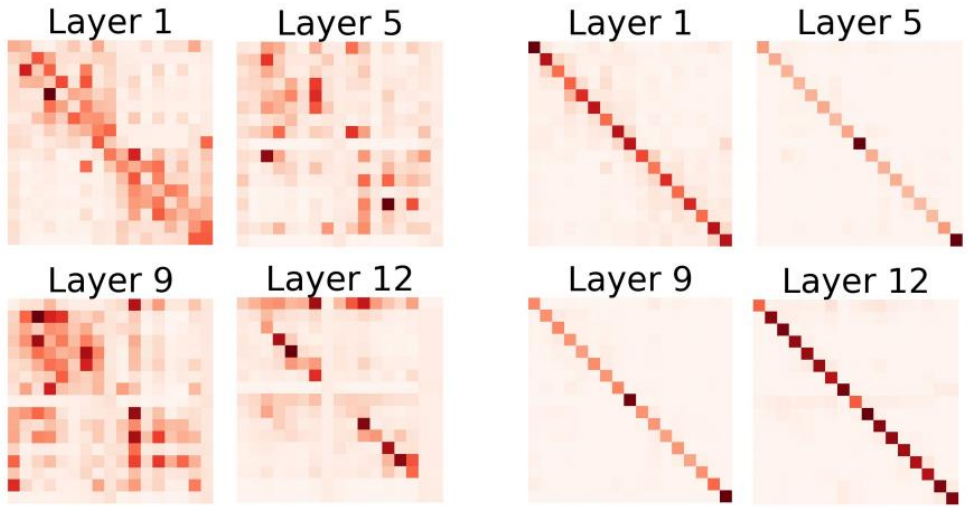
Attention-Norm





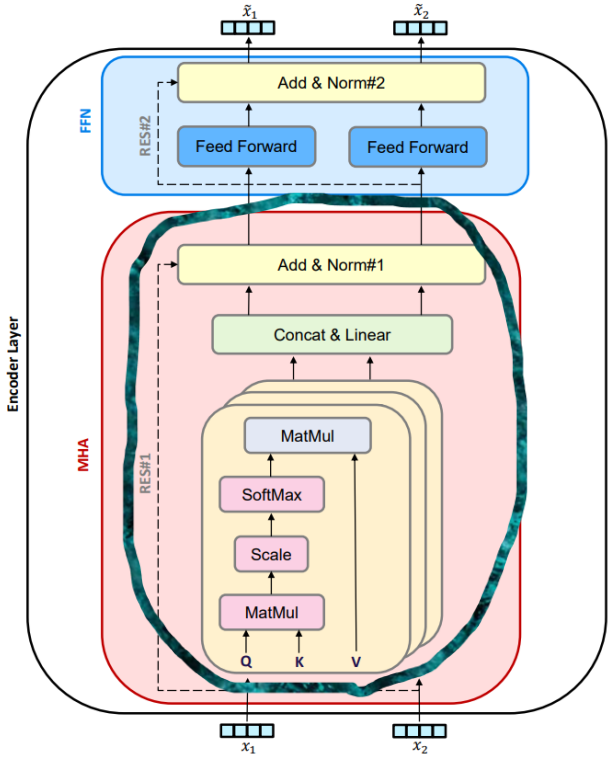
(Kobayashi et al., 2020)

Attention-Norm (extended)



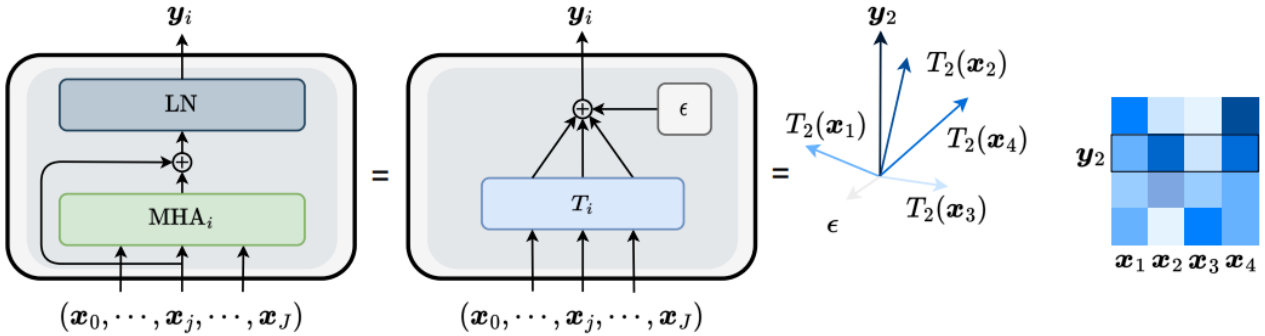
(a) Existing analysis focusing only on the multi-head attention (Kobayashi et al., 2020).

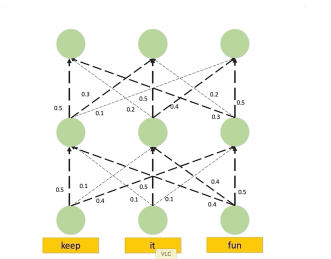
(b) Proposed method incorporating the whole attention block (i.e., multi-head attention, residual connection, and layer normalization) into the analysis.

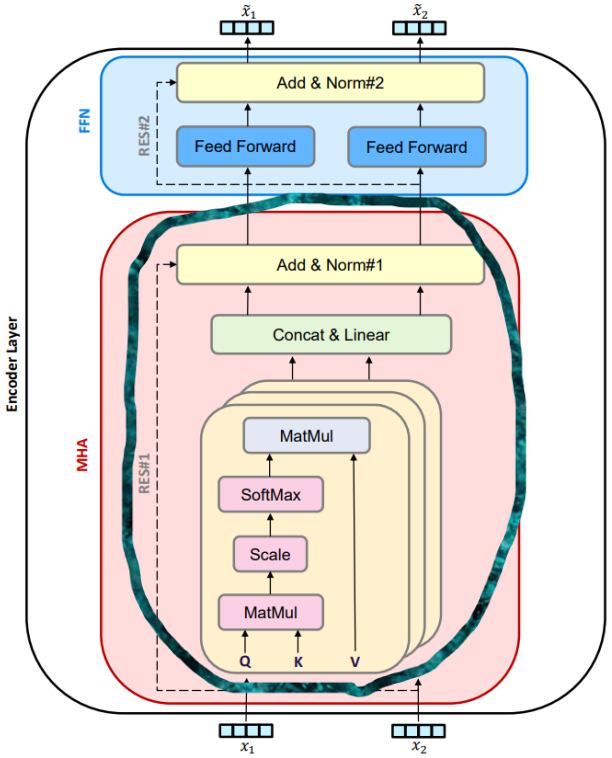


(Kobayashi et al., 2021)

ALTI

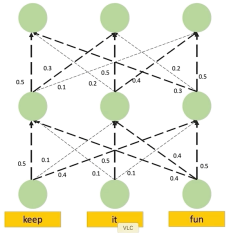






(Ferrando et al., 2022)

ALTI



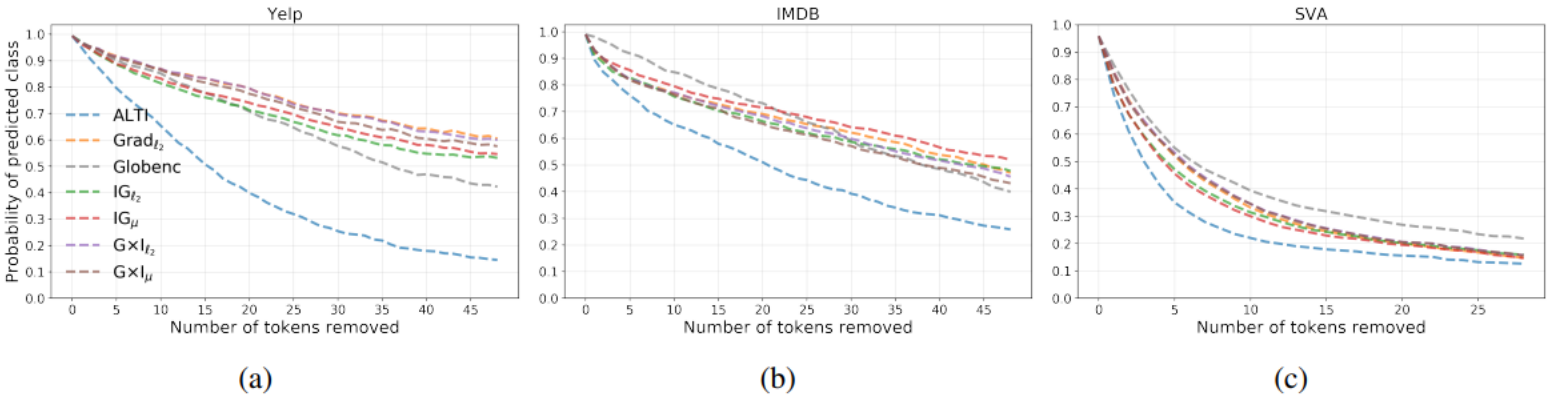
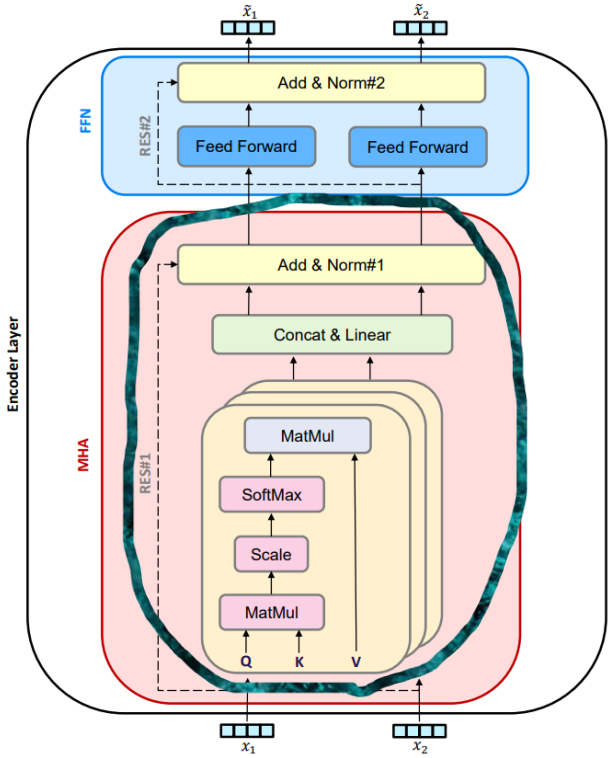


Figure 6: Probability drop in BERT predictions when removing important tokens, obtained by different interpretability methods. We show results on three datasets.



(Ferrando et al., 2022)

GlobEnc

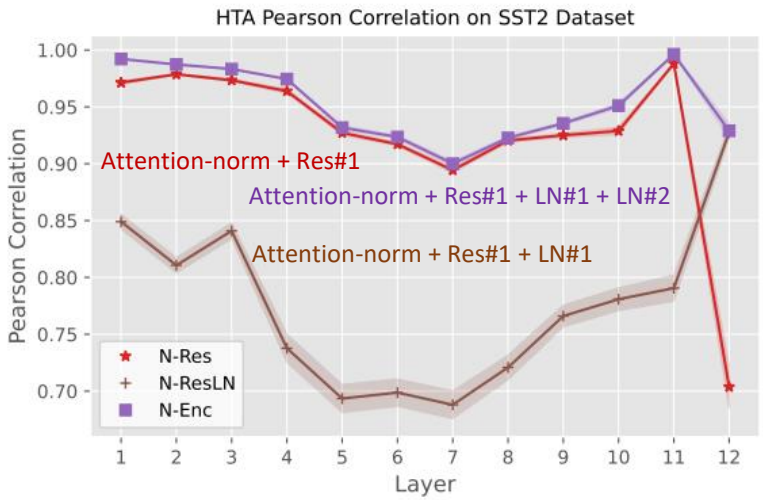


Figure 4: Single layer Pearson correlation of HTA maps with attribution maps. The 99% confidence intervals are shown as shaded areas around each line. $\mathcal{N}_{\text{RESLN}}$ shows considerably less association with HTA.

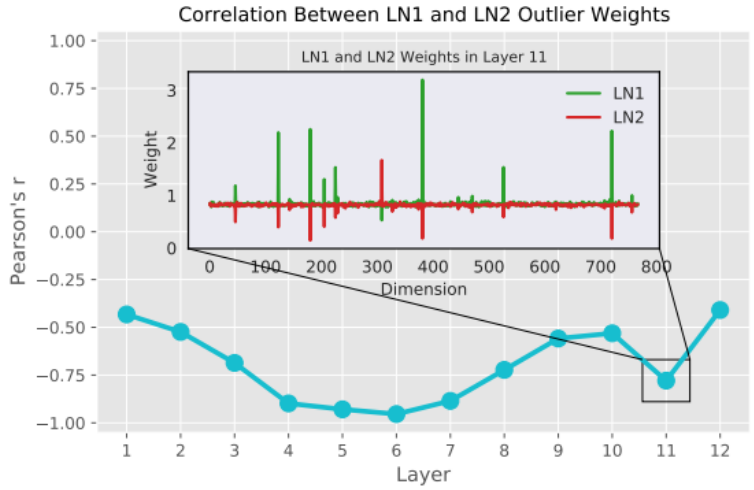
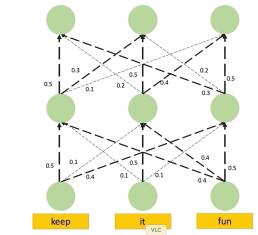
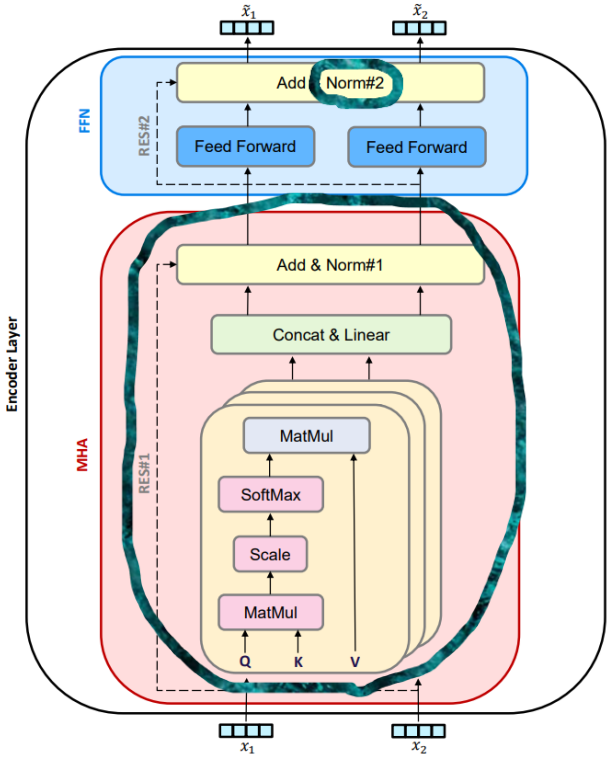


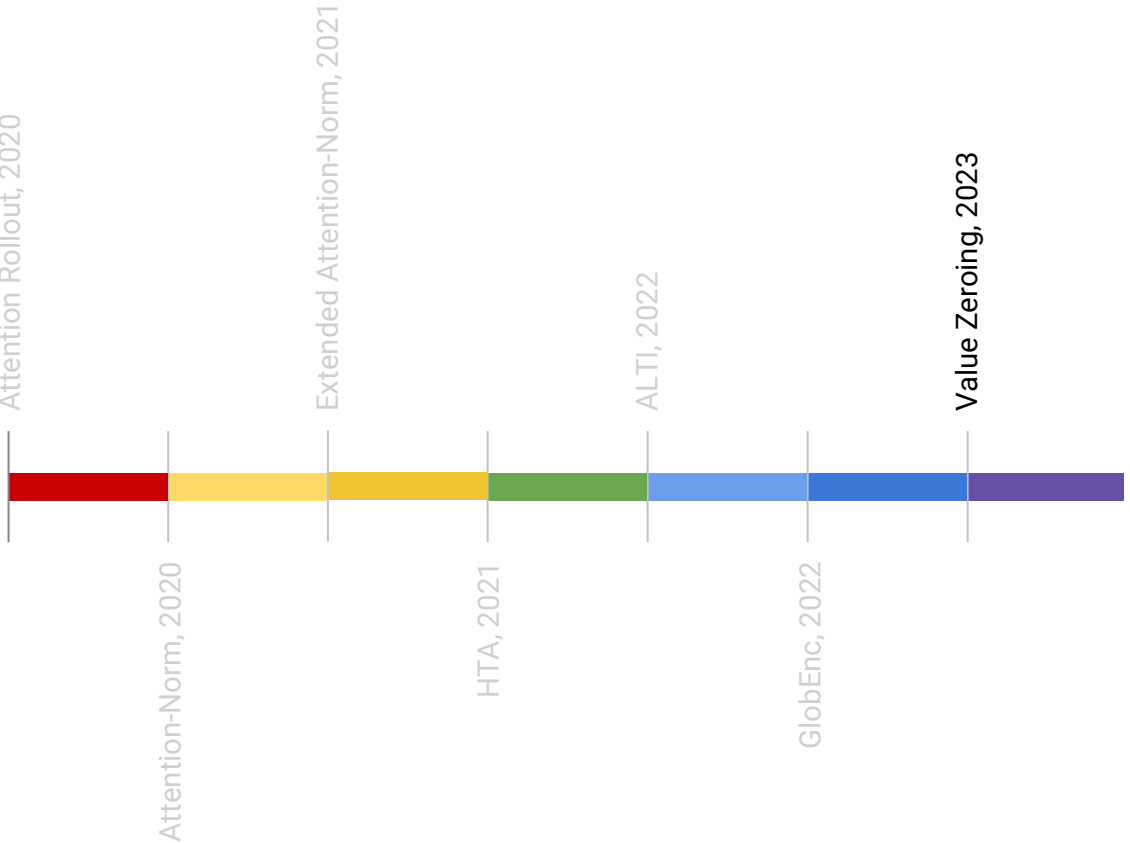
Figure 5: The Pearson correlation between outlier weights of LN#1 and LN#2 across layers. The weight values for layer 11 are shown as well.

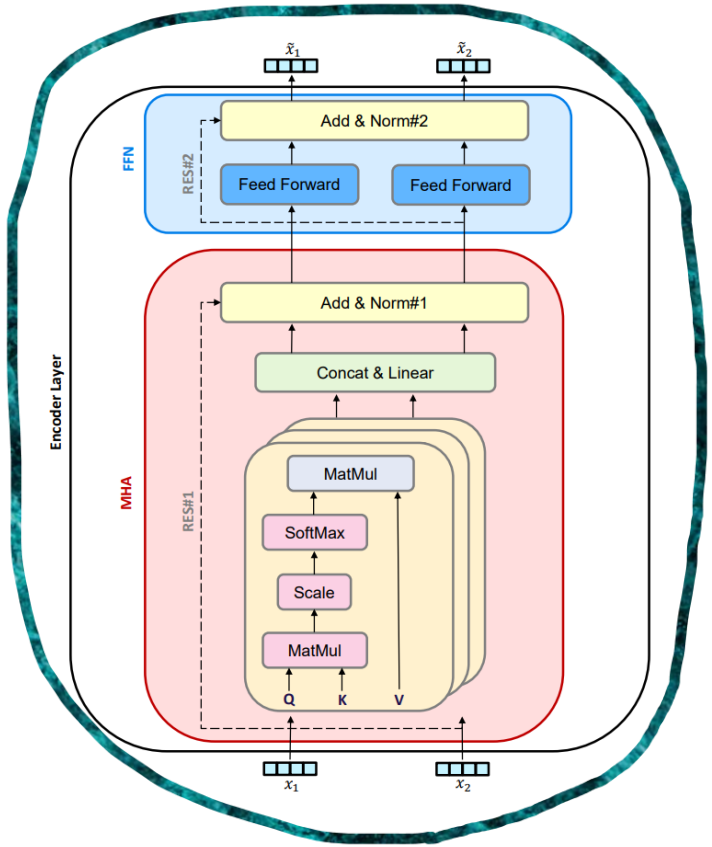




(Modarressi et al., 2022)

Value Zeroing





([Mohebbi et al., 2023])

Value Zeroing

$$(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$$

$$\left.\begin{aligned} \boldsymbol{q}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{h}} + \boldsymbol{b}_Q^h \\ \boldsymbol{k}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{h}} + \boldsymbol{b}_K^h \\ \boldsymbol{v}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{V}}^{\boldsymbol{h}} + \boldsymbol{b}_V^h \end{aligned}\right\} \alpha_{i,j} = \underset{\boldsymbol{x}_j \in \mathcal{X}}{\text{softmax}}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d}}\right) \in \mathbb{R}$$

$$\boldsymbol{z}_i^h = \sum_{j=1}^{n} \alpha_{i,j}^h \boldsymbol{v}_j^h$$

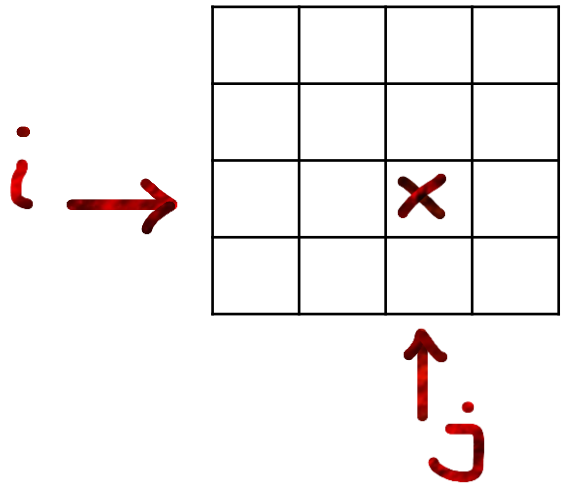
$$\boldsymbol{z}_i = \text{CONCAT}(\boldsymbol{z}_i^1, ..., \boldsymbol{z}_i^H)\boldsymbol{W}_O$$

$$\boldsymbol{z}_i = \text{LN}_{\text{MHA}}(\boldsymbol{z}_i + \boldsymbol{x}_i)$$

$$\tilde{\boldsymbol{x}}_i = \max(0, \boldsymbol{z}_i \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$\tilde{\boldsymbol{x}}_i = \text{LN}_{\text{FFN}}(\tilde{\boldsymbol{x}}_i + \boldsymbol{z}_i)$$

$$c_{i,j} = ?$$



Value Zeroing

$$(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$$

$$\left.\begin{aligned} \boldsymbol{q}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{h}} + \boldsymbol{b}_Q^h \\ \boldsymbol{k}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{h}} + \boldsymbol{b}_K^h \\ \boldsymbol{v}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{V}}^{\boldsymbol{h}} + \boldsymbol{b}_V^h \end{aligned}\right\} \alpha_{i,j} = \underset{\boldsymbol{x}_j \in \mathcal{X}}{\text{softmax}}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d}}\right) \in \mathbb{R}$$

$$\boldsymbol{z}_i^h = \sum_{j=1}^{n} \alpha_{i,j}^h \boldsymbol{v}_j^h$$

$$\boldsymbol{z}_i = \textsc{Concat}(\boldsymbol{z}_i^1, ..., \boldsymbol{z}_i^H)\boldsymbol{W}_O$$

$$\boldsymbol{z}_i = \text{LN}_{\text{MHA}}(\boldsymbol{z}_i + \boldsymbol{x}_i)$$

$$\tilde{\boldsymbol{x}}_i = \max(0, \boldsymbol{z}_i \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$\tilde{\boldsymbol{x}}_i = \text{LN}_{\text{FFN}}(\tilde{\boldsymbol{x}}_i + \boldsymbol{z}_i)$$

$$\mathcal{C}_{i,j} = ?$$

$$\boldsymbol{v}_j^h \leftarrow \mathbf{0}, \forall h \in H$$

$$\mathcal{C}_{i,j} = \tilde{\boldsymbol{x}}_i^{\neg j} * \tilde{\boldsymbol{x}}_i$$

(Mohebbi et al., 2023)

Value Zeroing

$(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$

No ablation here!

$$\left.\begin{aligned} \boldsymbol{q}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{h}} + \boldsymbol{b}_Q^h \\ \boldsymbol{k}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{h}} + \boldsymbol{b}_K^h \\ \boldsymbol{v}_i^h &= \boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{V}}^{\boldsymbol{h}} + \boldsymbol{b}_V^h \end{aligned}\right\} \alpha_{i,j} = \underset{\boldsymbol{x}_j \in \mathcal{X}}{\text{softmax}}\left(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^\top}{\sqrt{d}}\right) \in \mathbb{R}$$

$$\boldsymbol{z}_i^h = \sum_{j=1}^{n} \alpha_{i,j}^h \boldsymbol{v}_j^h$$

$$\boldsymbol{z}_i = \textsc{Concat}(\boldsymbol{z}_i^1, ..., \boldsymbol{z}_i^H)\boldsymbol{W}_O$$

$$\boldsymbol{z}_i = \text{LN}_{\text{MHA}}(\boldsymbol{z}_i + \boldsymbol{x}_i)$$

$$\tilde{\boldsymbol{x}}_i = \max(0, \boldsymbol{z}_i \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

$$\tilde{\boldsymbol{x}}_i = \text{LN}_{\text{FFN}}(\tilde{\boldsymbol{x}}_i + \boldsymbol{z}_i)$$

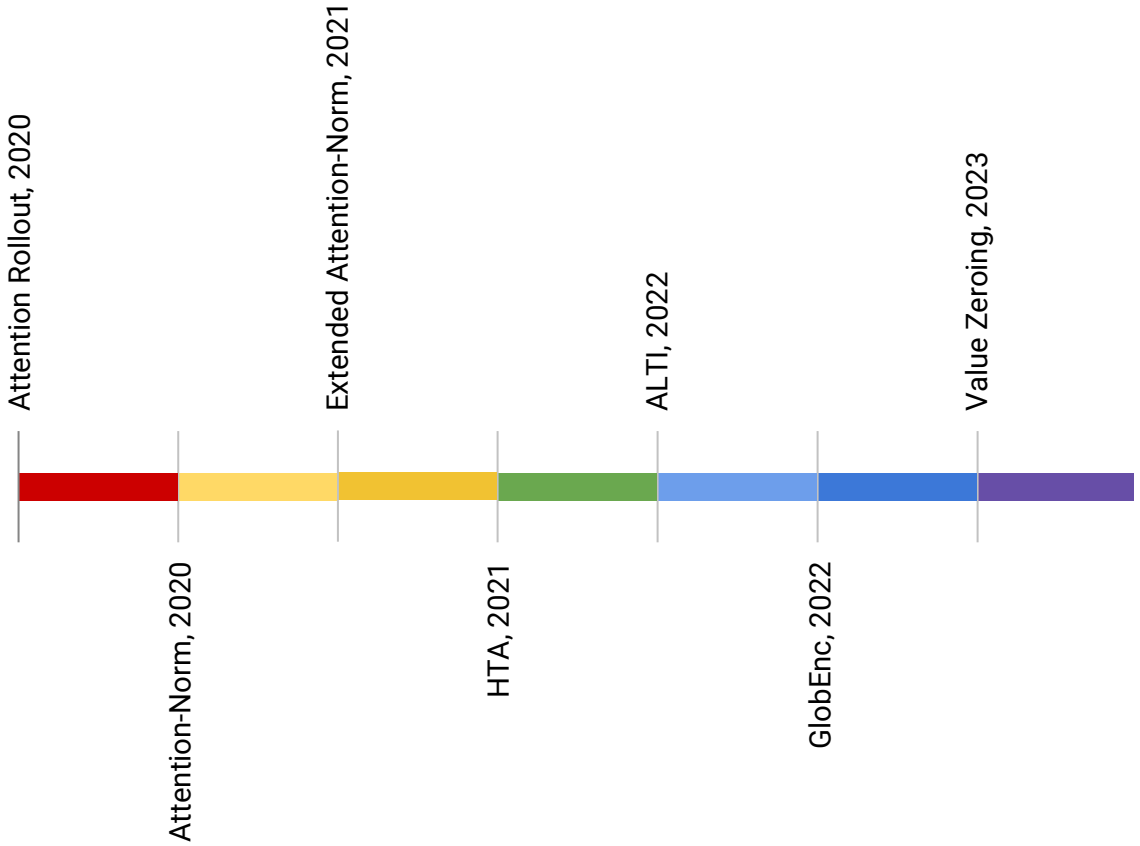
$$\mathcal{C}_{i,j} = ?$$

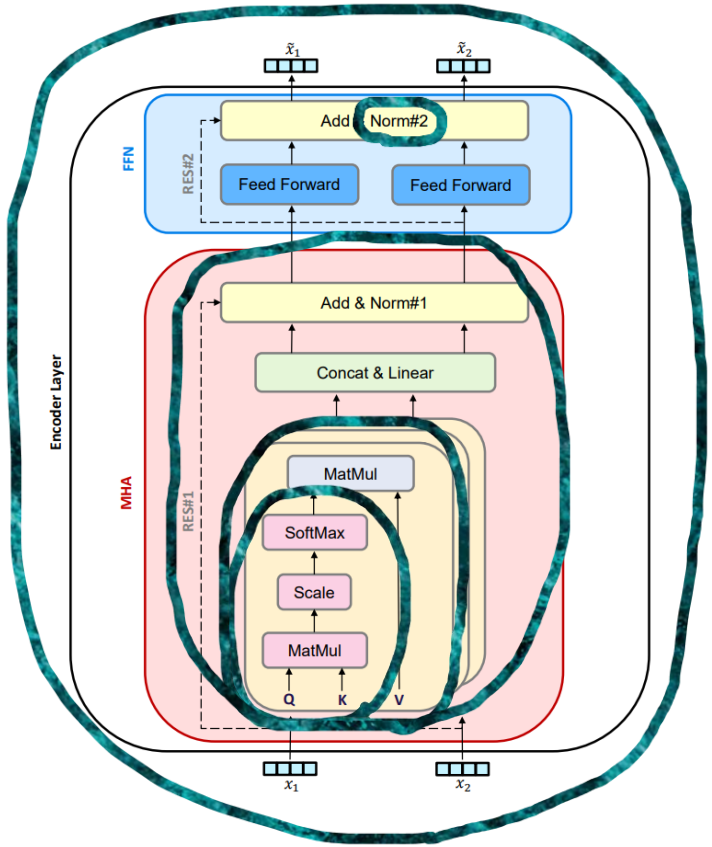
$$\boldsymbol{v}_j^h \leftarrow \mathbf{0}, \forall h \in H$$

$$\mathcal{C}_{i,j} = \tilde{\boldsymbol{x}}_i^{\neg j} * \tilde{\boldsymbol{x}}_i$$

(Mohebbi et al., 2023)

Let’s evaluate & compare





Evaluation in Text

Controlled task: grammatical agreements

The books in the library [MASK] read by many.

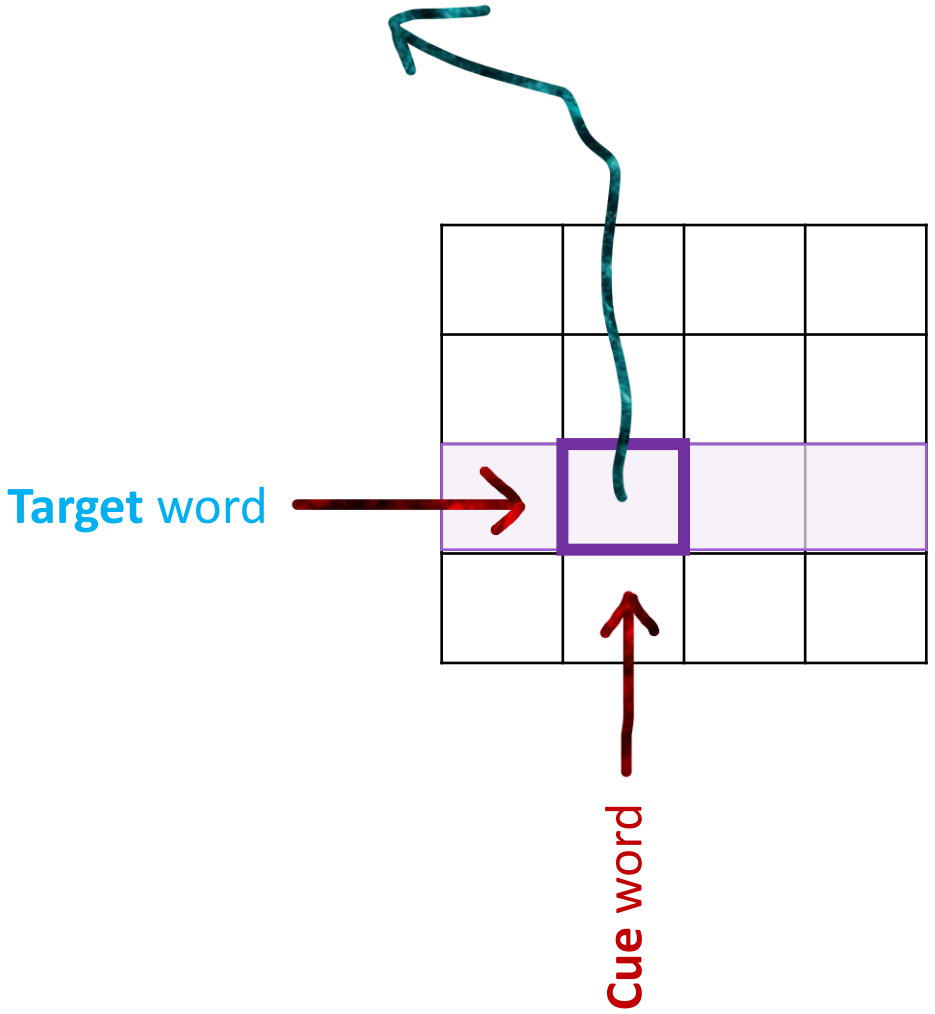
Evaluation in Text
Controlled task: grammatical agreements

Phenomenon	UID	Example	Target word	Foil word
Anaphor Number Agreement	ana	Many teenagers were helping [MASK].	themselves	herself
Determiner-Noun Agreement	dna	Jeffrey has not passed [MASK] museums.	these	this
	dnaa	Sara noticed [MASK] white hospitals.	these	this
Subject-Verb Agreement	darn	The pictures of Martha [MASK] not disgust Anne.	do	does
	rpsv	Kristen [MASK] fixed this chair.	has	have

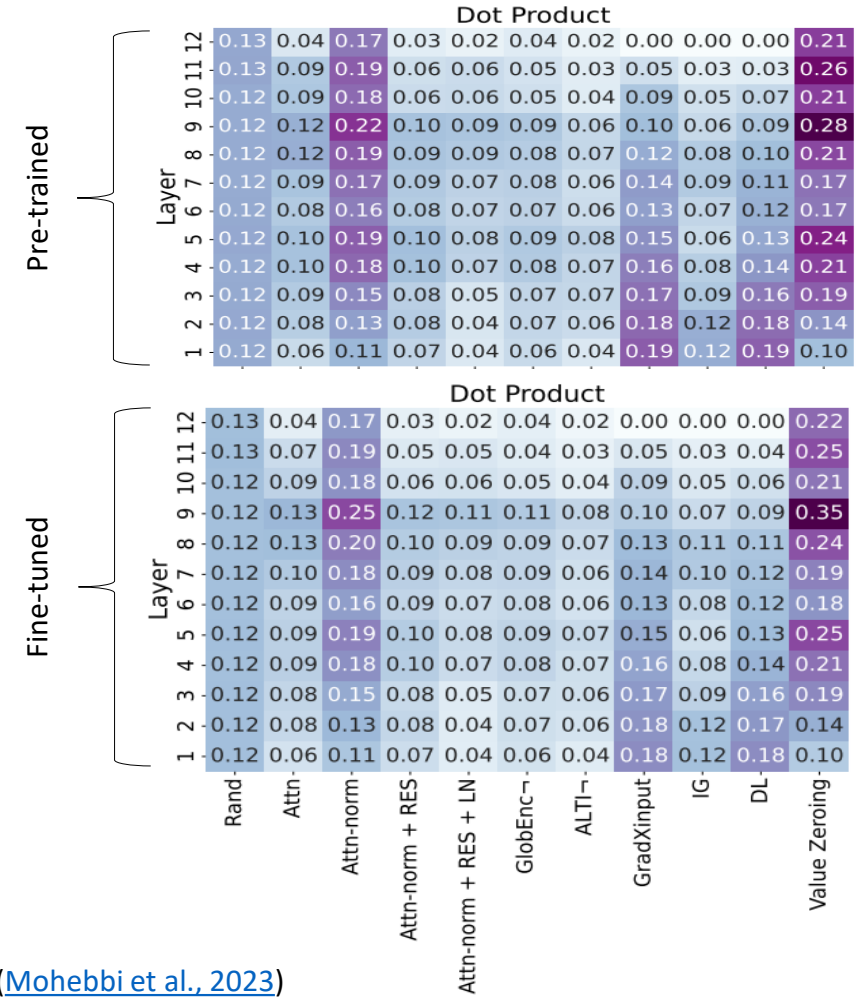
Table 1: Examples of the selected tasks with our annotations from the BLiMP benchmark (UIDs are unique identifiers used in BLiMP). *Cue* words are underlined.

(**Accuracy** for BERT-base-uncased is **0.96** in the pre-trained and **0.99** in the fine-tuned setup)

‘Cue Contribution’ score



Evaluation in Text



([Mohebbi et al., 2023](#))

Evaluation in Text

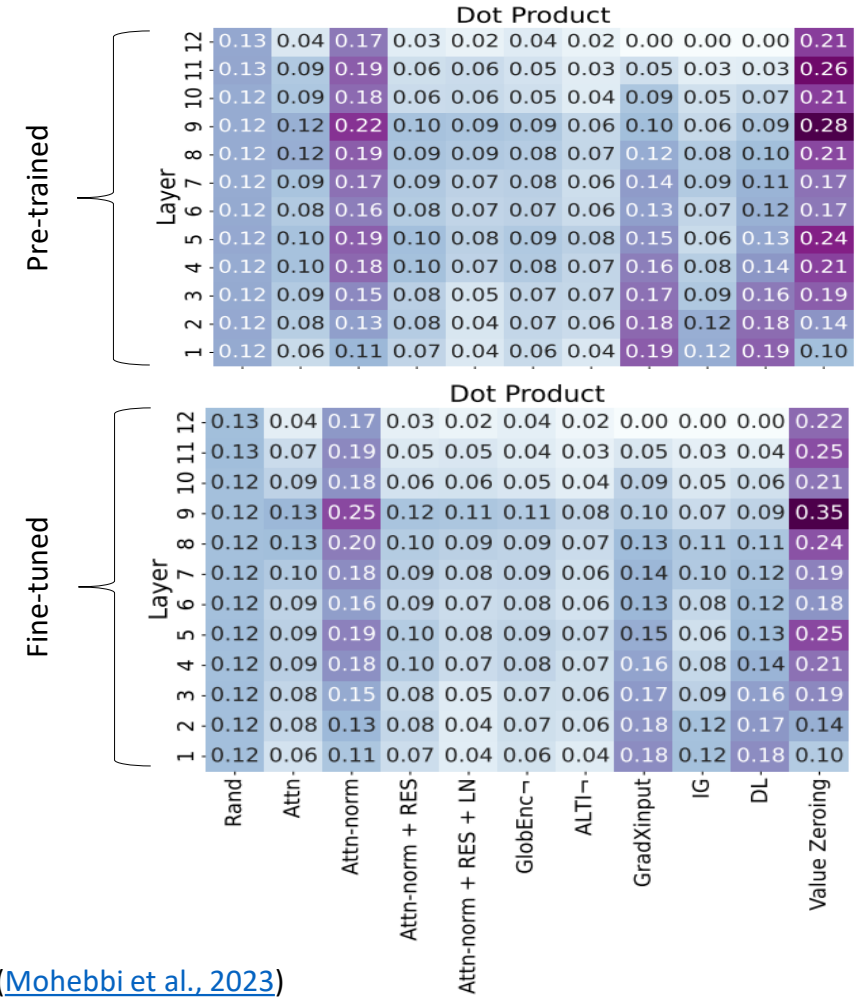
(Mohebbi et al., 2023)

Singular v.s. **Plural**

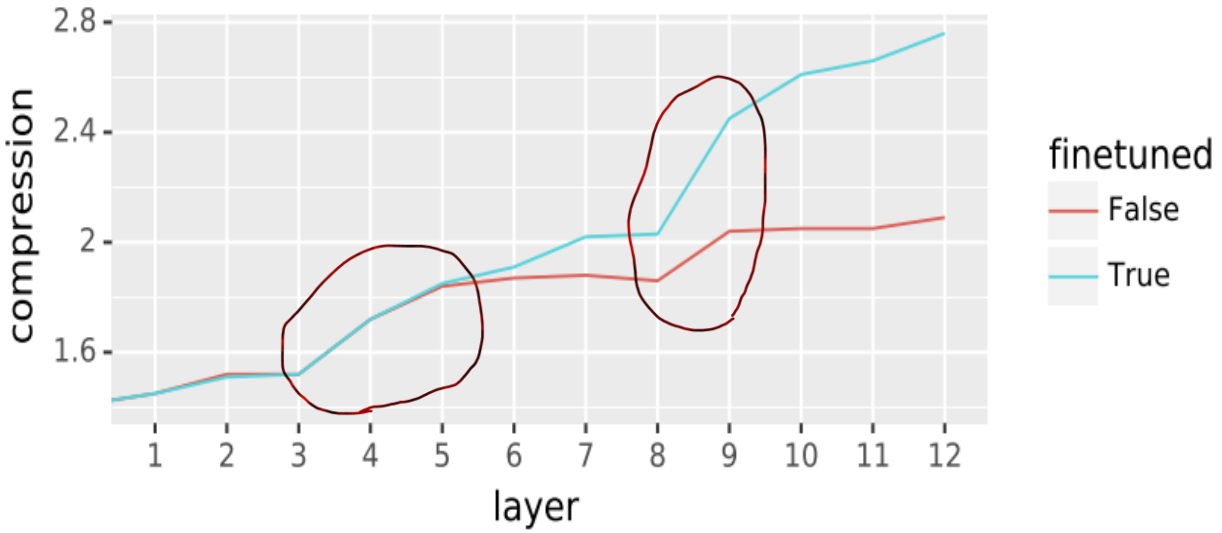
Probing classifier

The books in the library [MASK] read by many.

Cue Contribution v.s. Number encoding probing



(Mohebbi et al., 2023)



Aggregated scores for a SVA example

The pictures of some hat [MASK] scaring Marcus.

Aggregated scores for a SVA example

The pictures of some hat [MASK] scaring Marcus.

Attn:

Attn-norm:

Attn-norm+RES:

Attn-norm+RES+LN:

GlobEnc:

ALTI:

GradXinput:

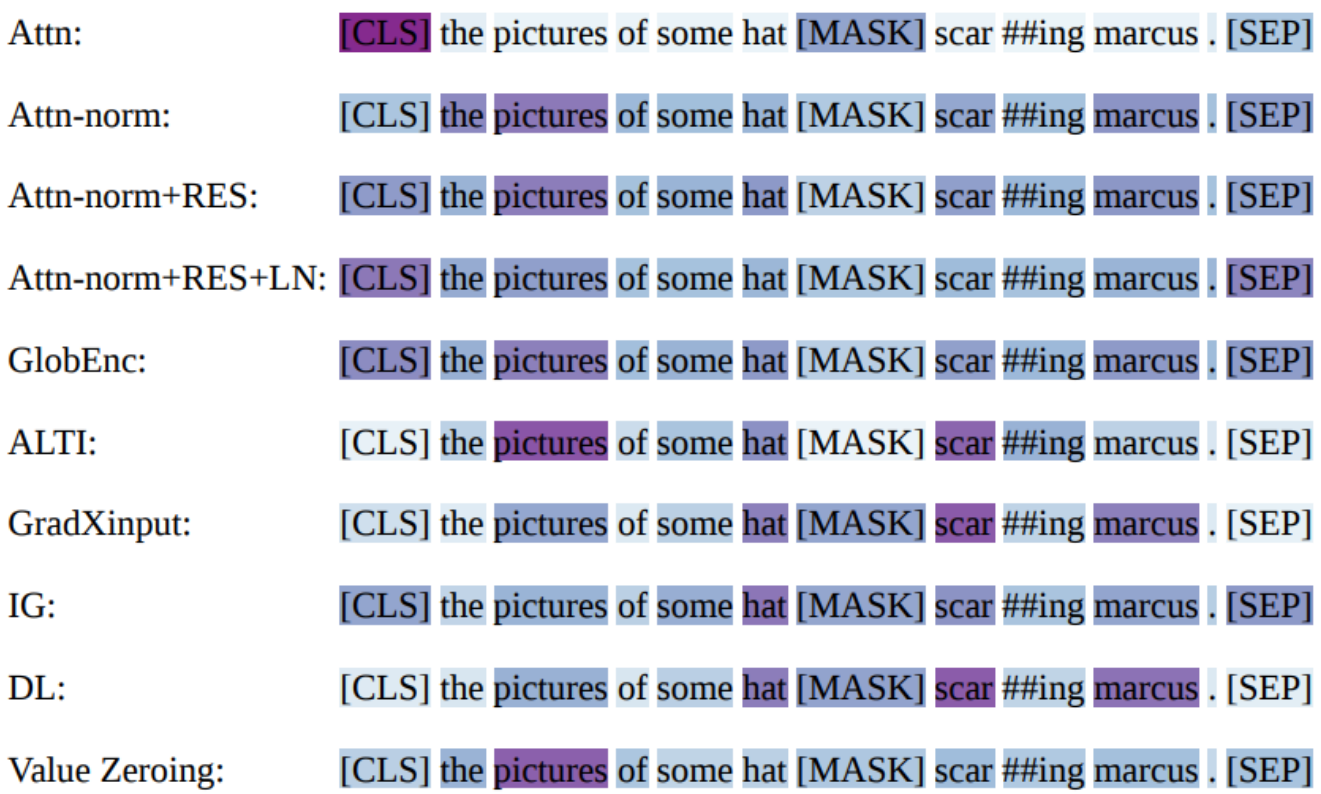
IG:

DL:

Value Zeroing:

Aggregated scores for a SVA example

The pictures of some hat [MASK] scaring Marcus.



Evaluation in Speech

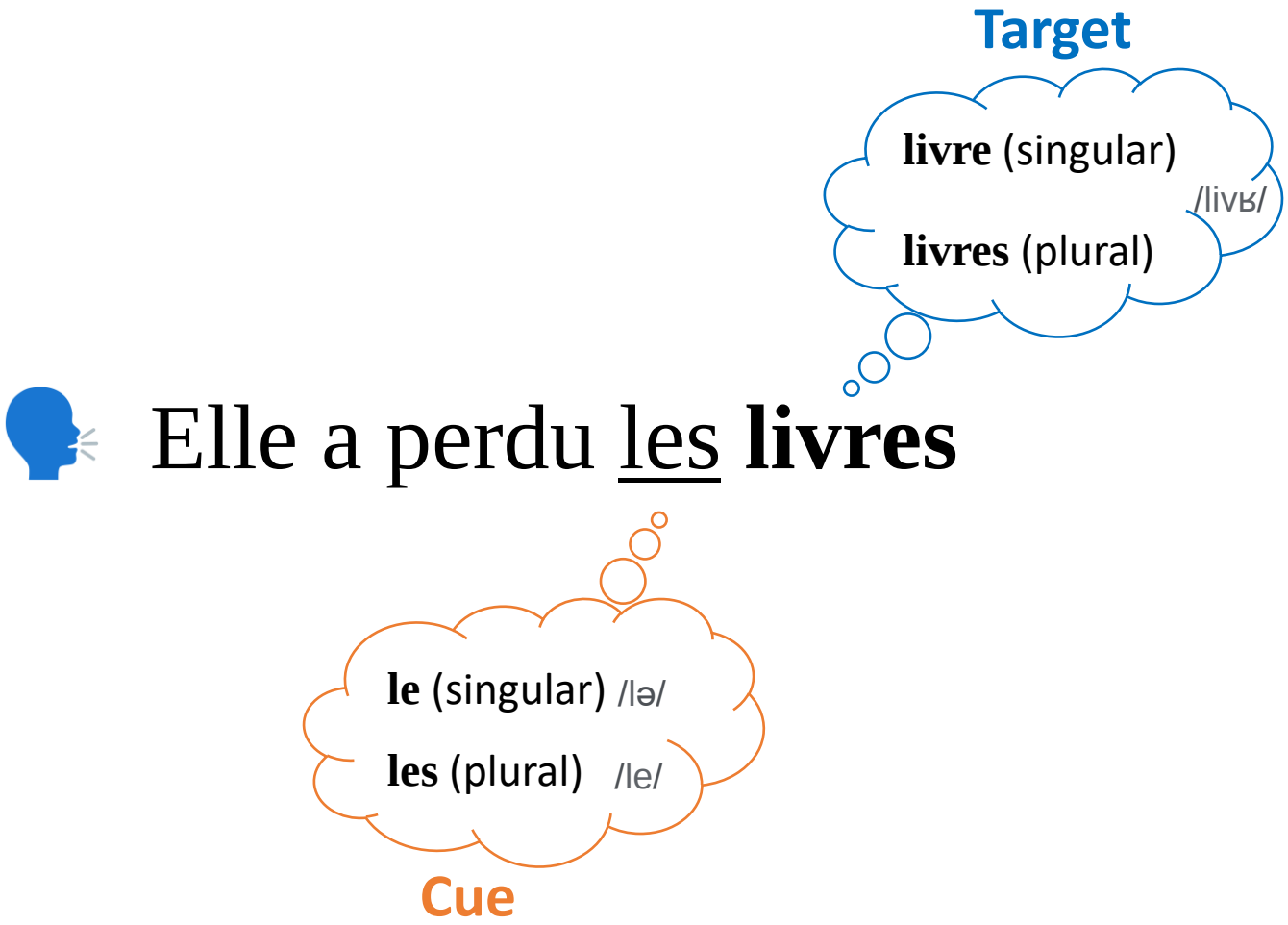
Controlled task: homophony in French

Elle a perdu les **livres**

(She lost the books)

Evaluation in Speech

Controlled task: homophony in French



Defined Templates

Pattern	Examples of transcription	#
Det_Noun	C'est le septième **titre** de champion de Syrie de l'histoire du club Il y mène une **vie** d'études et de recherches	720
Pronoun_Verb	Chaque jour, leurs concurrents les voient sortir de pistes dont ils **ignorent** l'existence On y **trouve** une plage naturiste	257
Det_Noun_Verb	Peu après cette élimination, le **club** et Alexander se **séparent** à l'amiable À la fin, les **enfants** se **révoltent** et détruisent l'école.	23

Table 1: Examples of the extracted audios from the Common Voice corpus based on defined patterns. Last column shows the number of examples obtained. Cue and Target words are underlined and **bolded**, respectively.

Evaluation in Speech

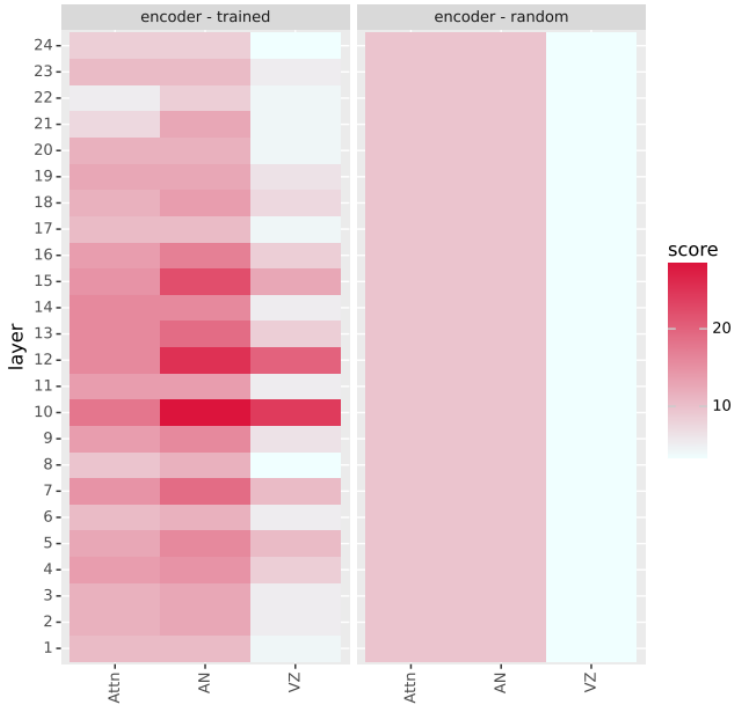


Figure 1: Layer-wise cue contribution according to different analysis methods averaged over all examples for XLSR-53, trained (left) vs. randomly initialized (right).

Evaluation in Speech

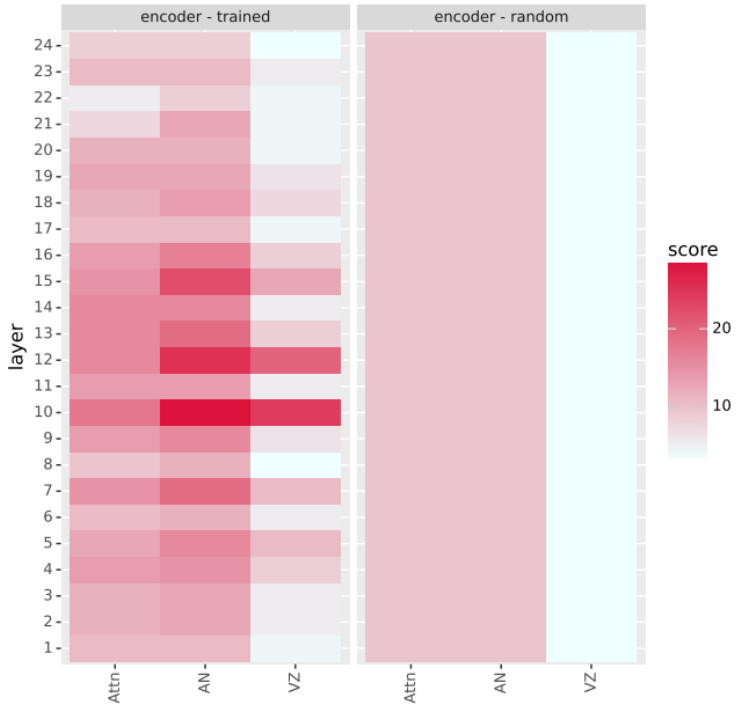


Figure 1: Layer-wise cue contribution according to different analysis methods averaged over all examples for XLSR-53, trained (left) vs. randomly initialized (right).

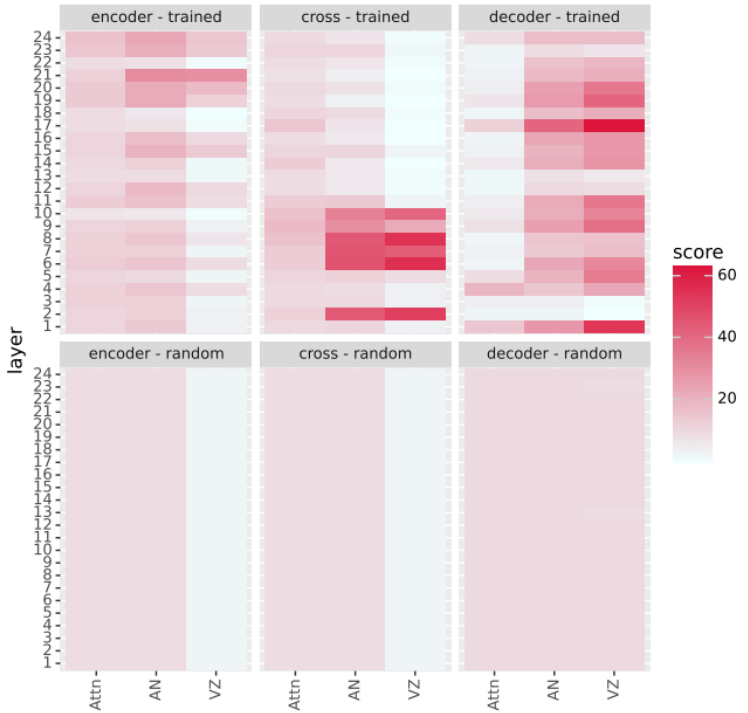


Figure 2: Layer-wise cue contribution according to different analysis methods averaged over all examples for Whisper-medium, trained (top) vs. randomly initialized (bottom).

Cue Contribution v.s. Number encoding probing

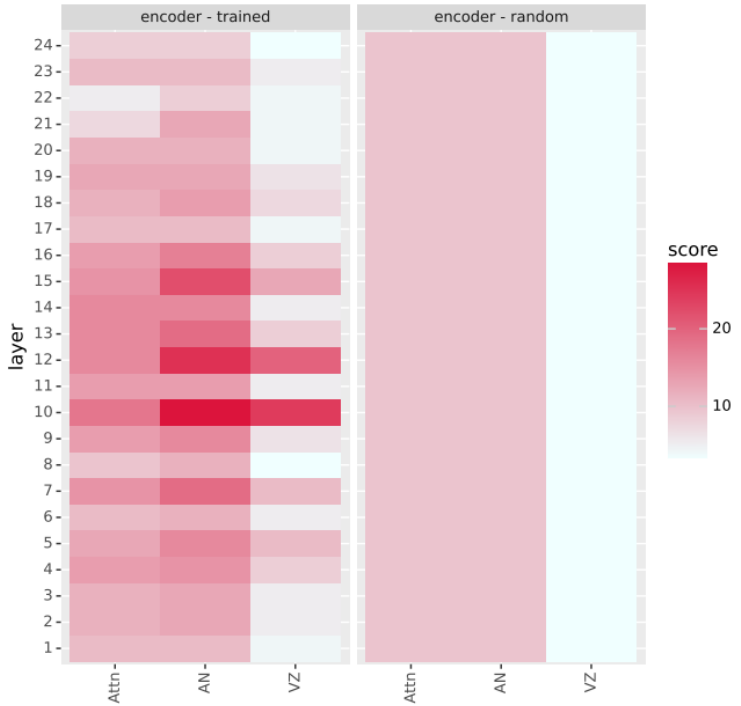


Figure 1: Layer-wise cue contribution according to different analysis methods averaged over all examples for XLSR-53, trained (left) vs. randomly initialized (right).

(Mohebbi et al., 2023)

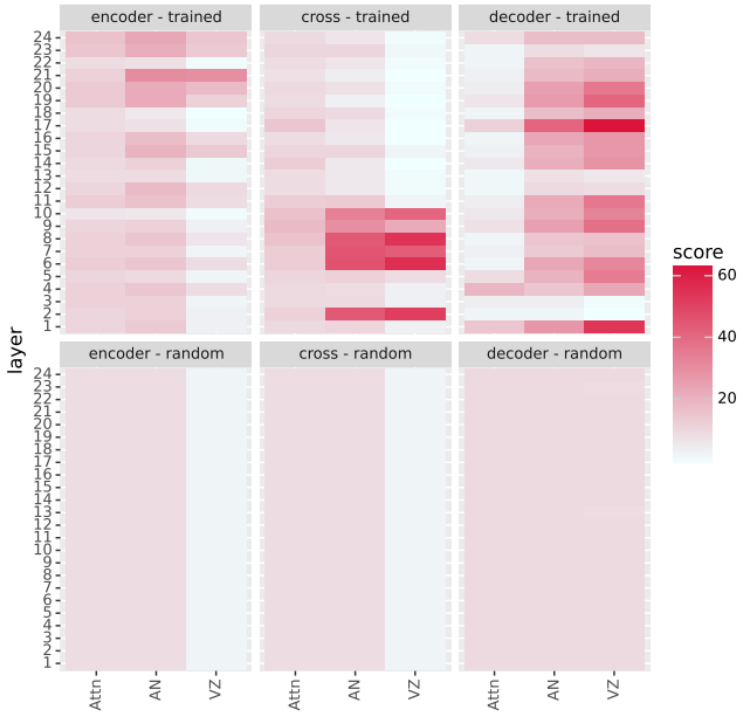


Figure 2: Layer-wise cue contribution according to different analysis methods averaged over all examples for Whisper-medium, trained (top) vs. randomly initialized (bottom).

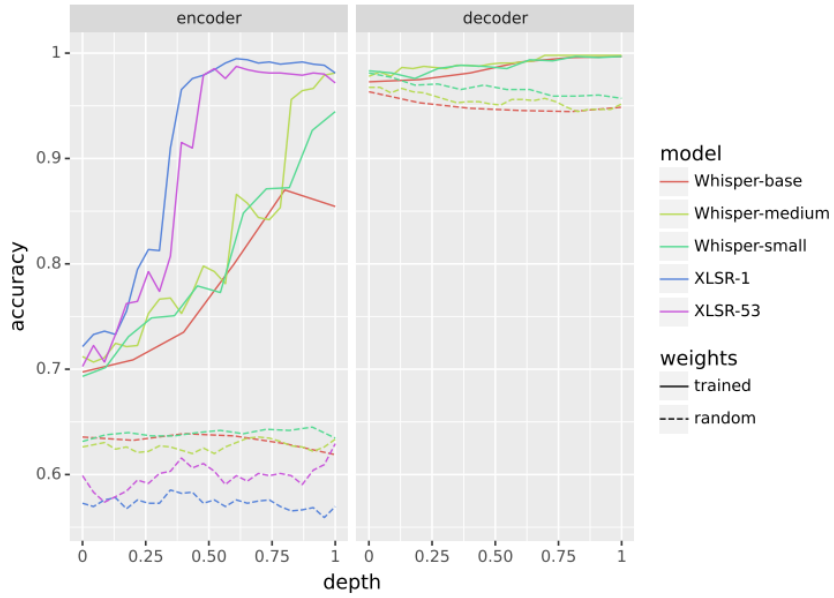


Figure 4: Accuracy of probing classifiers trained on frozen target representations obtained from various ASR models. The depth of Whisper-base (6) and Whisper-small (12), has been normalized to 1 to facilitate comparisons.

Logit-based methods

- LRP-based Attention
- ALTI-Logit
- DecompX

(Chefer et al., 2021; Ferrando et al., 2023; Modarressi et al., 2023)

Final thoughts

- Transformers are shared among different modalities; time to converge in our analysis methods to have methods that are:
  - tailored to the model architecture (Transformer)
  - irrespective to input data type (text, audio, music, image, etc.)
  - irrespective to training objective (language modeling, contrastive learning, denoising, etc.)
- We are interested in quantifying context mixing, something that Transformers are made for!
- Attention is Not enough for the purpose of context mixing measurement!

Value Zoroing in use


```
import inseq

model = inseq.load_model("Helsinki-NLP/opus-mt-en-fr", "value_zeroing")
out = model.attribute("A generative language models interpretability tool.")
out.show()
```

Source Saliency Heatmap

x: Generated tokens, y: Attributed tokens

	_Un	_outil	_d	'	interprétation	_de	_modèles	_de	_langage	_génér	atifs	.	</s>
_A	0.784	0.001	0.0	0.0	0.0	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0
_	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.0	0.0	0.0
gener	0.003	0.083	0.349	0.002	0.002	0.237	0.056	0.322	0.01	0.525	0.008	0.0	0.0
ative	0.0	0.0	0.032	0.0	0.0	0.016	0.0	0.036	0.0	0.232	0.945	0.0	0.0
_language	0.003	0.026	0.031	0.044	0.006	0.013	0.026	0.61	0.988	0.001	0.0	0.0	0.0
_models	0.033	0.029	0.027	0.0	0.012	0.72	0.916	0.012	0.0	0.001	0.0	0.001	0.0
_interpret	0.0	0.001	0.464	0.399	0.667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ability	0.003	0.001	0.091	0.331	0.31	0.006	0.001	0.0	0.0	0.0	0.0	0.007	0.0
_tool	0.119	0.859	0.001	0.002	0.001	0.001	0.0	0.0	0.0	0.0	0.0	0.002	0.0
.	0.048	0.0	0.002	0.02	0.0	0.003	0.0	0.018	0.0	0.239	0.0	0.979	0.588
</s>	0.006	0.001	0.002	0.201	0.001	0.002	0.001	0.001	0.002	0.001	0.046	0.01	0.412

(Inseq)

Thank you! 🙂

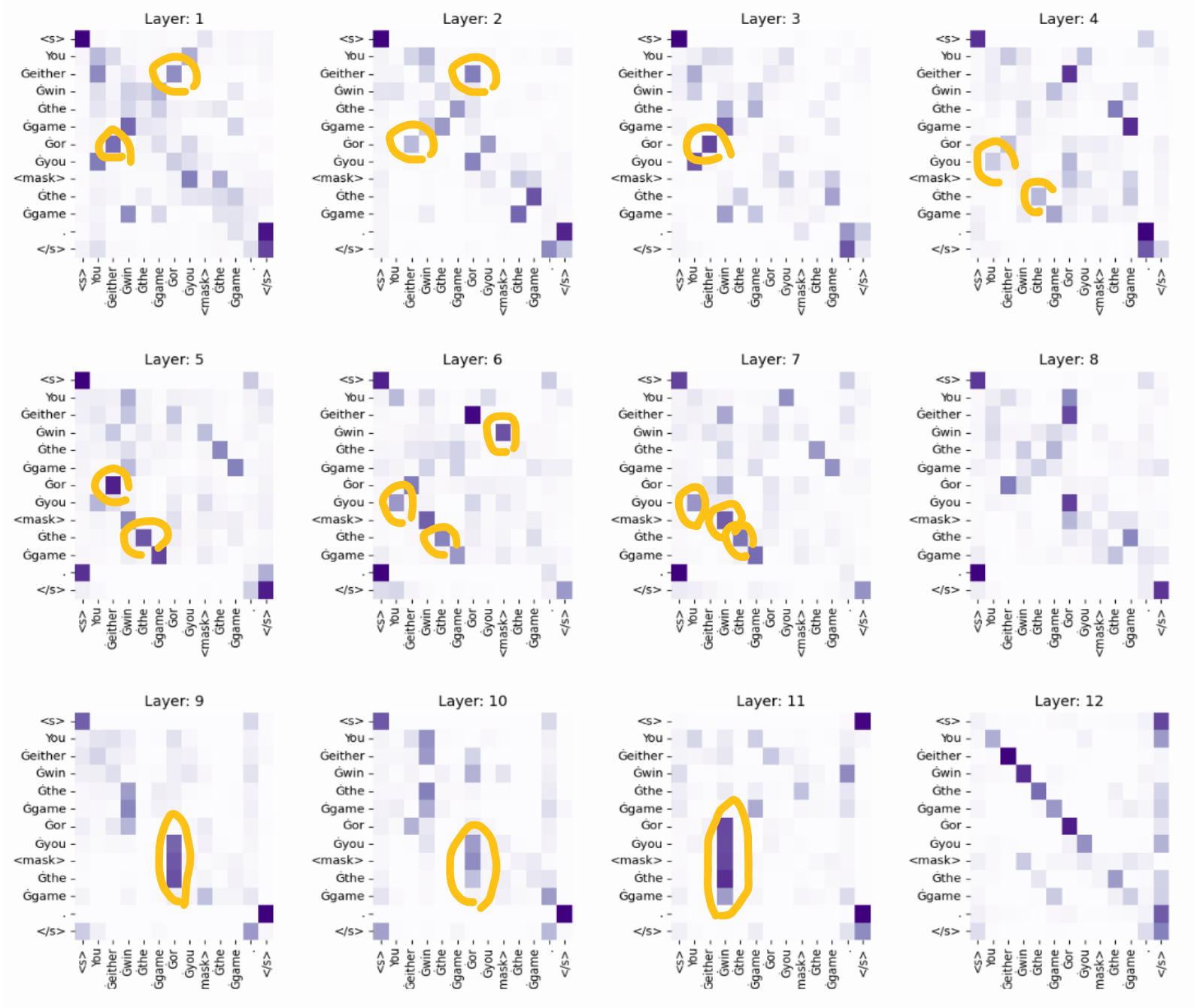
https://projects.illc.uva.nl/indeep/tutorial/

Let’s explore!

Either you win the game or you <mask> the game.

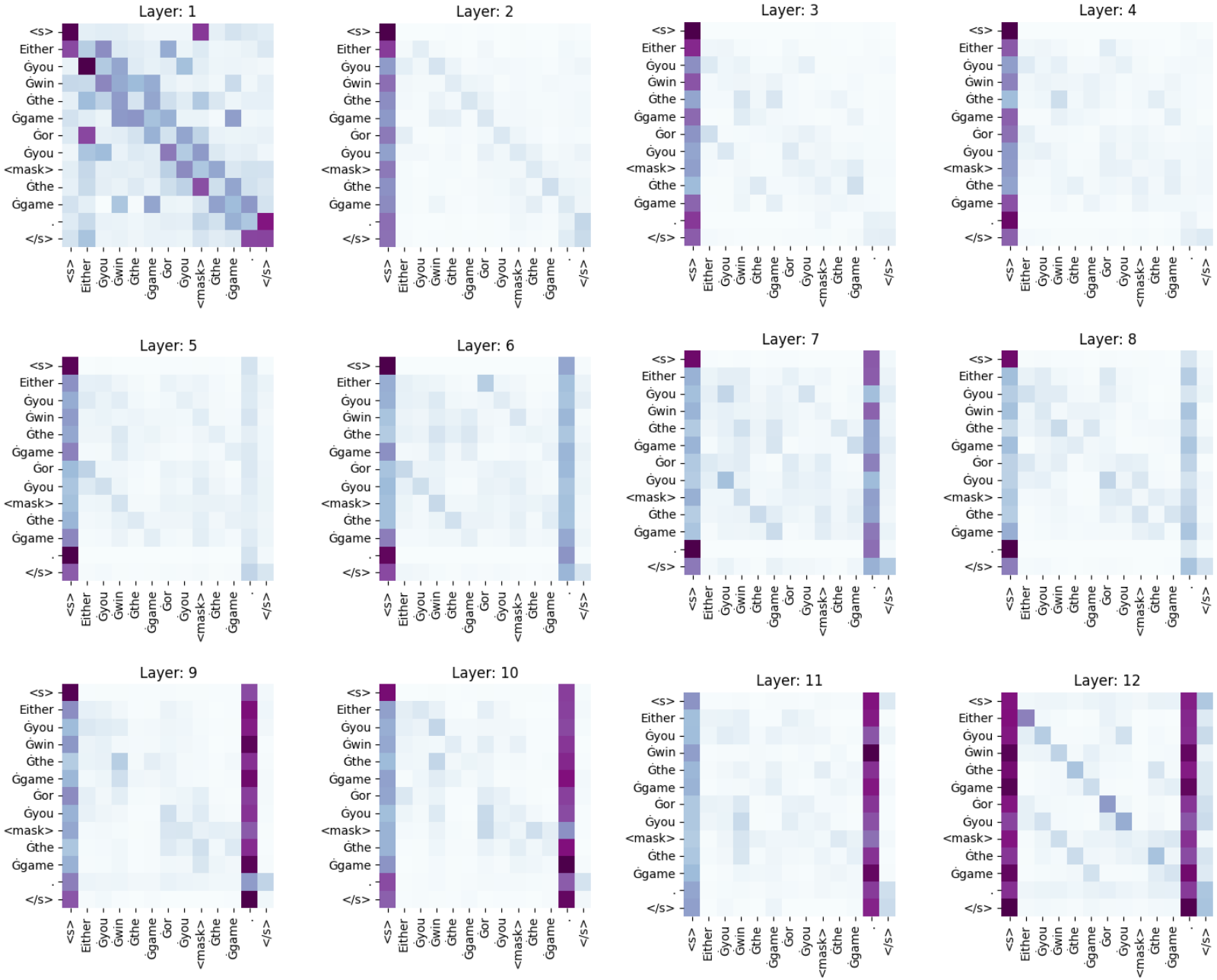
Value Zeroing

Either you win the game or you <mask> the game.



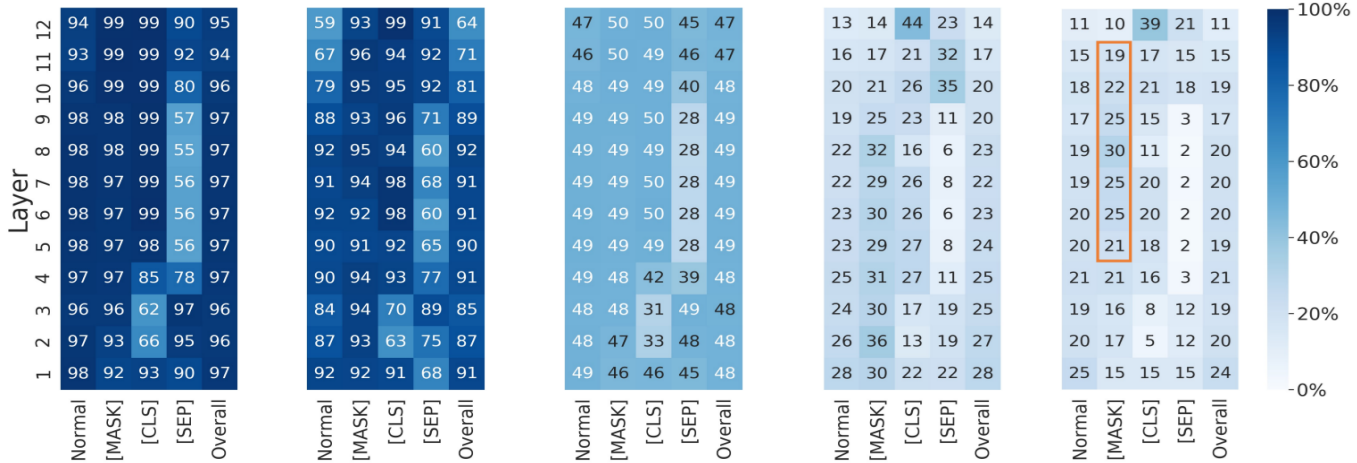
Attention

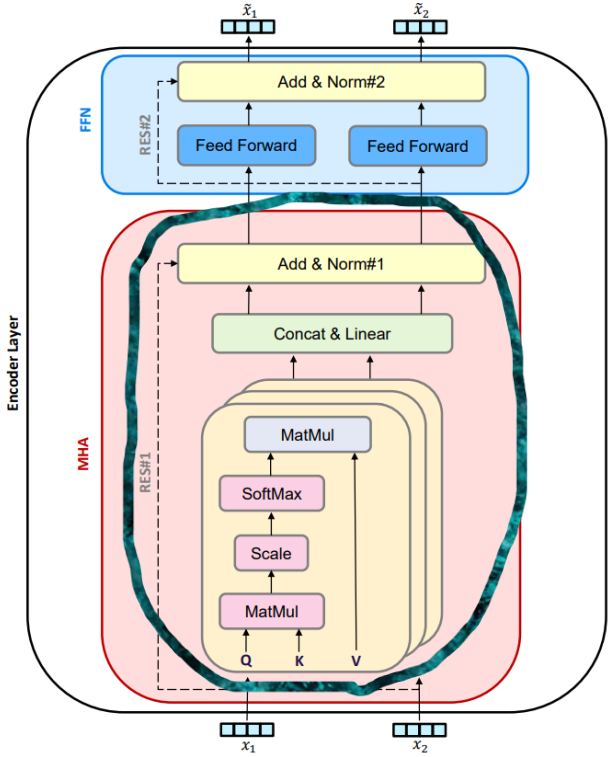
Either you win the game or you <mask> the game.



Attention-Norm (extended)

$$r_i = \frac{\|\widetilde{\boldsymbol{x}}_{i\leftarrow \text{context}}\|}{\|\widetilde{\boldsymbol{x}}_{i\leftarrow \text{context}}\| + \|\widetilde{\boldsymbol{x}}_{i\leftarrow i}\|}.$$





(Kobayashi et al., 2021)

HTA

$$c_{i,j}^{l} = \frac{||\nabla_{i,j}^{l}||_2}{\sum_{k=0}^{d_s}||\nabla_{k,j}^{l}||_2} \qquad \textbf{with} \quad \nabla_{i,j}^{l} = \frac{\delta \boldsymbol{e}_j^{l}}{\delta \boldsymbol{x}_i}$$