ACL 2025 VIENNA

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Learning to Look at the Other Side: A Semantic Probing Study of Word Embeddings in LLMs with Enabled Bidirectional Attention

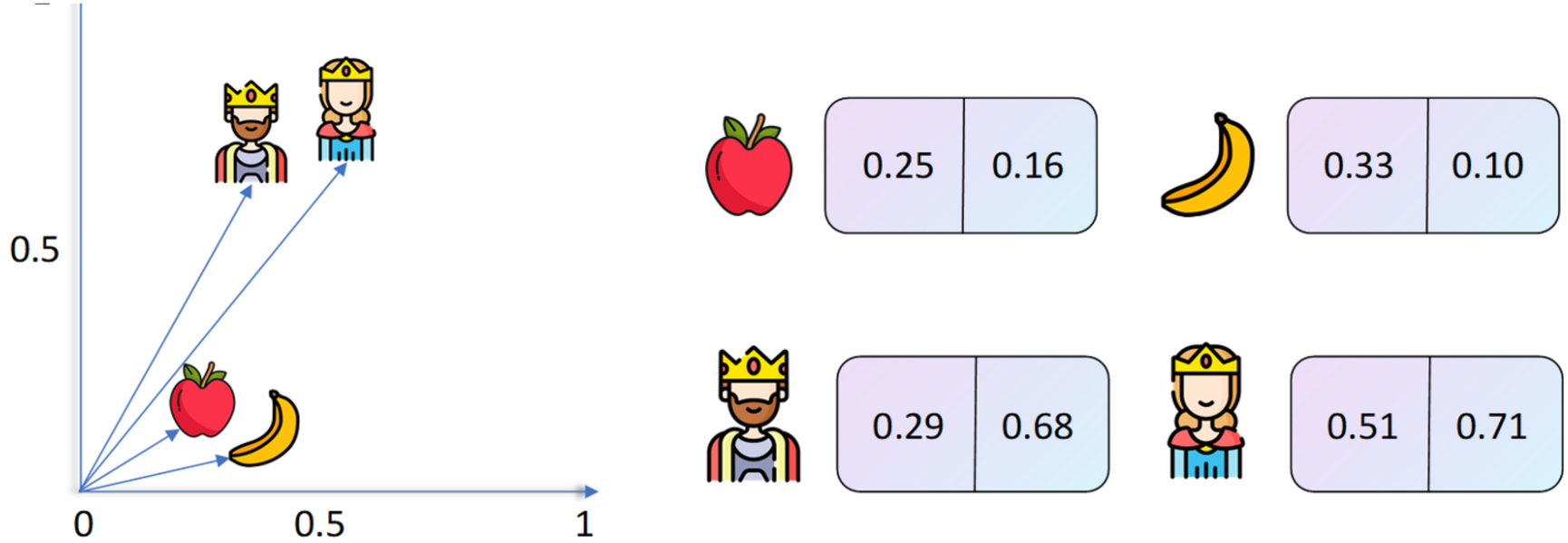
Zhaoxin Feng, Jianfei Ma, Emmanuele Chersoni , Xiaojing Zhao and Xiaoyi Bao

June 12th, 2025

One sentence to summarize what we do:

We use five semantic probing tasks, to examine how bidirectional attention influences LLMs' text embeddings on word level.

Text Embedding



https://towardsdatascience.com/deep-learning-for-nlp-word-embeddings-4f5c90bcdab5

- Text embedding: converting text into a *numerical vector representation* so that computers can better understand and process the text.
- Applications: text classification, clustering, and information retrieval.

How to get the text embedding?

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Language Models can use the ***context*** to extract the embedding of target word.

"bed" is the target word

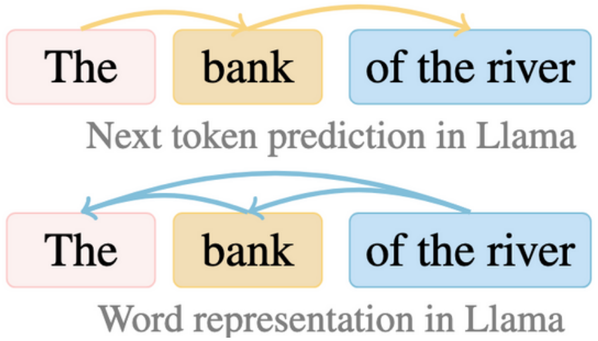
There is a lot of trash on the **bed** of the river

[x, y, z, ...] --> Embedding!

a high-dimensional vector as the embedding of "*bed*", which encodes the word semantics

Encoder-Only & Decoder-Only

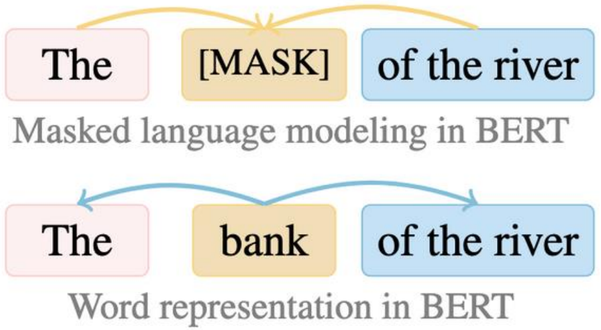
The ability of different types of language models to utilize context varies…



Decoder-Only:

Unidirectional Attention

Can only encode the *left-hand context* into the word embedding.



Encoder-Only:

Bidirectional Attention

Can encode the *both hands context* into the word embedding.

Encoder-Only & Decoder-Only

"bed" is the target word

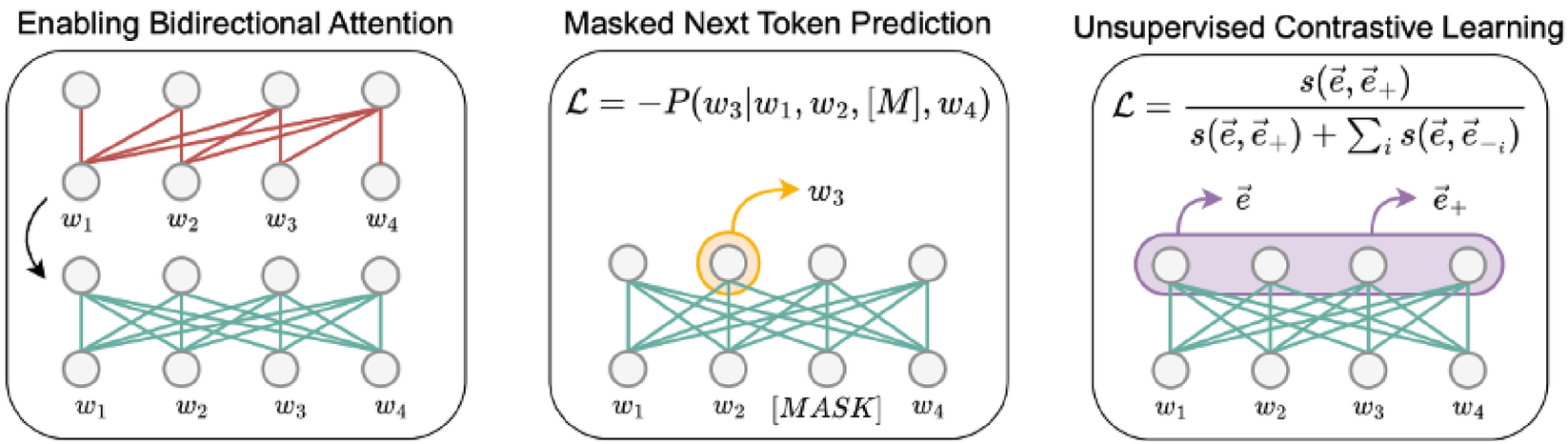
There is a lot of trash on the **bed** of the river

Encoder-only models can make use of both directions' context, but decoder-only models can only use the left-hand context.

However, we hope to use decoder-only models (autoregressive LLMs) to do this!

- Decoder-only architecture enables more efficient learning from all input tokens during pre-training, significantly improving sample efficiency compared to encoder-only models.

How to solve the problem in LLM?



Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In Proceedings of COLM.

Converts decoder-only LLMs into bidirectional encoders via three steps:

- Enabling bidirectional attention
- Masked next-token prediction
- Unsupervised or supervised contrastive learning

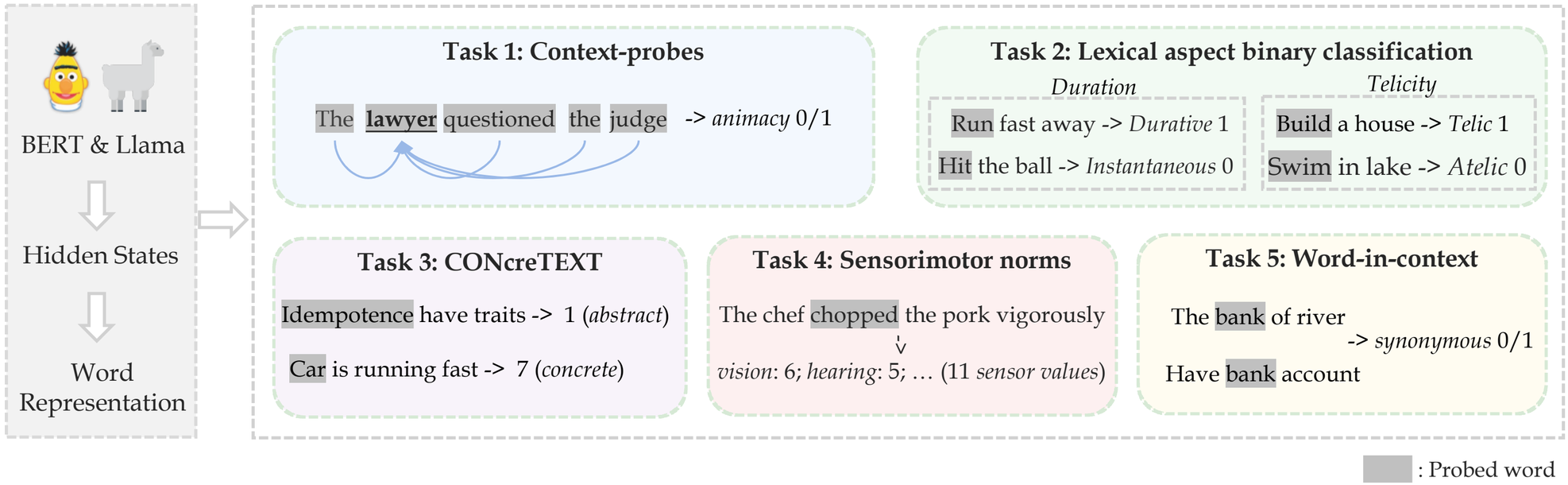
Research Gap

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Research Gap:

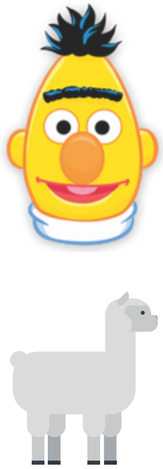
- On word level, LLM2Vec only evaluated out-of-context tasks (e.g., chunking, NER, POS).
- Only reveals the phenomenon (bidirectional attention are beneficial), but fails to explore the underlying causes and potential risks.
- Lacks the comparison with encoder-only models.
- Omits anisotropy analysis, critical for embedding quality assessment.

Our study tries to...

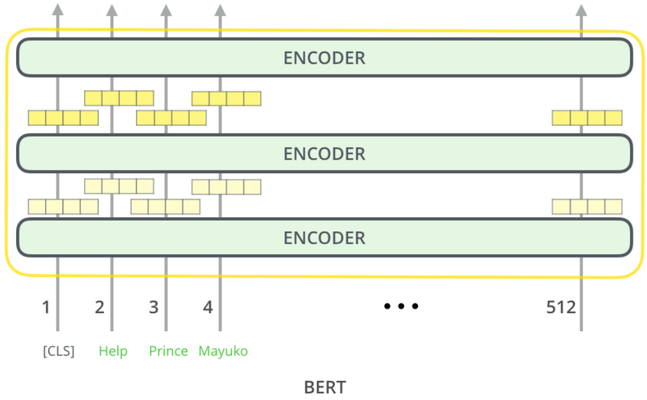


- Focusing on word semantics, using five semantic probing tasks, we examine how bidirectional attention influences Decoder-only models' text embeddings, evaluating its effects on context utilization, anisotropy, and contrastive learning's role on above two effects.

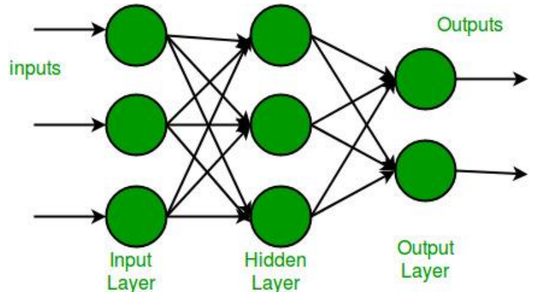
Methodology: probing



BERT & Llama



Use last hidden states as the contextualized embedding



"Probe": multi-layer perception

- Probing Linguistic Features in LLMs
  - the hidden states from the final layer -> contextualized embedding .
  - use a simple diagnostic model ("probe", MLP) to predict specific linguistic properties (e.g. animacy) from the embedding.
  - test Llama's text embedding before and after activating bidirectional attention, campare with BERT on five semantic tasks.

Methodology: five semantic tasks

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Task 1: Context-probes

The **lawyer** questioned the judge -> *animacy* 0/1

The lawyer **questioned** the judge -> *causative* 0/1

dynamic 0/1

Task 2: Lexical aspect binary classification

Duration

Run fast away -> *Durative* 1

Hit the ball -> *Instantaneous* 0

Telicity

Build a house -> *Telic* 1

Swim in lake -> *Atelic* 0

Task 3: CONcreTEXT

Idempotence have traits -> 1 (*abstract*)

Car is running fast -> 7 (*concrete*)

Task 4: Sensorimotor norms

The chef chopped the pork vigorously

↓

vision: 6; *hearing*: 5; … (11 *sensor values*)

Task 5: Word-in-context

The bank of river

Have bank account

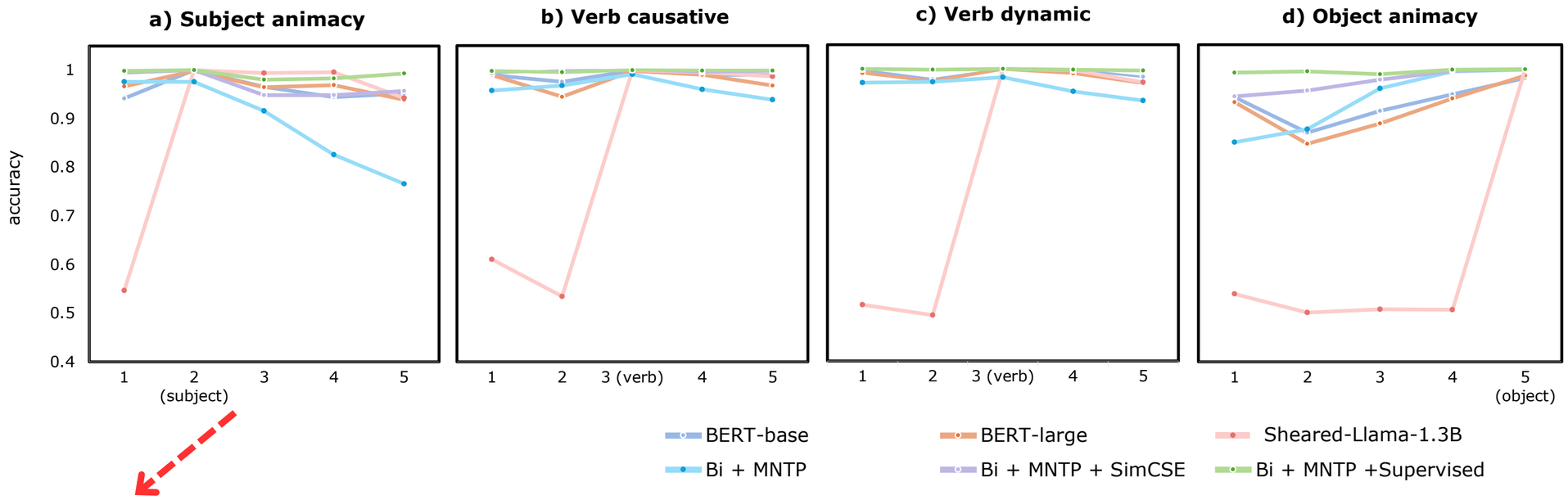
-> *synonymous* 0/1

: Probed word

Task 3 and 4 are regression tasks, the other tasks are all binary classification tasks.

Findings

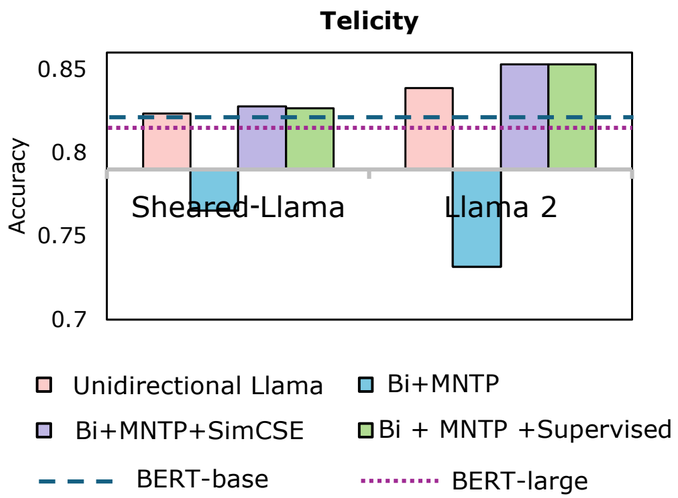
- Finding 1 (Task 1&2)
  - Bidirectional attention improves the LLMs' ability to represent subsequent context, but it also weakens the utilization of the previous context.
  - Contrastive learning techniques mitigate this trade off by enhancing the model's ability to balance contextual understanding in both directions.

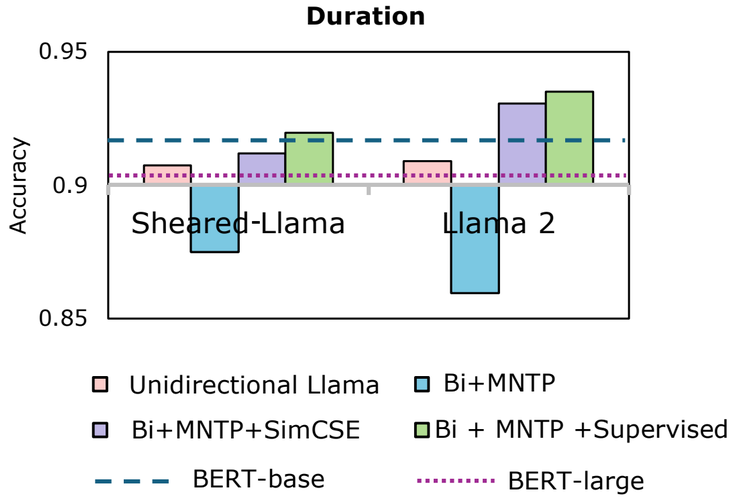


The pink line shows relatively stable accuracy in the latter half, while the blue line exhibits a noticeable decline.

Findings

- Finding 1 (Task 1&2)
  - Bidirectional attention improves the LLMs' ability to represent subsequent context, but it also weakens the utilization of the previous context.
  - Contrastive learning techniques mitigate this trade off by enhancing the model's ability to balance contextual understanding in both directions.



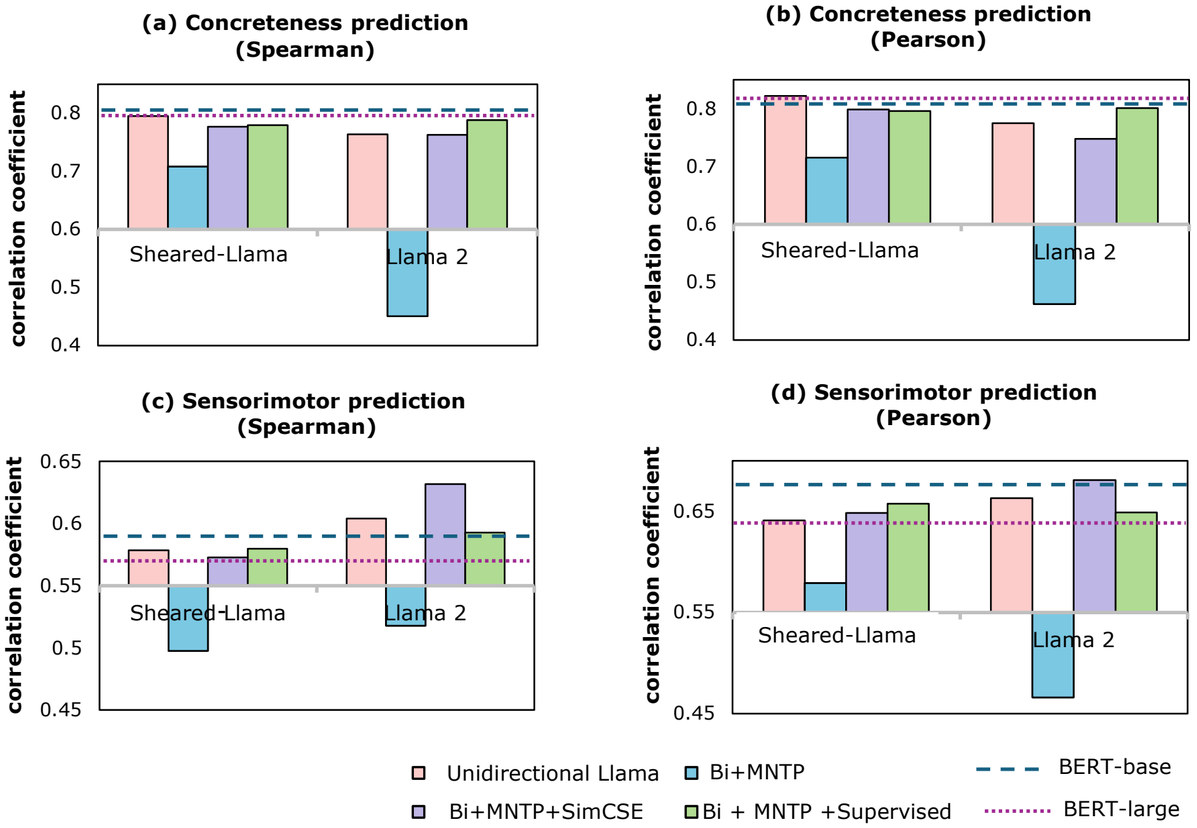


As previously observed:

- bidirectional attention alone reduces model accuracy
- while contrastive learning techniques consistently boost performance.

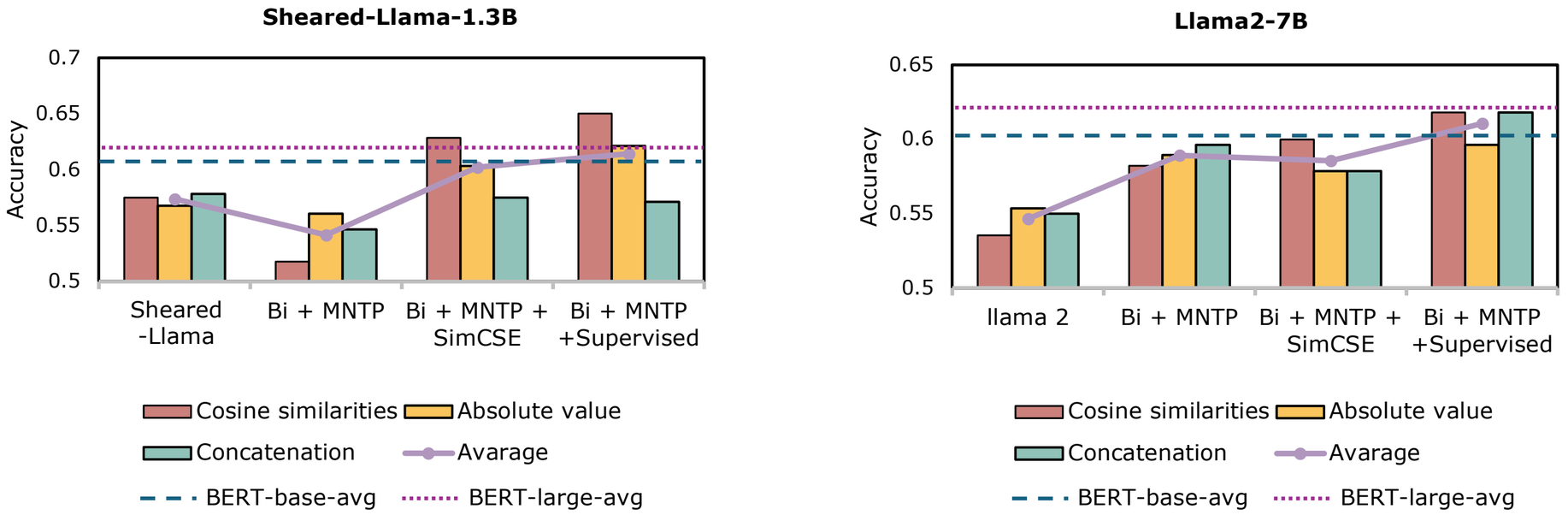
Findings

- Finding 2 (Task 3&4)
  - After enabling bidirectional attention and contrastive learning, decoder-only models can perform similarly or even better to encoder only models on regression probing tasks.



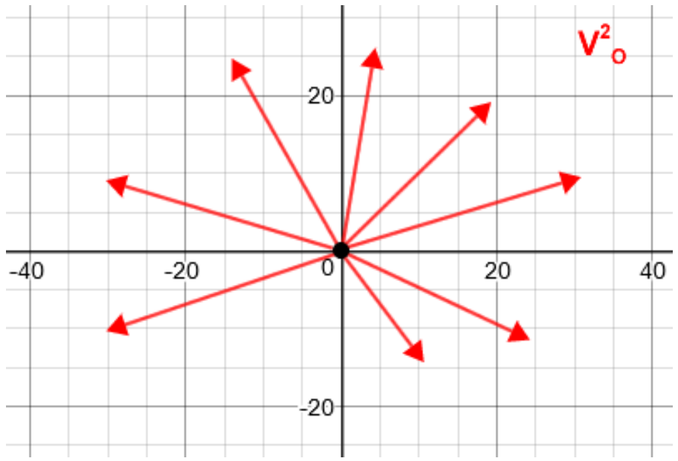
Findings

- Finding 3 (Task 5)
  - In the sense disambiguation task, contrastive learning methods improve the quality of embeddings from decoder-only models irrespective of the strategy for extracting probe features.

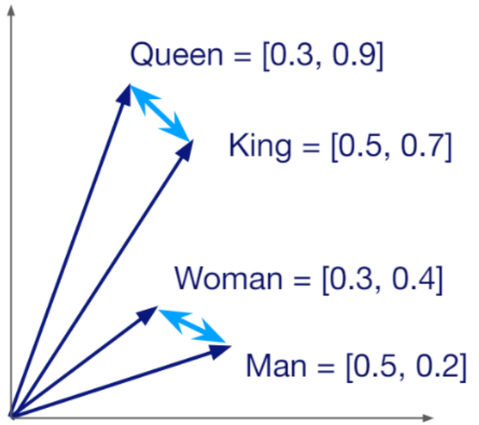


Anisotropy analysis

- Anisotropy issue
  - Despite the advantages of representing context meanings, contextualized embedding were shown to have a high level of anisotropy, i.e. they occupy just a narrow cone in the vector space, with the consequence that randomly-sampled words might also get high similarity values (Ethayarajh, 2019) and postprocessing techniques need to be applied to adjust the similarity metrics for anisotropy (Timkey and van Schijndel, 2021).



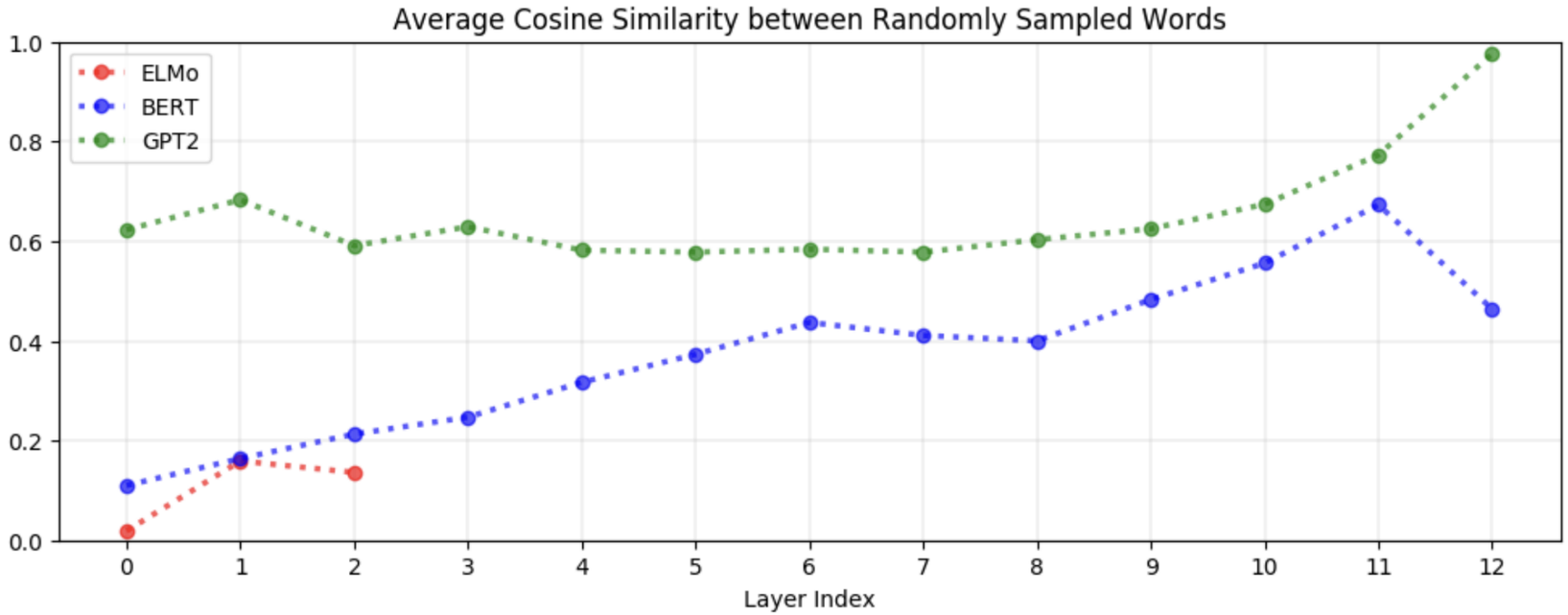
Lower anisotropy



Higher anisotropy

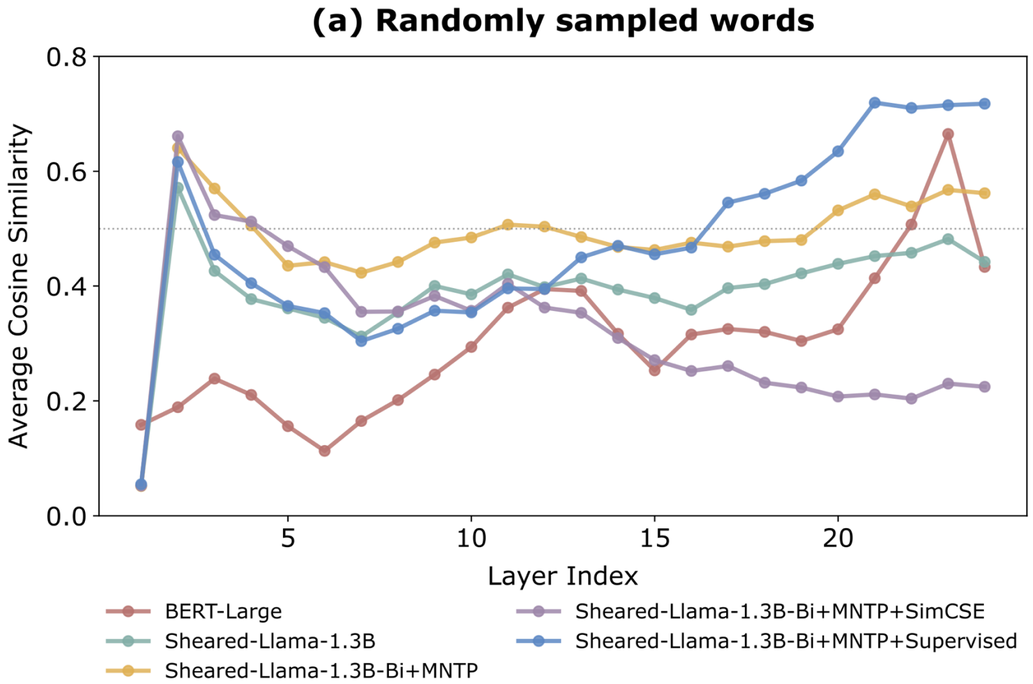
Anisotropy analysis

- Anisotropy issue
  - In the literature (Ethayarajh, 2019), encoder-only models were shown to exhibit lower anisotropy compared to autoregressive decoder-only models, possibly due to the impact of the attention mechanisms.



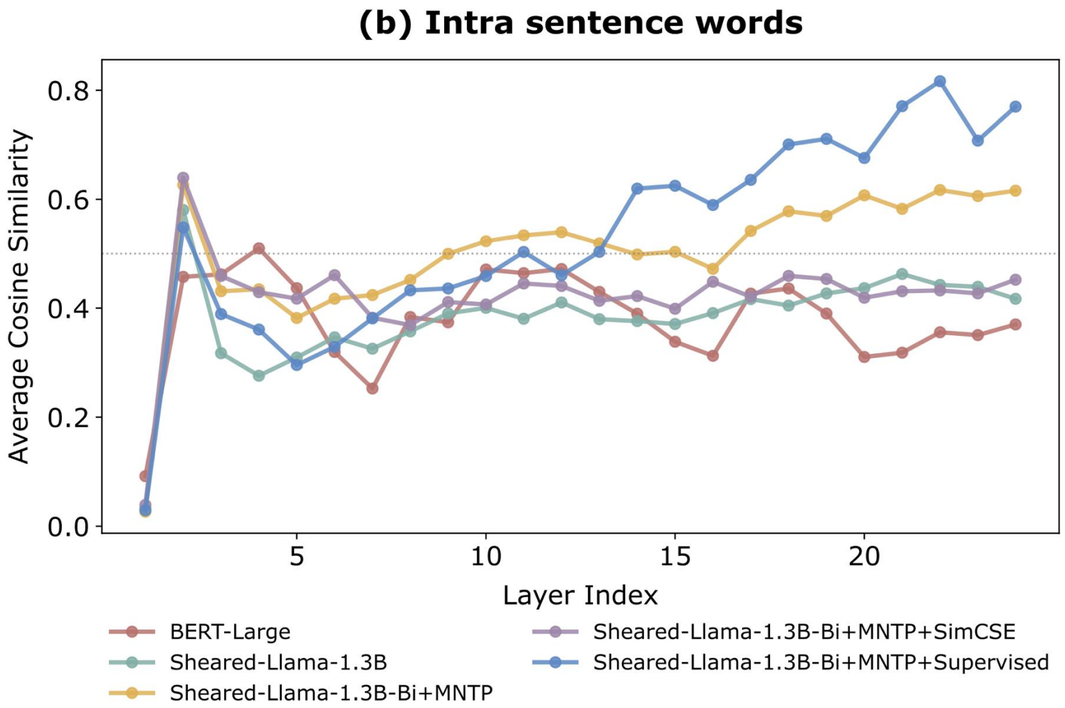
Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 55–65, Hong Kong, China. Association for Computational Linguistics

Anisotropy analysis



- Bidirectional attention increases isotropy level across all Llama layers.
- Among the contrastive learning strategies, the supervised one seems to increase the anisotropy level in the vector space, and the unsupervised one can mitigate the anisotropy issue .

Anisotropy analysis



- To investigate how word embedding within the same sentence evolve from shallow to deep layers of the model, we also extract words from individual sentences and perform layer-wise cosine similarity calculations to quantify intra-sentence anisotropy.
- We find that supervised contrastive learning bidirectional Llama and bidirectional only Llama models perform increasing anisotropy across layers.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

THANKS!

FENG Zhaoxin (Betty)