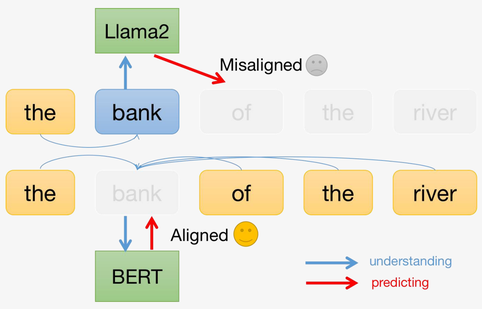
Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics

Zhu Liu, Cunliang Kong, Ying Liu, Maosong Sun
liuzhu22@mails.tsinghua.edu.cn

Introduction

How do generative LLMs encode lexical semantics?

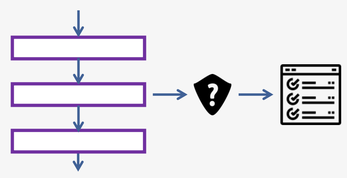
- Decoder-only model vs. Encoder-style model



1. Access only preceding context
How to disambiguate "bank"?
2. Misaligned objective of understanding and predicting
Understanding the current word, but target the next one?

- Interpretability from a representational view

1. Geometric probing without training a probing classifier
2. Layer-wise dynamics
3. Top-down interpretability



Aims

- 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

- Probing task: WiC (Word in Context)
  Whether words in two contexts have the same meaning?

Air pollution - Open a window and let in some air √
The bank of the river - the bank where you deposit ×

- Model
  Llama-2 7B base with various settings; BERT-Large

- Settings - Where to extract representations?

base	the **bank** of the river
repeat	the bank of the river the **bank** of the river
repreat_prev	the bank of the river **the** bank of the river
prompt	The bank in this sentence: "the bank of the river" means in one word **:**

to use all the context

represents the next token?

- Method - How to make the binary classification?

Results

Layer-wise performance shows two trends: ↗ and ↗↘

↗ Increasing trend: prompt; repreat_prev; BERT_large

↗↘ non-monotonic trend: base; repeat;

Indicating Llama encodes lexical semantics before predicting (meaning of the next token)

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	72.7	65.6
Llama2_prompt†(28)	71.1	68.9	**72.9**
Llama2_prompt (21)	**72.7**	**74.5**	72.1

- Llama2 (especially with prompting) has the potential for word-level understanding
- repeat strategy is comparable to prompting and outperforms the base strategy
- verbs are generally more challenging to disambiguate
- anisotropy removal is better

References

- Wang et al., 2023a: Label words are anchors: An information flow perspective for understanding in-context learning. *EMNLP*
- Touvron et al., 2023: Llama 2: Open foundation and fine-tuned chat models. *arxiv*
- Zou et al., 2023: Representation engineering: A top-down approach to ai transparency. *arxiv*
- Jiang et al., 2023: Scaling sentence embeddings with large language models. *arxiv*
- Ethayarajh, 2019: How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *EMNLP-IJCNLP*

Paper: https://arxiv.org/abs/2403.01509

Code: https://github.com/RyanLiut/LLM_LexSem

ACL 2024 Bangkok, Thailand
Tsinghua University