Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics

Zhu Liu, Cunliang Kong, Ying Liu, Maosong Sun

Tsinghua University

How Do LLMs Encode Lexical Semantics?

- GPT-like models
  - access only preceding context

the bank | along the river

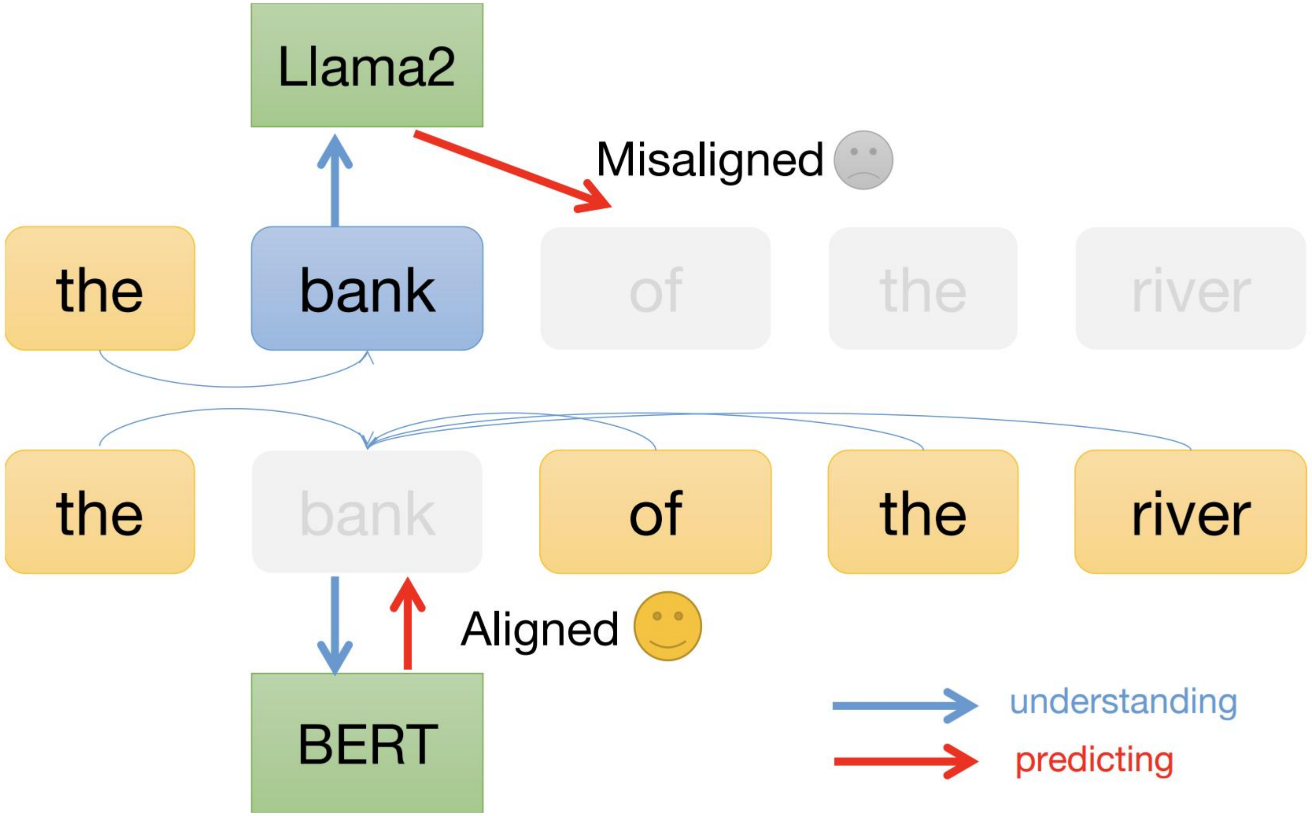
the bank | where you deposit

the two *bank* instances cannot be distinguished

  - utilize the objective of predicting the next token

different layers have varying understanding of contextual information and different abilities to predict the next word

How Do LLMs Encode Semantics?



Structural differences between
BERT and LLAMA2

How Do LLMs Encode Semantics?

- Research Question

To what extent and through which layer do LLMs encode lexical semantics?

- Hypothesis

GPT-like LLMs encode lexical semantics in shallow layers while making predictions, potentially leading to the forgetting of information related to current tokens in deep layers.

Method

- Word in Context (WiC) Task

a binary classification task

True ***Air*** pollution — Open a window and let in some ***air***
False the ***bank*** of the river — the ***bank*** where you deposit

- Method: 1) extract the layer-wise Llama2 representations with different settings.
  2) classify the pair according to cos-simiarity score by a learnable threshold.

base the **bank** of the river

to better utilize the context

repeat the bank of the river the **bank** of the river

repeat_prev the bank of the river the bank of the river

to show the prediction ability

prompt The bank in this sentence: “the bank of the river” means in one word:

Observations

- Llama2 has the potential for word-level understanding
- prompting is the most effective method for Llama2
- repeat strategy is comparable to prompting and outperforms the base strategy
- verbs are generally more challenging to disambiguate
- anisotropy removal improves the performance

Method	All	Noun	Verb
Human	80.0	-	-
Random	50.0	-	-
WSD	67.7	-	-
BERT_large†(23)	67.8	69.1	67.6
BERT_large (22)	71.0	70.7	71.5
Context2vec	59.3	-	-
Elmo	57.7	-	-
Llama2_base†(6)	60.9	63.7	58.3
Llama2_base (11)	63.6	66.8	58.7
Llama2_repeat†(9)	64.5	66.4	63.4
Llama2_repeat (8)	68.1	<u>72.7</u>	65.6
Llama2_prompt†(28)	<u>71.1</u>	68.9	**72.9**
Llama2_prompt (21)	**72.7**	**74.5**	<u>72.1</u>

Overall accuracy (%) on the WiC test set

Observations

- base & repeat
  - increase in shallow layers
  - decrease in deep layers
- BERT-Large
  - obtains the best performance in higher layers

- lower layers in Llama2 might encode lexical semantics

Layer-wise acc (%) for different settings

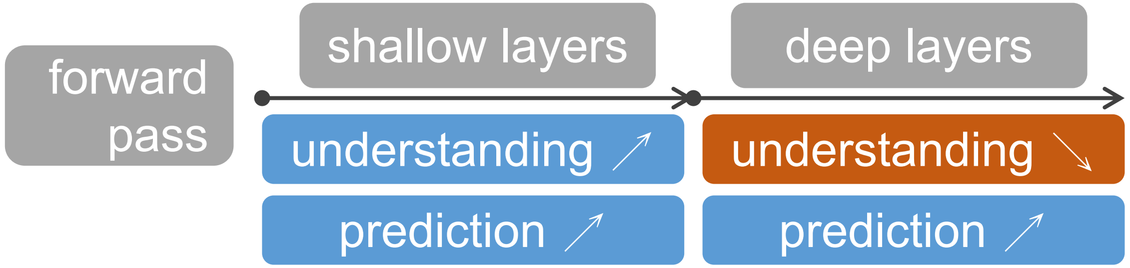
Observations

- repeat
  - increases in shallow layers
  - decreases in deep layers
- repeat_prev & prompt
  - monotonically increase
- while the understanding may diminish as layers go deeper, the prediction ability improves

Layer-wise acc (%) for Llama2 settings

Takeaways

- This study investigates how Llama2’s forward-pass layer-wise representations encode lexical semantics using the WiC dataset.
- Llama2 might prioritizes understanding before prediction as information flows from shallow to deep layers.
- These findings may offer practical guidance on extracting lexical representations.



Contact

Personal Website