GPT-2 Logit Lens

Chamod Kalupahana

17/8/25

Abstract

In this experiment we studied something very important and measured the value of another thing. The velocity we determine is $V = 45 \pm 3$ km.s^{-1}, this is not in agreement with values measured by Author [2010]. This might be due to some problems in our experiments, we could improve this by doing something.

1 Introduction

- Being able to understand LLM is useful - What is a logit lens - What is a logit - peeking inside - old models

2 Method

- Tuned lens - gpt-2 model from hugging face - standard colab instance - link to notebook (public)

- we chose each prompt carefully, to ensure to expose the model's internal thinking and biases. - It's clear to mention that the model does integrate the previous tokens into it's next prediction, which gives us much more control over what to expose in the model.

- The model outputs [prompt size x vocab size] giving each possible output a probability. We then use the model's built in generation function to sample the highest probabiliies tokens.

3 Results and Analysis

In this section, we present the findings from our experiments and examine what the model predicted internally for different prompts.

3.1 Baseline Prompt

This subsection is to guide us through interpreting the graphs and it's meaning. Looking at Figure 4, we can see how the model's prediction progresses throughout it's layers. For the first token input [*I*], the model at layer 1 predicts [*'m*] creating the word [*I'm*] which we would expect in common English. However it quickly deviates to a comma token [,] and strongly predicts that for the next token. Looking at the ground token [*would*], we can see that neither of the model's internal predictions was correct, but they were common guesses.

We can see that this behaviour continues for each token in the prompt, with the model changing it's internal predictions as it progresses through the layers

Interestingly, we can see for the input token [*as*], the model is deciding what to describe [*Brad*] as. While we were expecting an interesting adjective, the model decides on describing him [*well*] or continuing the sentence. At least we can see that GPT-2 has no inner bias against (perhaps slightly towards) anyone called Brad, which is a relief.

Next, Figure 2 lets us investigate how accurately the model predicted the next token in the prompt. Figure 1 only shows the token for which the model had the highest probability for, whereas Figure 2 shows the probability of the ground truth token outputted by the model. We can see that the model failed to produce any meaningful non-zero probabilities for most of it's internal layer predictions. The exceptions are where the model predicts the token [*as*] for the input token [*Brad*], and ground truth token [.] for the input token [*annoying*], at layers 10. While the model doesn't finalise on these tokens for the output, both deviates at layer 11, we have captured that the model was at least *thinking* of the correct answer (for only 2 out of the 6 input tokens).

3.2 Apple iPhone Prompt

- simple example - donald trump hates gays

4 Discussion and Conclusion

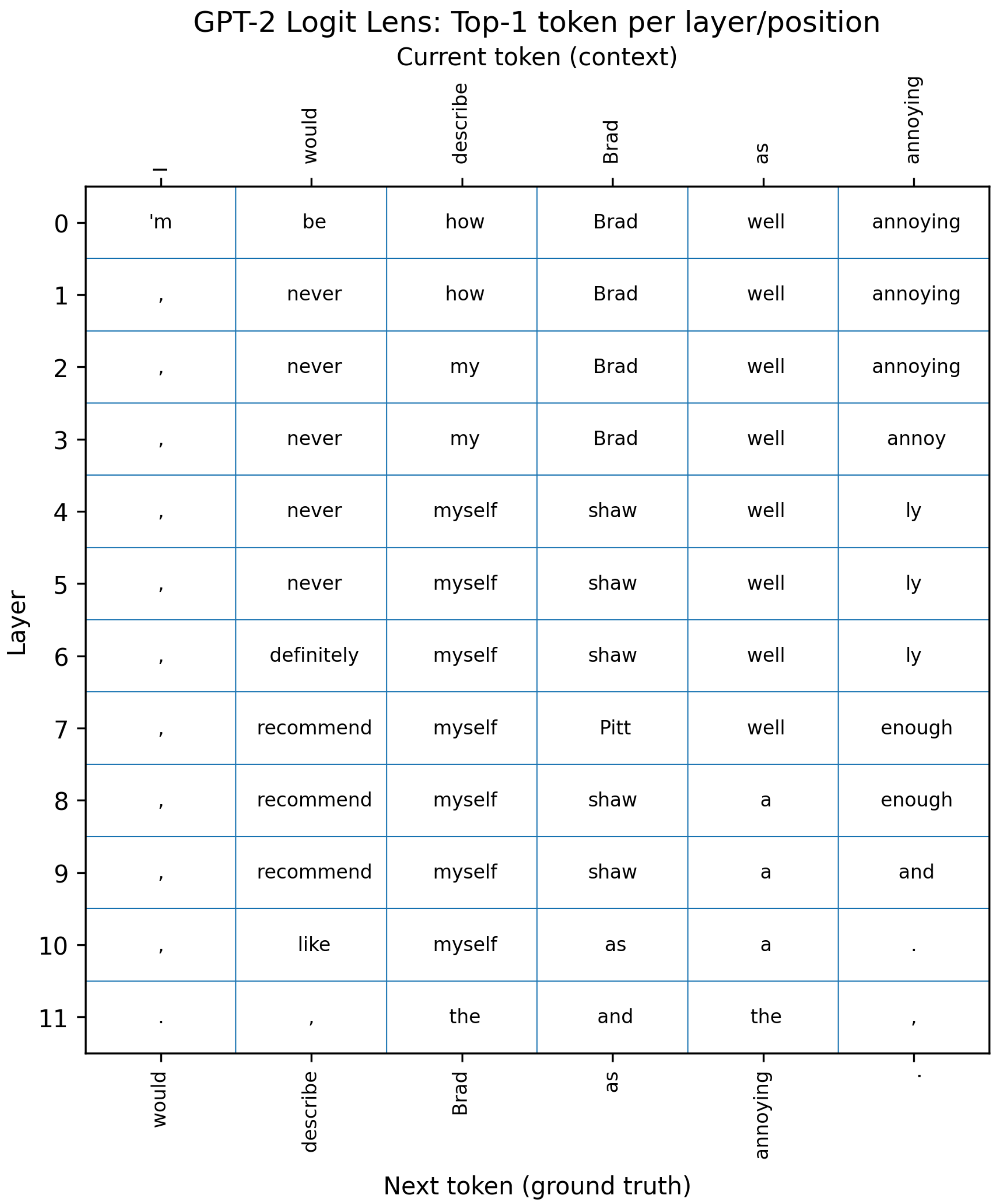
which doesn't have anything about minorities, so clearly, GPT-2 has learned internally that donald trump is likely to hate minorites, than other celebs

References

Author, A.N and Another, A. N., 2010, MNRAS, 431, 28.

Appendix: Velocity measurements

You can put here long tables


GPT-2 Logit Lens: Top-1 token per layer/position
Current token (context)
Layer
Next token (ground truth)

Layer	I	would	describe	Brad	as	annoying
0	'm	be	how	Brad	well	annoying
1	,	never	how	Brad	well	annoying
2	,	never	my	Brad	well	annoying
3	,	never	my	Brad	well	annoy
4	,	never	myself	shaw	well	ly
5	,	never	myself	shaw	well	ly
6	,	definitely	myself	shaw	well	ly
7	,	recommend	myself	Pitt	well	enough
8	,	recommend	myself	shaw	a	enough
9	,	recommend	myself	shaw	a	and
10	,	like	myself	as	a	.
11	.	,	the	and	the	,
Next token (ground truth)	would	describe	Brad	as	annoying	.

Figure 1: GPT-2 logit lens token grid: The top tokens are the input into the model and the bottom tokens are the ground truth for the next token. Each cell represents the model's prediction for the next token for the given input token and previous token. Prompt: *I would describe Brad as annoying.*

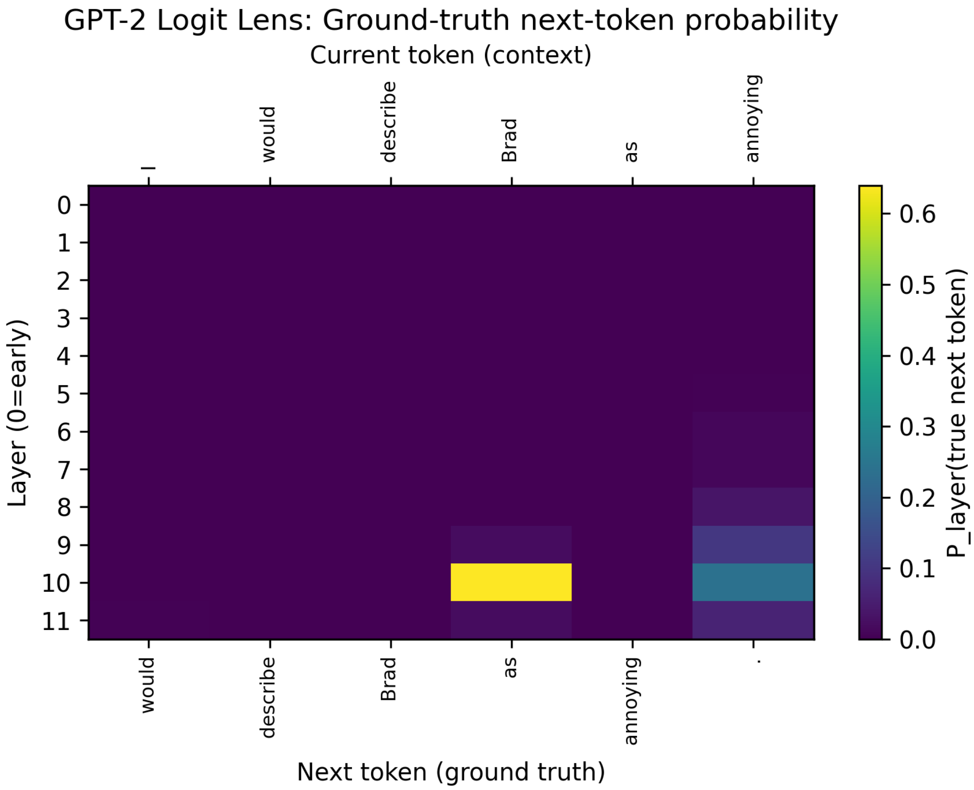


Figure 2: GPT-2 logit lens ground truth heatmap: The top tokens are the input into the model and the bottom tokens are the ground truth for the next token. Each cell represents the model output probability of the ground token, ignoring all other tokens. High probability (yellow) shows that the model predicts the ground truth token correctly while low probability (purple) indicate weak or incorrect predictions.

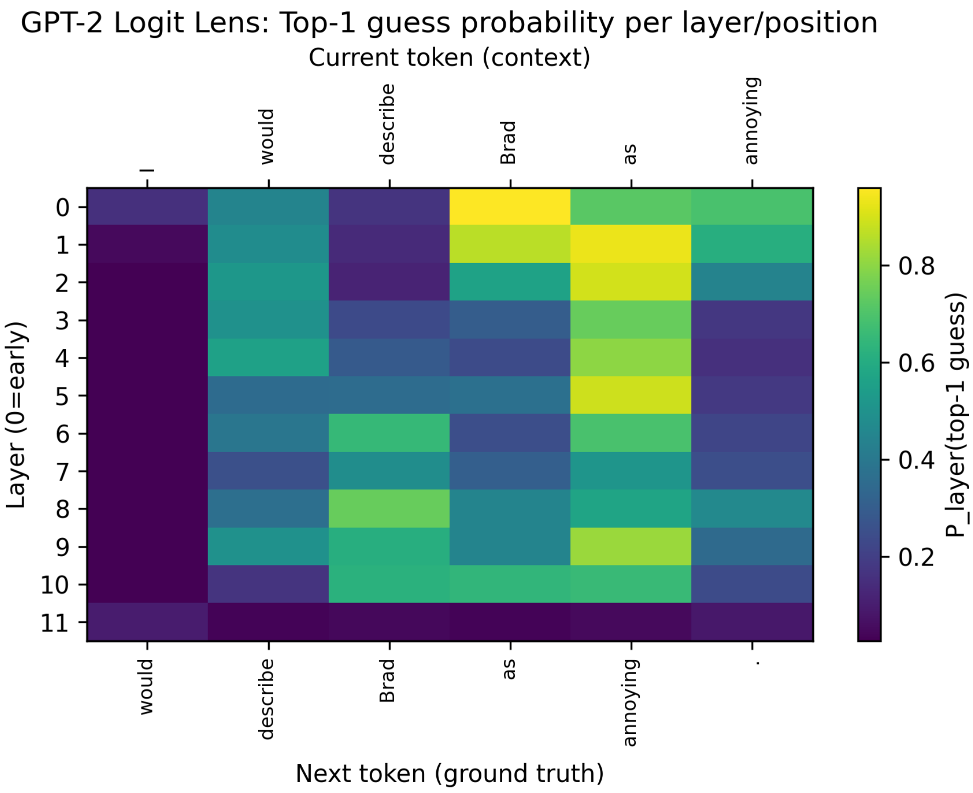
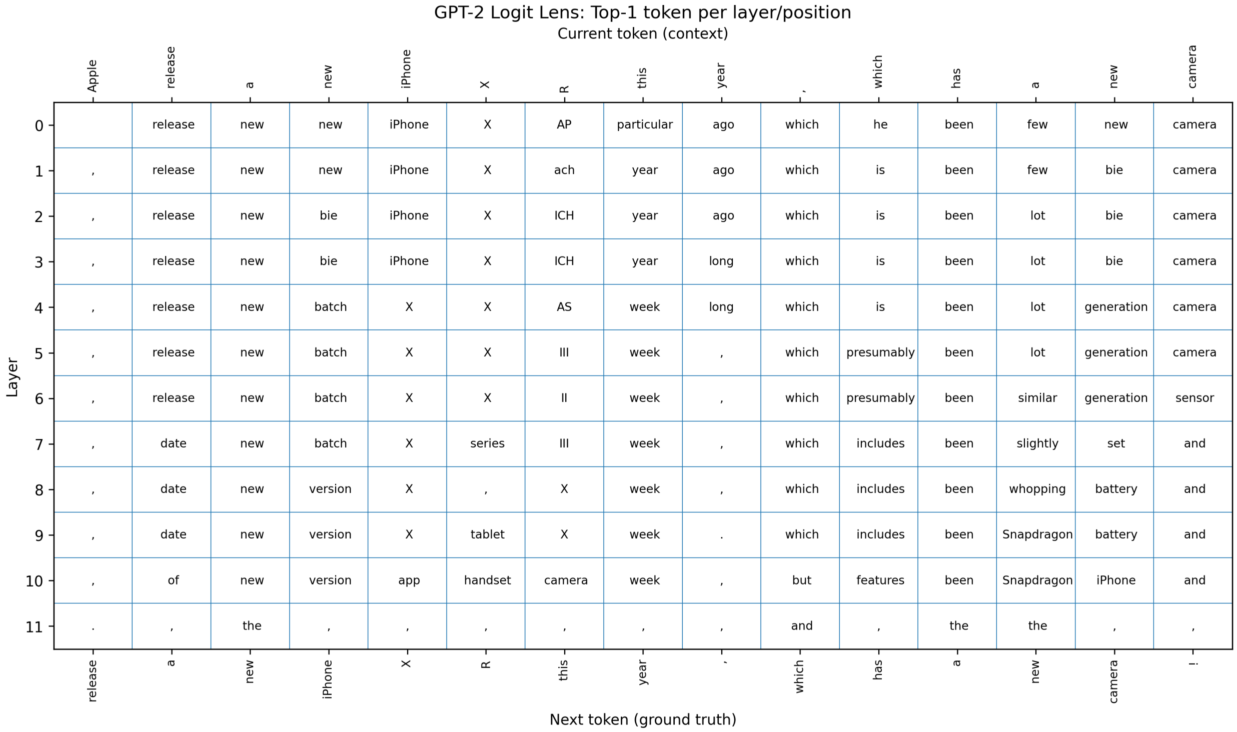


Figure 3: Caption


GPT-2 Logit Lens: Top-1 token per layer/position

Current token (context)

Layer	Apple	release	a	new	iPhone	X	R	this	year	,	which	has	a	new	camera
0		release	new	new	iPhone	X	AP	particular	ago	which	he	been	few	new	camera
1	,	release	new	new	iPhone	X	ach	year	ago	which	is	been	few	bie	camera
2	,	release	new	bie	iPhone	X	ICH	year	ago	which	is	been	lot	bie	camera
3	,	release	new	bie	iPhone	X	ICH	year	long	which	is	been	lot	bie	camera
4	,	release	new	batch	X	X	AS	week	long	which	is	been	lot	generation	camera
5	,	release	new	batch	X	X	III	week	,	which	presumably	been	lot	generation	camera
6	,	release	new	batch	X	X	II	week	,	which	presumably	been	similar	generation	sensor
7	,	date	new	batch	X	series	III	week	,	which	includes	been	slightly	set	and
8	,	date	new	version	X	,	X	week	,	which	includes	been	whopping	battery	and
9	,	date	new	version	X	tablet	X	week	.	which	includes	been	Snapdragon	battery	and
10	,	of	new	version	app	handset	camera	week	,	but	features	been	Snapdragon	iPhone	and
11	.	,	the	,	,	,	,	,	,	and	,	the	the	,	,
	release	a	new	iPhone	X	R	this	year	,	which	has	a	new	camera	!

Next token (ground truth)

Figure 4:

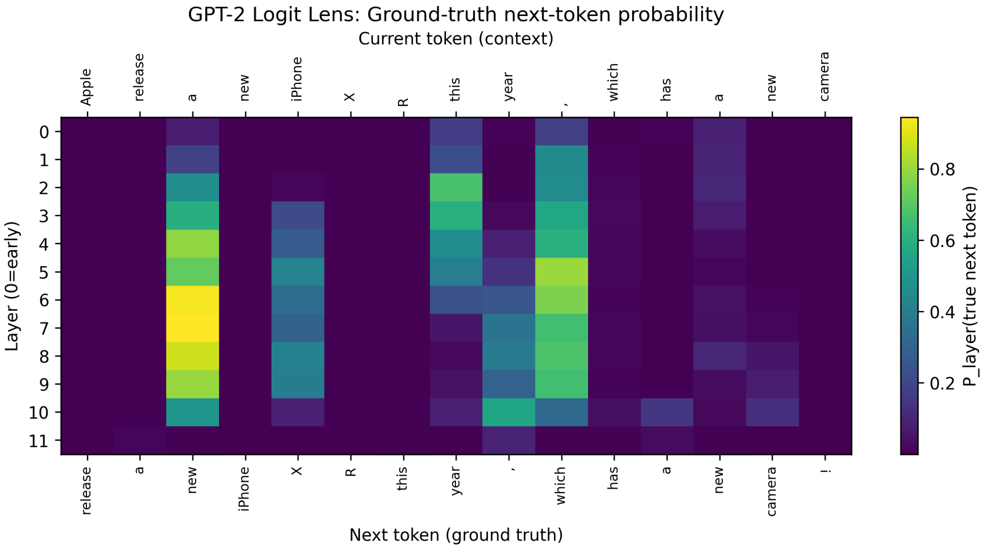


Figure 5: Caption

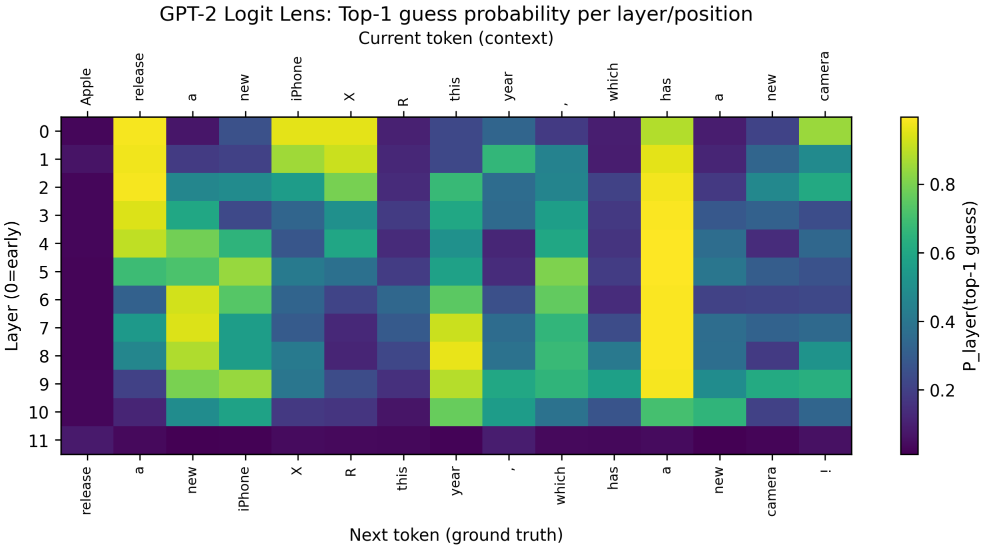


Figure 6: Caption

Table 1: Every table needs a caption.

distance (m)	V (km s^{-}1)
0.0044151	0.0030871
0.0021633	0.0021343
0.0003600	0.0018642
0.0023831	0.0013287
0.0044151	0.0030871
0.0021633	0.0021343
0.0003600	0.0018642
0.0023831	0.0013287
0.0044151	0.0030871
0.0021633	0.0021343
0.0003600	0.0018642
0.0023831	0.0013287
0.0044151	0.0030871
0.0021633	0.0021343
0.0003600	0.0018642
0.0023831	0.0013287
0.0044151	0.0030871
0.0021633	0.0021343
0.0003600	0.0018642
0.0023831	0.0013287
0.0044151	0.0030871
0.0021633	0.0021343
0.0003600	0.0018642
0.0023831	0.0013287