

University of Stuttgart Institute of Industrial Automation and Software Engineering





Interpret the Black Box of Large Language Models (LLMs)



Black box

- What is happening during the text generation? (*Transparency*)
- Is its output what we really want?
 (*fairness* and *beneficial outcomes*)
- Model Understanding and Improvement
 - Understand the inner mechanisms of Model
 - Improvement of model development

Agenda

- Motivation & Basis
- Conceptual Design
- Implementation
- Evaluation
- Summary and Outlook

Large Language Model Transformer Architecture



- The Transformer Model architecture was introduced in the paper "Attention is all you need," which consists of "Encoder" and "Decoder."
- Encoder-only LLM example: BERT (embedding)
- Decoder-only LLM example: GPT4, LlaMA-2. (text generation)

Post-Normalization and Pre-Normalization Architecture



Post-Normalization and Pre-Normalization Architecture



Post-Normalization and Pre-Normalization Architecture



Post-Normalization and Pre-Normalization Architecture



Post-Normalization and Pre-Normalization Architecture



*RMS (Root Mean Square) Normalization is a simplification of the original LayerNorm.

Understand Large Language Model Thoughts



- Interpreting outputs of each hidden layer's attention mechanism output and block output.
- Attention mechanism output what texts does the LLM focus on?
- Block output what results does the LLM generate?

Probing Techniques – Logit Lens

- **Flow:** Transforms input to embeddings, processes through layers to output logits, and probes its intermediate results.
- Probing Intermediate States: Directly using the model's unembedding layer to decode each hidden layer's attention mechanism output and block output.

To understand how information evolves and contributes to the final output. [4]



Probing Techniques – Tuned Lens



- Introduced in "Eliciting Latent Predictions from Transformers with the Tuned Lens"
- Similar to the **Logit Lens** implementation.
- Tuned Lens
 The Tuned Lens refines the Logit Lens by introducing "translators," which use machine learning to train.
 - These translators adjust the hidden states layer by layer, ensuring they more closely match those expected at the final output layer.

	Logit Lens (Left)	Tuned Lens (Right)
Lens Architecture	Projection	Transformation
Interpretability	Direct	Refined Token
Methodology	Mapping	Training (Machine Learning)
Transparency	High and Straightforward	High but close aligned with the final output



System Overview

- Instruction Prompt + Question Text
- Probing Technique on LLM's Hidden Layers
 - Logit Lens
 - Tuned Lens
- Analyzing the Hidden Layers' Outputs and make further processing.



Question Text

Probing Technique in LlaMA 2 architecture



Instruction Prompt with Question Text

"""<s>[INST] <<SYS>>

Role and goal:

Your role is to act as an intelligent problem-solver, tasked with selecting the correct answer from a set of multiple-choice options. Your goal is to analyze the question carefully and each of the provided options, applying your extensive knowledge base and reasoning skills to determine the most accurate and appropriate answer.

Context:

The input text is a question with multiple-choice options. The correct answer is indicated by the option label A, B, C, or D.

1. Question: A clear query requiring a specific answer.

2. Options: A list of possible answers labeled with possible answer A.First possible answer B.Second possible answer C.Third possible answer D.Fourth possible answer

Instructions:

Analyze the question and options provided.

Use your knowledge to assess each option.

Employ reasoning to eliminate clearly incorrect options.

Identify the most accurate answer based on the information given.

Conclude by justifying your selection, clearly indicating your choice by referencing the option label A, B, C, or D. You should only output one capitalized letter indicating the correct answer.

Example:

Input: // you will receive the question and options here.

Output: The correct answer is {one of A, B, C, D} // you will output the correct answer, replace {one of A, B, C, D} with the correct option label A, B, C, or D.

Now you can start to answer the question with given options to give the correct answer. $<\!<\!/{\rm SYS}\!>\!>$

Input: {{inputText}}[/INST]
Output: The correct answer is """

- Role and goal
- Context
- Instructions
- Example
- Input

"{{inputText}}" with the.

Question Text

Output

Question Text – Multiple Choices Question

Question: An industry that create a lot of waste is? Options: A.Solar Power Companies B.Recycling companies C.water production companies D.Cleaning Companies.

Question: Which of the following is NOT a characteristic of perfectly competitive industry? Options: A.Free entry into the industry B.Product differentiation C.Perfectly elastic demand curve D.Homogeneous products.

Question: Which of the following elements is not part of Porter's Five Forces model for industry competitiveness? Options: A.Threat of substitutes B.Threat of suppliers C.Power of buyers D.Threat from government.

Question: Which of the following is an example of a bulk-gaining industry? Options: A.Steel B.Bottled orange juice C.Paper D.Copper.

Question: The industry that makes plastic army figures uses a small fraction of the plastic demanded for all purposes. On this basis, we can conclude that the army-figures industry is most likely a(n)? Options: A.increasing-cost industry constant-cost industry B.decreasing-cost industry C.profit-making industry.

Question: Which of the following has the highest predictive validity in personnel selection in industry? Options: A.A projective technique B.An objective personality inventory C.An interview by the personnel manager D.A biographical inventory.

Wrappers for Attention Mechanism and Block Outputs Interpretation



LLaMa 1 layer



Wrappers for Attention Mechanism and Block Outputs Interpretation



Wrappers for Attention Mechanism and Block Outputs Interpretation



Wrappers for Attention Mechanism and Block Outputs Interpretation



Wrappers for Attention Mechanism and Block Outputs Interpretation









```
self.block_output = self.norm(output[0])
```

```
attn_output = self.block.self_attn.activations
```

```
self.attn_mech_output = self.norm(attn_output)
```

```
attn_output += args[0]
```

```
self.intermediate_res = self.norm(attn_output)
```

```
mlp_output = self.block.mlp(self.post_attention_layernorm(attn_output))
```

```
self.mlp_output = self.norm(mlp_output)
```

return output

Visualizing the CSV data in Heatmaps

 Create interactive heatmaps based on each token's probabilities.

Layer	Attention mechanism
layer0	[('–ø—É—Ç–∞', 37.25), ('archivi', 22.86), ('textt', 5.72), (',ñÅkonn', 3.91), (',ñÅpartieller
layer1	[('%%%%', 2.4), ('pick', 1.4), ('kin', 1.36), ('ol', 1.05), (',ñÅPick', 0.82), ('TRAN', 0.67), ('rie
layer2	[('werb', 0.74), (',ñÅkernel', 0.72), ('√∂lker', 0.68), ('cv', 0.53), (',ñÅtun', 0.51), ('<0xA6>'
layer3	[(',ñÅsudden', 1.18), (',ñÅnoise', 0.91), (',ñÅfalse', 0.74), (',ñÅCha', 0.64), (',ñÅveget', 0.4
layer4	[(',ñÅquestions', 4.22), (',ñÅs√≠', 2.34), (',ñÅyes', 2.19), (',ñÅanswers', 1.64), (',ñÅQues
layer5	[('oro', 0.69), (',ñÅchron', 0.68), (',ñÅbear', 0.68), ('BS', 0.62), ('ach', 0.52), ('lat', 0.5), ('
layer6	[(',ñÅ_', 1.26), (',ñÅrather', 1.21), (',ñÅanswer', 1.16), (',ñÅmost', 0.53), (',ñÅdepends', 0
layer7	[('bst', 2.14), ('witz', 1.5), (',ñÅalberga', 0.96), ('acci', 0.91), (',ñÅdich', 0.65), ('ïãù', 0.59
layer8	[(',ñÅrather', 3.15), (',ñÅhands', 1.66), (',ñÅmix', 0.81), (',ñÅcul', 0.7), ('phere', 0.67), (',ŕ
layer9	[(',ñÅanswer', 1.4), ('√¥', 1.03), ('jsp', 0.97), (',ñÅlsa', 0.89), (',ñÅsim', 0.88), (',ñÅGran',
layer10	[('inas', 1.94), (',ñÅGe', 1.42), (',ñÅindeed', 1.25), (',ñÅmiddle', 0.94), (',ñÅthird', 0.52), (
layer11	[('uge', 1.62), ('phia', 0.9), ('ingsområ', 0.83), ('idos', 0.7), ('veg', 0.69), ('erdings', 0.56)
layer12	[('iemann', 2.24), ('yal', 1.22), (',ñÅthree', 1.14), ('iman', 1.03), ('alion', 0.72), ('-º-æ', 0.
layer13	[(',ñÅm√©r', 1.3), ('line', 0.87), ('ico', 0.8), (',ñÅcorrectly', 0.77), (',ñÅspot', 0.49), (',ñÅs
layer14	[('fô', 0.89), ('jest', 0.82), ('que', 0.58), ('ille', 0.58), (',ñÅign', 0.57), ('agnet', 0.49), (',ñÅo
layer15	[('psum', 1.75), (',ñÅexternos', 1.53), ('-æ—Ç-≤-µ—Ç', 1.52), ('scheid', 0.81), ('osos', 0
layer16	[(',ñÅD', 1.04), ('feld', 0.88), ('ategy', 0.64), (',ñÅdrum', 0.63), ('A', 0.59), ('iast', 0.52), (',i
1	WE ADON WERE COMPANIES OF WERE OCT WAY OCT HERE OF A

Block Output Visualization - Set 1

90

80



	0	1	2	3	4	5	6	7	8	9
1	esterni	adoop	1	bid	_	ogeneous	rgba	ependant	SELECT	_Dios
р	•	Einzelnach	mathchar	_Попис	illaume	<0xF3>		_(utos	giore
9	<0xE2>	<0×F0>	<0×EF>	~	2		<0×EE>	<0xE1>	*	•
в	P	_^	_в	_<	_c	_д	_^			-
z			turns	Receive	turno	380	aterra	_`€	бра	Id
5	P	_^	_c	B		_"			_E	
5	_//		_<	_///		////	///	_	///////	_^
4	_^		_^			at	RS	TT	el	refer
3	éric		гер	Wein	коли	férences	familjen	Night	Bien	pieler
2	_^	_externes	iformes	良	atol	房	_8	abord	CTYPE	udni
1	P	_c			_A	B	_4	Cap	_Deep	CD
D.	P		A	d	д	DE	DE	DS		DR
9		_^	_P			_(
в	_alphabet	_#	_letter	_none	None	<0xBB>	_letters	_^	Option	none
Z		_P	fün	_ker	дей		_д	unker	_lungo	TD
5	_P	feld	ategy	drum		iast	Social		swer	_None
5	psum	_externos	ответ	scheid	osos	isseur	_onderwerp	einmal	шин	Bat
4		jest	que	ille	ign	agnet	odd	ouse	dienst	nier
3	mér	line	ico	correctly	spot	strugg	ovie	rela	_skill	정
2	iemann	yal	three	iman	alion	мо _	_gepubliceer	1_Schön	条	Wol
1	uge	phia	ingsområ	idos	veg	erdings	uchte	baut	ius	Paolo
D	inas	_Ge	_indeed	middle	third	umar	iker .	_externas	centre	_above
9	answer	ô	jsp	_Isa	sim	_Gran	_Zug _	_answered	Ch	ago
в	rather	_hands	mix	_cul	phere	Mitchell	Hinweis	sources	_properly	нин
Z	bst	witz	_alberga	acci	dich	식	hat	crisis	_number	izza
5		rather	answer	most	depends	_sens	destination		audi	decimal
5	oro	chron	bear	BS	ach	lat		arms	Ts	ellow
4	_questions	_sí	_yes	_answers	_Question	Answer	answer	satisf	Answer	_answered
3	sudden	noise	false	Cha	_veget	mér	kar	Filter	Bal	_log
2	werb	_kernel	ölker	cv	tun	<0xA6>	itself	ote	_im	olf
1	%%%%%	pick	kin	ol	Pick	TRAN	ric		ing	osa



	Logit Lens	Tuned Lens
Attention Mechanism	Both look similar on each layer's attention	on mechanism output. However
	 The option labels begin (A, B, C, D) to be focused in layers 17-20 More unrecognized words or non-English words, difficult to interpret 	 The option labels begin to be focused in layers 16-18 More readable English words
Block Output	 Random word outputs before layers 17 - 20 The option label outputs after layers 17 - 20 	 The option number(1, 2, etc.) outputs before layers 17 - 18 The option label outputs after layers 17 - 18

	Logit Lens	Tuned Lens
Attention Mechanism	Both look similar on each layer's attentio	n mechanism output. However
	 The option labels begin (A, B, C, D) to be focused in layers 17-20 More unrecognized words or non-English words, difficult to interpret 	 The option labels begin to be focused in layers 16-18 More readable English words
Block Output	 Random word outputs before layers 17 - 20 The option label outputs after layers 17 - 20 	 The option number(1, 2, etc.) outputs before layers 17 - 18 The option label outputs after layers 17 - 18

	Logit Lens	Tuned Lens					
Attention Mechanism	Both look similar on each layer's attention	on mechanism output. However					
	 The option labels begin (A, B, C, D) to be focused in layers 17-20 	The option labels begin to be focused in layers 16-18					
	 More unrecognized words or non- English words, difficult to interpret 	More readable English words					
Block Output	 Random word outputs before layers 17 - 20 	• The option number(1, 2, etc.) outputs before layers 17 - 18					
	 The option label outputs after layers 17 - 20 	The option label outputs after layers 17 - 18					

	Logit Lens		Tuned Lens
Attention Mechanism	Both look similar on each layer's attentio	n r	nechanism output. However
	 The option labels begin (A, B, C, D) to be focused in layers 17-20 	•	The option labels begin to be focused in layers 16-18
	 More unrecognized words or non- English words, difficult to interpret 	•	More readable English words
Block Output	 Random word outputs before layers 17 - 20 	•	The option number(1, 2, etc.) outputs before layers 17 - 18
	The option label outputs after layers 17 - 20	•	The option label outputs after layers 17 - 18

Result Comparison of Logit Lens and Tuned Lens

	Logit Lens		Tuned Lens					
Attention Mechanism	Both look similar on each layer's attentio	n m	echanism output. However					
	 The option labels begin (A, B, C, D) to be focused in layers 17-20 	•	The option labels begin to be focused in layers 16-18					
	 More unrecognized words or non- English words, difficult to interpret 	•	More readable English words					
Block Output	 Random word outputs before layers 17 - 20 	•	The option number(1, 2, etc.) outputs before layers 17 - 18					
	The option label outputs after layers 17 - 20	•	The option label outputs after layers 17 - 18					

• Easier Question: Question: Which city is the capital of Germany? Options: A.London B.Paris C.Berlin D.Amsterdam

Evaluation Question: Which city is the capital of Germany? Attention Mechanism Output (Easier Question) Options: A.London B.Paris C.Berlin D.Amsterdam

Attention Probing Visualization - Set 1

Layer

Token

Logit Lens

	0	1	2	3	4	5	6	7	8	9				0	1	2	3	4	5	6	7
31		1					answer	concaten	BE	correspond			31	_	rgba	_concaten		—	unicí	vens	én
30	ços	giore	bine	бен	chte <u>n</u>		ló	lá	`\$				30	ços	bine	évrier	giore	chten	бен		lå
29			٠		scheidung	berger	8.#	ató	ст	٠			29	<0x0A>		<0xF0>	<0xE2>	8.#		B	schei
28	D	С	Α	в	с	Α	D	в	д	Ber	90)	28	_0	_c	_^	_в	с	A	D	E
27	the	turns	based		the	THE	enjo	The	following				27	<0x0A>	the	_turns	based	_THE	_the	enjo	!</th
26	Ber	Ber	ber	Berlin	BER	Бер	Berl	D	ber	бер			26	Ber	Ber	Berlin	_ber	_P	_^	BER	_6
25	//	//		////	///	://	/	W	upon	///////	80)	25	_//	//	_///	////	://	///		_
24	Ber	ber	Ber	Berlin	Бер	Berliner	BER	Berl	ber	berry			24	Ber	Ber	ber	Berlin	_Бер	Berliner	BER	be
23	Bien	chter	={{	édia	night	Buen	asto	Night	phantom	parallel			23	Bien	night	édia	={{	Night	chter	parallel	{
22	A	ш	Cs	##	//	penas	//	terminal	#,	heimer	70)	22	_^	_//		##	_Cs	//	penas	heir
21	D	ber	ber	Ber	Ber	berga	Cong	Berl	BER	Theater			21	_0	ber	ber	Ber	Ber	_Cong	BER	C
20	D	correct	D	Correct	correct	д	Capital	DR	capital	д			20	D	D	_correct	_Correct	A	correct	_d	_Ca
19	omial	oir	emberg	Bau	ovis	ensional	German	WM	zos	sterd	60		19	emberg	oir	ovis	omial	Bau	_German	Balt	na
18	chen	ilia	blatt	maj	oma	٠	patr	hus	idea	fF			18	chen	ilia	blatt	<0x81>	Bedeut	patr	oma)
17	pse	Pla	={{	//	touch	lyph	properly	bmatrix	не	дро			17	pse	={{	Pla	//	touch	lyph	bmatrix	pma
16	feld	ouv	п	ategy	aient	nos	irc	ël	unci	Уg	50	Yer	16	feld	ategy	unci		nos	Уg	ouv	in
15	onderwerp	ответ	ibile	einmal	Car	öd	Jung	beh	уче	Ziel	50	La	15	Car	ibile	einmal	ответ	_в	_Jung	_onderwer	Р _b
14	cios	odd	patri	icia	til	NER	bold	改	monarch	rapper			14	cios	perty	rapper	icia	NER	改	odd	bin
13	uga	références	reve	пора	atro	eni	atem	yles	organ	root			13	uga	reve	_organ	_root	spot	fit	_swe	_he
12	Hint	ane	jan	PC	qué	Hier	dere	ji	Außer	ор	40		12	qué	Hint	Außer	PC	jan	dere	ane	H
11	cov	itare	tid	ucci	дела	mad	bage	Graphics	cov	房			11	cov	itare	tid	房	дела	schau	bage	_c
10	素	gepubliceerd	prep	ymnasium	lès	externas	externos	asm	orum	hi			10	prep	_hi	_prepar	素	asm	Guy _	_gepublicee	rd ita
9	rör	<u>.</u>	宇	kill	pool	tiny	Pool	bare	anie	answer	30		9	rör	based	_pool	answer	宇		firm	ki
8	ede	٠	Freder	pal	лыи	ée	fraction	Wol	iske	Invoke			8	љи	<0x82>	ść	_Kilometer	TeX	_Freder	ede	Wiki
7	av	Colors	div	anz	ct	ť	gresql	bst	rgb	dar			7	av	gresql	_Cultura	_Colors	_alberga	ť	平	bs
6	лет	gresql	÷	iwers	lette	Render	rendering	correl	Kor	render	20)	6	pron	vers	_jud	rendering	correl	_lean	_Kor	
5	luss	endorf	Om	Haus	moz	пе	nem	ове	cout	**			5	Om	nem	luss	пе	_Haus	endorf	_**	_0
4	lius	arte	род	superfic	mund	Haupt	Bog	eben	Lund	arrison			4	lius	arte	tight	род	mund	arrison	_Haupt	B
3	動	Års	penas	narrow	това	ített	vo	Point	ított	управ	10)	3	_more	_we	_or	Great	_to	first	narrow	s
2	werb	٠	led	asp	comm	ret	Janeiro	derived	adin	ander			2	werb	<0xAB>	led	ander	comm	asp	ret	w
1	refer	Hace	Cult	cus		Pick	pick	refer	ing	ol			1	ol	refer	cus	oc	osa	ula	pick	_H
0	penas	пута	sier	trightarrow	щ	<s></s>	hely	archivi	ímp	gat	0		0	пута	penas	_sier	archivi	trightarrow	씨	sime	k
							-						-								

Attention Probing Visualization - Set 1

Token

Tuned Lens

University of Stuttgart, IAS

9

eground Ιó

berger __Ber

> Turn _в

_upon

berry <0xF4>

> A _Berl

DR

ensional

nem

не

aya öd

arod

release

LIC _mad

alem

tiny îne

dar

name ano -> related cri aie partiellement totalité

ainx

ст

ber ///////

__Berl

cius terminal

_Des

canita Z05

idea

ąd

aient

лка 호

références

ucci mnasium

_Pool

render

Invoke arina

_д The 100

80

60

40

20

Evaluation Block Output (Easier Question)

Block Output Visualization - Set 1

Layer

Token

	0	1	2	3	4	5	6	7	8	9
1	С		A		в	BER	{	Ber	//	
D	С	A	D		в	С	D	Α	в	//
9		с	D	D	с	A	в		в	ber
8	D	с	A	Α	С	D	в	в	Ber	Ber
7		с	A	Α	с	Berlin	//	Ber	D	Ber
5		с	A	Berlin	Ber	//	Berl	Ber	ber	с
5	D	с	//	Α	//	С	Berl	{	capital	в
ļ	D	с	Α	Berl	recens	с	D	landa	capital	Berlin
3	D	recens	Capital	capital	landa	DI	Cic	maven	ą	Ā
2	D	capital	ą	Capital	THE	с	D	recens	Cic	THE
l	D	Capital	Cic	DI	capital	С	THE	recens	ieg	D
þ	capital	D	landa	Capital	aze	а	ieg	maven	emb	diam
9	ieg	landa	jet	nings	aze	land	capital	drum	diam	hyp
3	landa	land	ieg	nings	inton		champion	mana	li	capital
7	ds	ieg	uda	landa	inton	INE	hyp	hn	éta	noc
5	hyp	hn	coff	ds	landa		concentr	li	Herzog	mana
5	ds	backend	hn	otte	么	uda	rav	realized	concentr	diam
ł	ds	ulle	backend	bir	uda	George	MY	ison	nam	capital
ł	ulle	TD	lon	nea	vd	egg	uda	ison	VD	ital
2	٠	([ital	egg	backend	Sch	oro	nea	fol
	penas	&=\	ital	chosen	joint	tid	vd	Sch	от	со
þ	penas	ital	avia	Hat	•	&=\	ols	bes	BD	estaven
9	penas	oro	avas	BD	differ	espèce	comprom	backend	Марр	avia
3	Pat	eta	hands	Hat	oro	object	rain	ital	mob	•
7	٠	single	espèce	confident	Pat	arina	zza	avas	Ñ	önig
5	totalité	ahu	ling	٠	espèce	pó	rot	avas	penas	候
5	penas	totalité	urt	aut	mat	estaven	ling	ach	superfic	fle
ŧ	ilib	avas	penas	ellig	ph	totalité	ченко	attend	ahu	égl
3	penas	iennes	totalité	kne	sight	automatically	ner	Hay	avas	ph
2	ph	penas	cas	kne	totalité	식	Cas	iennes	repub	евич
L	penas	conde	ə	рó	х	totalité	θ	maste		ченко
0	IIYIa	penas	씨	sier	archivi	Licencia	ímp	embros	conde	瀨

Block Output Visualization - Set 1

50

40

Layer

20

10

Token

	0	1	2	3	4	5	6	7	8	9
31	С		Α	<0x0A>	в	BER	{	Ber	//	
30	С	A	D	в	<0x0A>	The	the	//	_D	_c
29	A		в	D	<0x0A>	_^	The	_P	the	_c
28	A		D	в	_D	_^	_c	<0x0A>	The	м
27		с	в	<0x0A>	D	The	м	G	_D	the
26	A	с	в	D	<0x0A>	_P	_^	The	м	_c
25		<0x0A>	в	Α	D	The	//	_^	the	_P
24		A	D	<0x0A>	_P	в	_^	The	_c	the
23		A	D	<0x0A>	в	The	the	_D	(d
22	с		_D	<0x0A>	A	в	The	the	_c	R
21	D	_P	с	<0x0A>	The	в	A	the	(d
20	The	<0x0A>	the	в	A	D	с	к	_D	w
19	The	<0x0A>	к	в	the	с	w	A	R	N
18	The	A	в	с	к	the	F	land	R	<0xF0>
17		<0x0A>	The	the	<0x09>	(This	к	А	
16		<0x0A>	The	к	A	the	N	м	-	He
15		<0x0A>	Α	к	(в	а	My	Com	The
14		<0x0A>	the	My	ds	(This	x	Α	a
13		<0x0A>	I	He	в	the	Sch	(а	The
12	(Sch	_(<0x0A>	_[_{	а	_	_*
11	<0x0A>	Sch	в	(а	in	we	the	_(d
10	<0x0A>	(в	Sch		m	k	_	1	the
9	<0x0A>			(_(L	в	k	Sch
8	<0x0A>	1		в		(_	0	_[_(
7	<0x0A>	1	3	_		4	0	[L	_[
6	1	[4	3	L	_	7	0	ling	2
5	1	4	0	3	ach	2	7	urt	aut	ľ
4	for	1	at	on	in	0	t	g	тм	I
3	1	in	m	0	р	to	be	_kne	t	Your
2	1	In	<0x09>	in	9	ph	_kne	to	[We
1	1	In	6	0	This	With	2	If	The	7
0	пута	penas	_sier	archivi	_konn _	partiellement	씨	virtuel	_cí	_Licencia

University of Stuttgart, IAS

Tuned Lens

60

50

40

30

20

10

Summary and Outlook

Summary and Outlook

Comparative Effectiveness in Model Layers

Attention Mechanism – Tuned Lens yields more predictive and understandable token interpretations than Logit Lens in complex prompts.

Block Output – The performance of the Tuned Lens is more consistent across various hidden layers compared to that of the Logit Lens.

Training own Tuned Lens

The decoding result could be improved by **training own Tuned Lens** for different LLMs.

Limitations and Model Capacity

Llama-2-7B might not be able to handle complicated queries effectively. Upgrading to more robust models like Llama-2-13B or Llama-2-70B could enhance the accuracy of both intermediate and final results.



University of Stuttgart Institut of Industrial Automation and Software Engineering

Thank you!



Lun-Yu Yuan

e-mail <u>st184762@stud.uni-stuttgart.de</u> phone +49 (0) 711 685fax +49 (0) 711 685-

University of Stuttgart Institut of Industrial Automation and Software Engineering Pfaffenwaldring 47, 70550 Stuttgart, Germany



Source

[1] <u>https://pub.aimind.so/demystifying-the-black-box-interpretable-machine-learning-6388bef4aeca</u>

- [2] <u>https://arxiv.org/pdf/1706.03762</u>
- [3] <u>https://cameronrwolfe.substack.com/p/llama-2-from-the-ground-up</u>

[4] <u>https://www.lesswrong.com/posts/fJE6tscjGRPnK8C2C/decoding-intermediate-activations-in-llama-2-7b#Future_work</u>

[5] <u>https://arxiv.org/pdf/2303.08112</u>