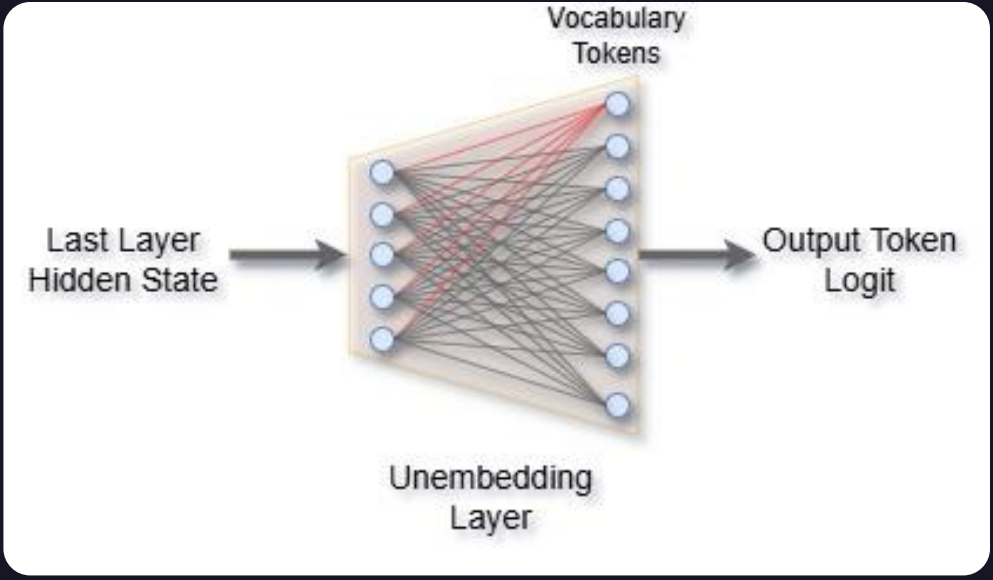
LLM-based Embeddings: Attention Values Encode Sentence Semantics Better Than Hidden States

Yeqin Zhang, Yunfei Wang, Jiaxuan Chen, Ke Qin, Yizheng Zhao, Cam-Tu Nguyen

Motivation

- **Compute dot-product similarity:** Use the last-layer hidden state (HS) and the vectors in the vocabulary unembedding layer to compute similarity.
- **Maximize the true probability:** After applying Softmax, the optimization objective is to maximize the probability of the ground-truth token.



By comparing the negative log-likelihood loss (NLL) of autoregressive models with the InfoNCE loss in contrastive learning, we find that the two have highly similar mathematical forms.

Autoregression (Negative Loglikelihood Loss)

$$\begin{aligned}\mathcal{L}_{\text{NLL}}(x_t, x_{<t}) &= -\log p(x_t \mid x_{<t}) \\ &= -\log \frac{\exp\left((\mathbf{x}_{t-1}^{L})^\top \mathbf{v}_{x_t}/\tau\right)}{\sum_{x\in V}\exp\left((\mathbf{x}_{t-1}^{L})^\top \mathbf{v}_{x}/\tau\right)}\end{aligned}$$

Contrastive Learning (InfoNCE Loss)

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\left(\text{sim}(q, k^+)/\tau\right)}{\sum_{k\in\mathcal{K}}\exp\left(\text{sim}(q, k)/\tau\right)}$$

Representation-space Misalignment.

Findings: Space Misalignment

Directly pooling hidden states inherits a space optimized for token-level discrimination, rather than a space optimized for sentence-level similarity, leading to a clear misalignment between the two.

Solution: Redefine Sentence Semantic

Revisiting the definition of sentence representations: In truth-conditional semantics, the meaning of a sentence is defined as the set of truth values it takes across different possible worlds. Sentence similarity is then defined as the proportion of possible worlds in which two sentences share the same truth value.

Core method: Value vectors and their aggregation strategy (VA)

1. What are value vectors?

At each Transformer layer l and for each token n, the value vectors from all attention heads h are concatenated to form a token embedding vector.

$$v_n^l = [v_n^{l,1}; \ldots; v_n^{l,H}] \in R^d.$$

2. How are value vectors aggregated?

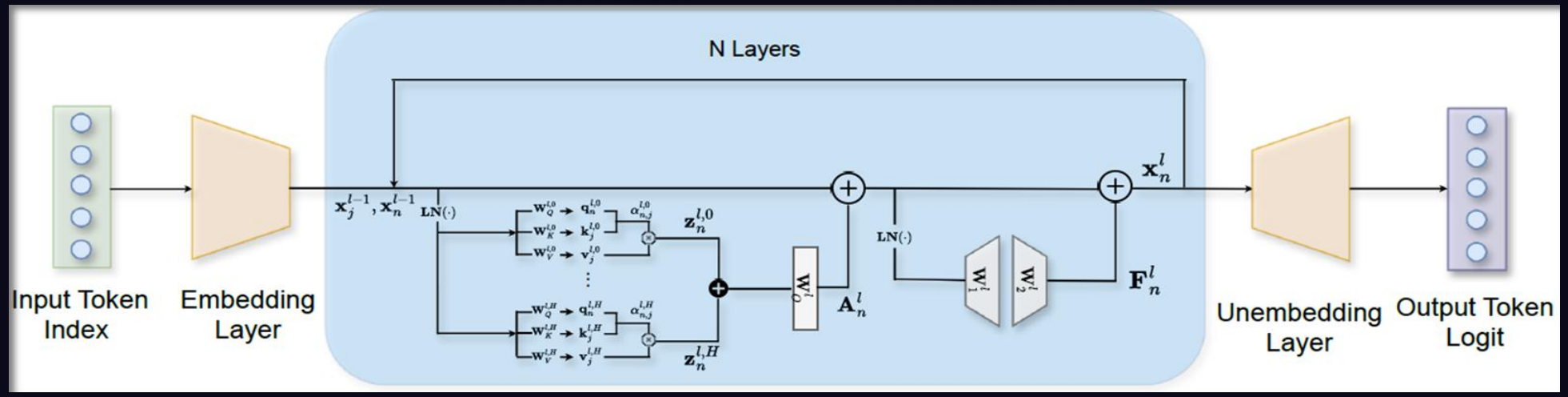
Step 1: Sentence-level average pooling

Average all tokens in each layer to obtain the sentence representation for that layer.

Step 2: Layer-level average pooling

Average over the selected layers S to obtain the final embedding.

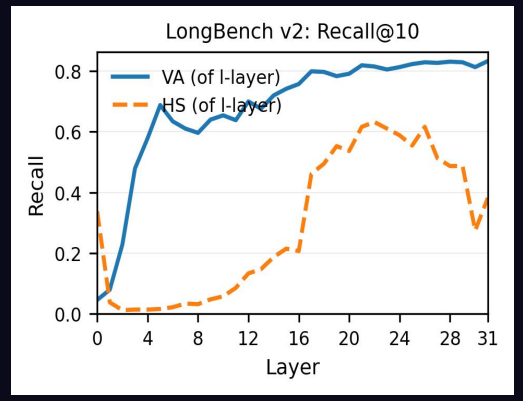
$$\widehat{v^l} = \frac{1}{N}\sum_{n=1}^{N} v_n^l \in R^d \ , \quad V_{\text{agg}}(x_{1:N}) = \frac{1}{|S|}\sum_{l \in S} \widehat{v^l} \in R^d.$$



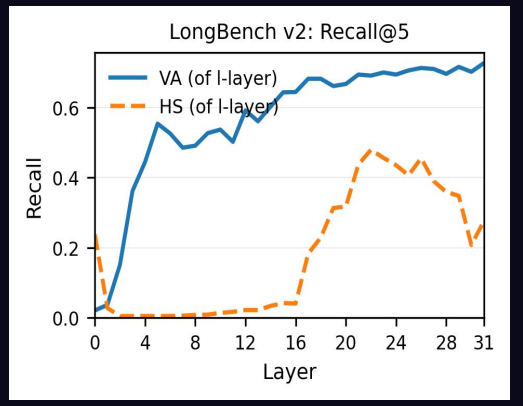
Experimental validation 1: value vectors vs hidden states (HS)

Experimental Tasks and Conclusions

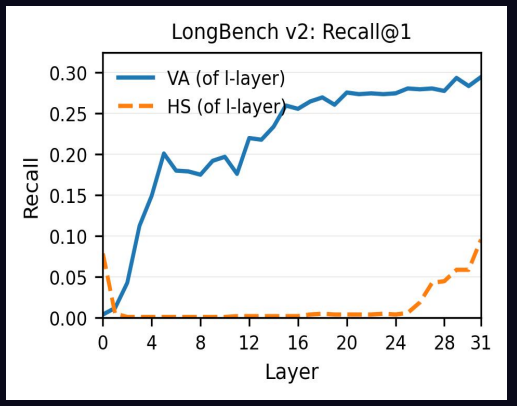
- Task: Compare the performance of different signals on the segment matching task.
- Conclusion: Value vectors (VA) consistently outperform hidden states (HS) across layers.



Recall@10



Recall@5



Recall@1

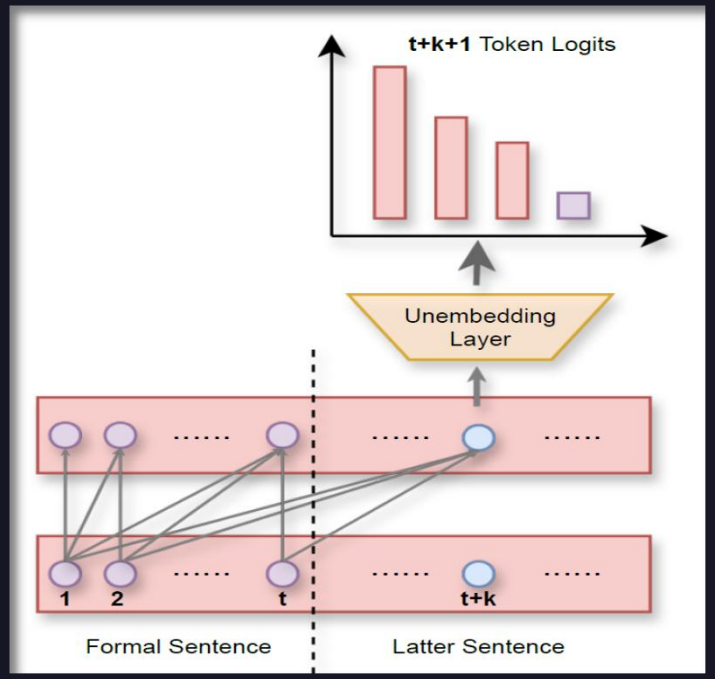
Core Finding: the superiority of value vectors

The figure clearly shows that the blue curve, representing value vectors, almost always lies above the orange curve, representing hidden states. This indicates that, in the segment matching task, value vectors can capture more information related to subsequent content and greatly improve recall.

Experimental validation 2: the predictive ability of value vectors

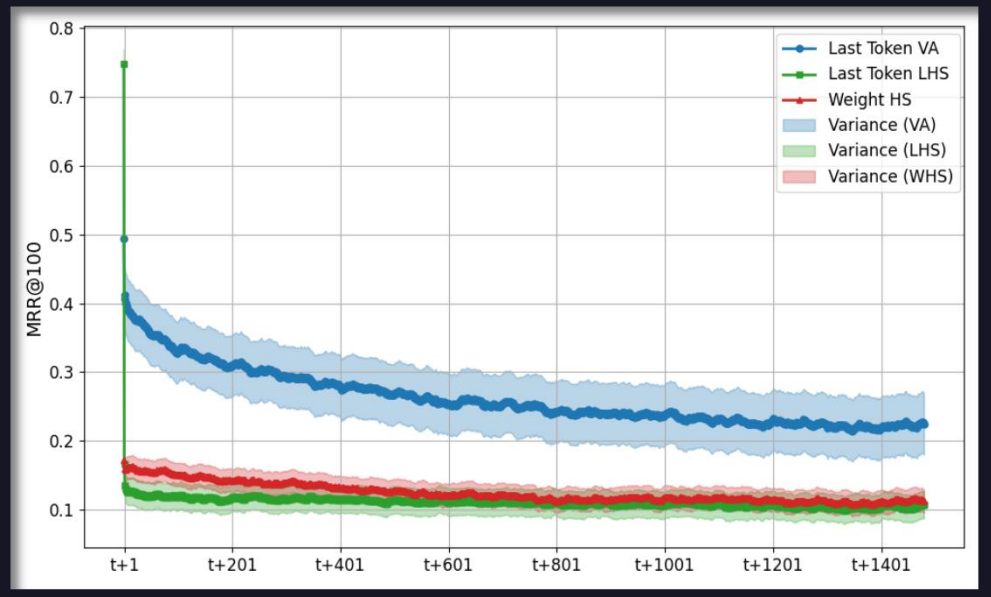
Task Design and Rationale

Using only the first (t) tokens and their attention weights to position (t+k), predict the token at position (t+k+1), thereby verifying whether value vectors help predict subsequent tokens.



Prediction Accuracy Comparison

The experimental results show that value vectors (blue line) achieve much higher prediction accuracy than hidden-layer representations (green and red lines), demonstrating that they contain richer semantic information for guiding the prediction of subsequent tokens.



Overall Performance Evaluation: prompt-free methods and inference cost

Prompt-free method with zero additional inference cost

Value aggregation (VA) is a prompt-free method that does not increase sentence encoding time during inference, maintaining very high computational efficiency.

Compared with MetaEOL: clear improvement in efficiency

MetaEOL requires eight forward passes, with an encoding time approximately eight times that of VA, whereas VA can be completed in a single pass.

The largest gains are observed on retrieval tasks

In the experimental evaluation, the value-vector aggregation method shows the most significant performance improvement on retrieval tasks.

Model	Dim	Backbone	Clustering			Retrieval			STS			Classification			Reranking		Avg.
			Bior.	Medr.	Twen.	SciF.	NFCo.	Argu.	STS17	SICK-R	STSB.	Bank.	Emot.	Spri.	Stac.	SciD.	
Prompt-free Methods																	
HS (Full)	4096	Llama-2	13.69	19.86	5.67	0.10	1.43	44.11	43.90	44.46	11.18	64.24	**39.34**	58.41	32.38	58.69	31.25
HS (Half)	4096	Llama-2	19.86	22.68	6.30	0.23	1.46	47.8	44.21	46.64	12.15	67.58	38.67	61.83	33.04	60.79	32.37
HS	4096	Llama-2	**20.28**	**23.01**	6.04	0.26	1.52	**47.51**	44.17	46.46	12.33	66.46	38.93	**62.89**	33.03	**61.12**	33.14
HS (Full)	4096	Qwen-3	7.42	17.47	5.27	0.17	1.58	40.71	61.05	38.97	23.54	58.95	34.85	46.83	28.41	53.08	29.88
HS (Half)	4096	Qwen-3	11.55	19.57	5.77	0.67	1.75	43.27	62.53	40.94	27.84	60.61	34.69	53.14	29.25	54.98	31.90
HS	4096	Qwen-3	12.87	20.73	5.72	0.71	1.87	42.50	**64.11**	42.54	28.06	61.61	34.62	56.40	31.16	58.10	32.93
LT.	4096	Llama-2	15.99	17.42	**15.96**	2.17	1.31	14.24	57.8	55.63	45.72	**68.65**	29.85	47.01	32.07	58.83	33.05
WMP	4096	Llama-2	19.73	19.47	14.54	**38.89**	**6.13**	33.59	63.91	**57.52**	**58.01**	66.42	30.97	58.48	**37.74**	61.05	**40.46**
LT	4096	Qwen-3	13.56	16.22	13.57	5.06	1.57	8.42	38.56	41.48	28.53	55.82	29.80	9.77	25.25	47.35	23.93
WMP	4096	Qwen-3	9.44	15.82	8.19	5.15	1.41	22.12	51.10	44.17	34.08	60.52	30.01	28.67	32.04	49.74	28.03
Prompt-based Methods																	
EE	4096	Llama-2	22.94	23.15	25.74	25.61	9.97	25.24	80.51	70.18	71.94	81.79	45.00	68.48	40.79	60.15	46.54
PromptEOL	4096	Llama-2	22.49	21.14	31.47	27.16	13.59	11.65	79.67	73.82	75.32	76.37	47.13	26.08	37.65	66.22	43.55
MetaEOL	4096	Llama-2	**30.95**	26.56	**40.03**	40.59	16.41	21.75	**82.29**	**76.88**	**76.87**	**82.26**	51.05	48.24	39.87	**77.91**	50.83
PromptEOL + CP	4096	Llama-2	22.72	21.26	28.98	33.42	17.43	14.57	80.93	72.69	74.73	77.51	47.70	28.42	37.54	66.55	44.60
PromptEOL	4096	Qwen-3	23.43	23.57	27.70	18.27	4.56	12.12	72.84	67.98	67.80	70.06	47.22	36.49	38.91	73.85	41.77
MetaEOL	4096	Qwen-3	29.21	**27.01**	36.74	**47.59**	12.90	**30.81**	80.72	74.24	71.65	81.90	**52.55**	**73.88**	**41.41**	77.53	**52.72**
PromptEOL + CP	4096	Qwen-3	27.66	24.67	34.76	27.35	**18.13**	14.84	74.14	69.38	73.22	71.14	48.66	22.15	40.60	75.90	44.47
Our Methods																	
VA (Full)	4096	Llama-2	31.69	28.30	26.34	51.28	21.00	42.75	74.51	61.47	60.94	73.89	39.58	71.42	41.63	76.16	50.07
VA (Half)	4096	Llama-2	32.45	28.65	27.95	52.41	23.52	44.26	74.08	61.49	61.72	75.19	39.54	73.75	41.51	76.70	50.94
VA	4096	Llama-2	**33.13**	**29.56**	**30.59**	54.58	**25.89**	**45.76**	75.37	61.92	63.03	76.15	**39.85**	75.97	42.08	**77.59**	**52.25**
VA (Full)	1024	Qwen-3	26.85	24.65	20.59	58.37	18.69	41.41	71.70	60.38	60.77	75.72	33.20	81.92	44.72	70.45	49.24
VA (Half)	1024	Qwen-3	26.94	24.29	21.92	58.29	18.80	41.49	71.83	60.47	60.91	75.71	32.91	**82.10**	44.79	70.44	49.35
VA	1024	Qwen-3	31.46	26.61	25.35	**58.93**	21.84	43.11	**76.28**	**62.98**	**63.24**	**76.49**	35.65	81.29	**45.51**	73.48	51.59

Overall Performance Evaluation: the impact of different weighting strategies

Model	Dim	Backbone	Clustering			Retrieval			STS			Classification			Reranking		Avg.
			Bior.	Medr.	Twen.	SciF.	NFCo.	Argu.	STS17	SICK-R	STSB.	Bank.	Emot.	Spri.	Stac.	SciD.	
Baselines																	
PrompEOL	4096	Llama-2	22.49	21.14	31.47	27.16	13.59	11.65	79.67	73.82	**75.32**	76.37	47.13	26.08	37.65	66.22	43.55
VA	4096	Llama-2	**32.45**	**28.65**	27.95	**52.41**	**23.52**	**44.26**	74.08	61.49	61.72	75.19	39.54	73.75	**41.51**	76.70	50.94
MetaEOL	4096	Qwen-3	29.21	27.01	**36.74**	47.59	12.90	30.81	**80.72**	**74.24**	71.65	**81.90**	**52.55**	**73.88**	41.41	**77.53**	**52.72**
Weighted Value Aggregation																	
WVA (LT)	4096	Llama-2	15.83	16.33	11.48	19.75	3.04	16.39	63.58	57.01	58.69	62.28	29.31	13.68	28.78	55.62	32.27
WVA (PromptEOL)	4096	Llama-2	27.40	22.14	30.67	45.32	**24.22**	31.27	85.02	**73.98**	**77.11**	81.15	53.66	57.21	**43.07**	75.44	51.98
WVA (FutureEOL)	4096	Llama-2	**30.46**	**25.21**	**35.04**	**51.80**	21.54	**37.95**	**85.04**	72.60	74.64	**81.54**	**53.86**	**75.83**	42.86	**77.22**	**54.69**
WVA (LT)	1024	Qwen-3	22.99	21.31	29.38	35.89	10.74	14.03	56.37	60.07	54.75	68.92	32.90	20.10	30.25	64.44	37.30
WVA (PromptEOL)	1024	Qwen-3	21.82	21.93	27.22	17.87	13.42	11.16	73.86	67.80	70.73	70.20	49.03	16.87	38.42	69.60	40.71
WVA (FutureEOL)	1024	Qwen-3	25.38	22.34	24.33	24.42	10.69	16.02	75.63	62.73	68.23	74.21	51.75	28.28	38.65	63.61	41.88
Aligned Weighted Value Aggregation																	
AlignedWVA (PromptEOL)	4096	Llama-2	28.82	23.45	31.20	45.31	25.13	32.44	83.33	73.80	76.74	82.09	52.68	60.93	43.52	75.76	52.51
AlignedWVA (FutureEOL)	4096	Llama-2	31.25	26.39	33.82	51.40	25.32	39.67	83.38	71.54	73.36	**82.62**	52.78	**78.42**	43.39	76.44	54.98
AlignedWVA (PromptEOL)	4096	Qwen-3	**35.56**	**29.73**	**48.04**	55.44	31.55	34.53	83.06	**76.14**	**77.48**	81.52	**54.18**	59.54	**45.88**	**83.01**	56.83
AlignedWVA (FutureEOL)	4096	Qwen-3	33.62	28.03	44.51	**62.98**	**32.04**	**42.88**	**84.31**	68.55	75.53	82.06	51.75	69.50	43.26	76.79	**56.84**

Simple Weighting

Directly use the attention weights generated by the last token for aggregation as the basic comparison baseline.

WVA (FutureEOL)

In terms of average performance, it outperforms the strong MetaEOL baseline by about 2 points, showing the effectiveness of the WVA strategy.

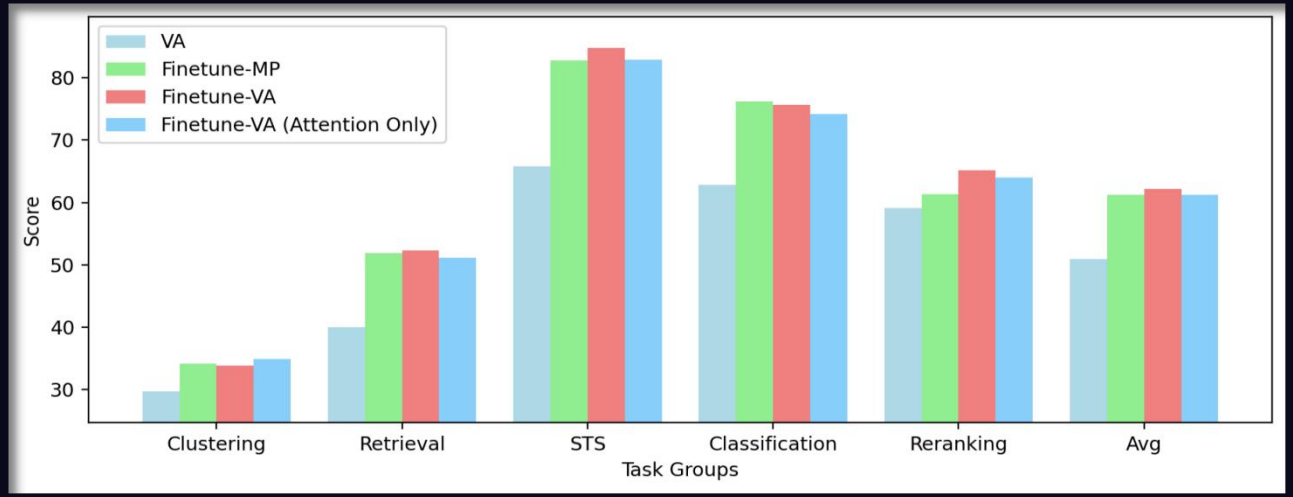
Aligned WVA (FutureEOL) – Best Strategy

By using a prompt to explicitly guide the model to focus on predicting the next token, the overall performance is further improved, highlighting the key role of the prompt.

Overall Performance Evaluation: comprehensive performance comparison

Figure Interpretation

Finetune-VA improves VA embeddings and achieves consistent gains over Finetune-MP. These improvements are most notable on reranking and STS tasks, with gains of 3.8 and 2.02 points, respectively. Finetune-VA (Attention Only) achieves nearly the same effect while training only one quarter of the parameters, showing the advantage of finetuning VA.



Conclusion

After fine-tuning, although the training objective is defined on the final-layer output, VA performs about 2 points lower than last-layer hidden-state pooling. Under grouped-query attention, the value projection is shared across four query groups. As a result, VA produces a 1024-dimensional embedding, whereas last-layer hidden-state pooling produces a 4096-dimensional embedding. Although the dimensionality is reduced by three quarters, VA is still only about 2 points behind last-layer hidden-state pooling on the same 8B model, while outperforming the 1024-dimensional last-layer hidden-state embedding in Qwen3-0.6B-Embedding by about 2 points.

Model	Dim	Backbone	Clustering			Retrieval			STS			Classification			Reranking		Avg.
			Bior.	Medr.	Twen.	SciF.	NFCo.	Argu.	STS17	SICK-R	STSB.	Bank.	Emot.	Spri.	Stac.	SciD.	
VA	4096	Llama-2	32.45	28.65	27.95	52.41	23.52	44.26	74.08	61.49	61.72	75.19	39.54	73.75	41.51	76.70	50.94
Finetuning Methods																	
Finetune-MP	4096	Llama-2	34.41	31.19	36.69	71.47	35.56	48.81	88.19	77.71	82.83	85.77	50.55	92.28	47.87	74.60	61.28
Finetune-VA	4096	Llama-2	33.30	29.97	38.26	71.86	38.21	46.88	89.76	79.97	84.62	82.92	49.32	94.67	49.29	80.98	62.14
Finetune-VA (Attention Only)	4096	Llama-2	34.12	31.02	39.64	70.53	31.12	51.75	88.57	77.63	82.56	85.37	47.24	90.07	48.20	79.70	61.25
LLM-based Pretrained Embedding Models																	
LLM2Vec	4096	Llama-2	34.81	31.37	51.04	77.30	40.33	56.53	90.63	83.01	88.72	88.17	51.71	96.83	51.02	84.03	66.11
VA (Llm2vec)	4096	Llama-2	33.87	30.73	47.82	74.95	29.51	56.98	88.48	81.12	87.07	83.32	52.75	97.07	52.31	83.02	64.21
Qwen3-Embedding-0.6B	1024	Qwen-3	39.98	36.76	51.20	70.65	36.06	70.65	93.32	84.69	91.23	85.19	59.77	**97.48**	54.22	86.78	68.43
VA (Qwen3-Embedding-0.6B)	1024	Qwen-3	39.23	36.05	49.84	68.21	32.74	67.02	89.67	80.17	87.14	84.34	59.46	96.50	54.17	85.90	66.46
Qwen3-Embedding-8B	4096	Qwen-3	**44.57**	**40.56**	**63.40**	**78.61**	**41.39**	**76.30**	**95.82**	**88.46**	**93.59**	**89.71**	62.82	97.46	**58.86**	**89.96**	**72.97**
VA (Qwen3-Embedding-8B)	1024	Qwen-3	44.29	39.97	60.40	75.16	34.48	72.71	94.15	83.94	91.00	89.40	**63.93**	96.06	58.19	88.76	70.89


THANK YOU