Learning to Rank and Classify Romanian Text: Embedding Strategies and Neural Architectures

Chiruș Mina-Sebastian, Bivol Mihai, coordinator: Florin Brad*, Marius Ninel Cătălin*
University of Bucharest, Romania
*Bitdefender, Romania

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

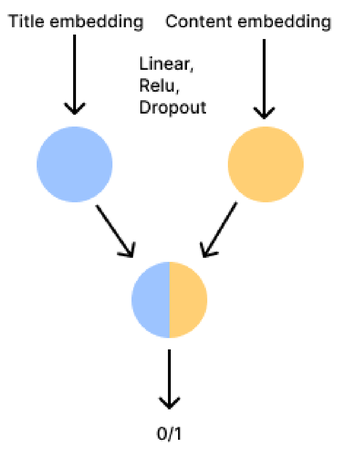
1. Introduction

As large language models continue to rise, observing how well they understand semantics and how do they reason remains an open challenge. In this paper, we evaluate and compare several Romanian-capable LLMs by analyzing their text embeddings on two distinct tasks: **sentiment understanding** using LAROSeDa [5] dataset and **logical reasoning** on RoARC Challenge [3].
We extract embeddings from **Llmic 3B [1], mGPT 1.3B [4], Romanian Llama2 7B [3], Llama 3.1 8B [2] , Gemma 3 1B [6]** using 3 pooling strategies, then train lightweight neural networks to perform classification or ranking. For RoARC we apply different embedding strategies to get the most out of the representations. We experimented various models and chose the best results for this paper.

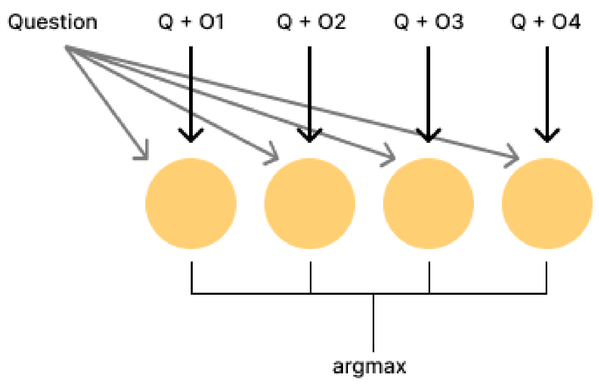
2. Dataset

We use two Romanian-language datasets: LAROSeDa, focused on sentiment analysis, with 12,000 training and 3,000 test samples. RoARC, a Romanian adaptation of the ARC Challenge benchmark, with 1,928 training and 644 test question-answer pairs.
In total, we generated **642,900 text embeddings**, occupying approximately **5.64 GB**.

3. Our Models



We design for **laroseda** a dual-branch MLP to process title and content embeddings separately. Each branch maps its input to a shared hidden space. The outputs are concatenated and passed through a final classifier for binary prediction. This structure helps capture distinct semantic roles of title and content.



We design for **RoARC** a single-layer scoring model that computes a score for each concatenated pair of question and answer option embeddings ($question \parallel option_i$). We then select the answer with the highest score using arg max. For a given question with a set of answer options, let s_y be the score assigned to the correct answer y, and s_j the score of an incorrect answer j. The loss for a single example is defined as:

$$\ell = \sum_{j \neq y} \max\left(0,\ 0.5 - (s_y - s_j)\right)$$

References

[1] Bădoiu et al., LLMic, 2025.
[2] Grattafiori et al., LLaMA 3 Herd, 2024.
[3] Masala et al., Vorbești Românește?, 2024.
[4] Shliazhko et al., mGPT, 2022.
[5] Tache et al., "LaRoSeDa Embedding Clustering", in: *arXiv* (2021).
[6] Gemma Team, "Gemma 3", in: *Kaggle* (2025).

4. Results

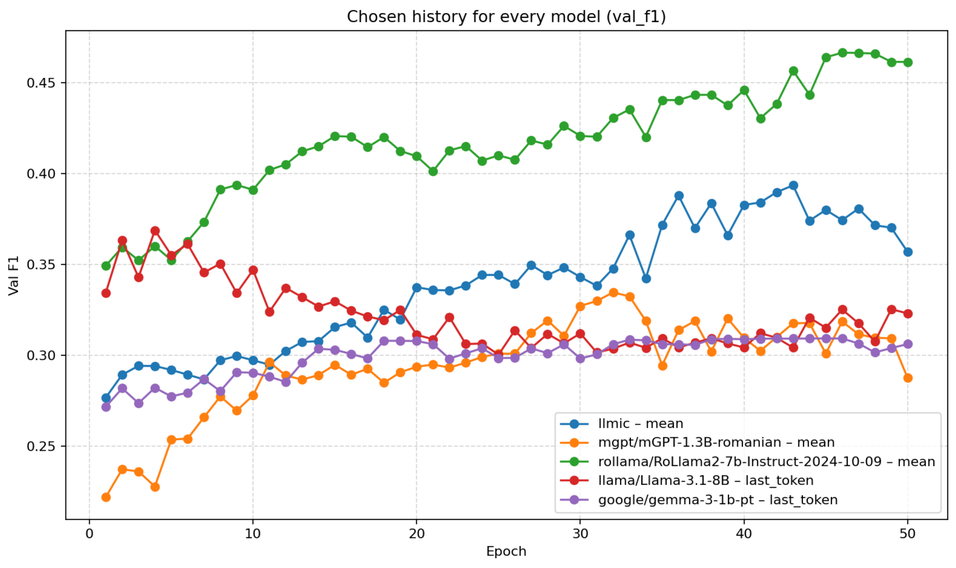
Table 1: Performance across pooling strategies for LAROSeDa

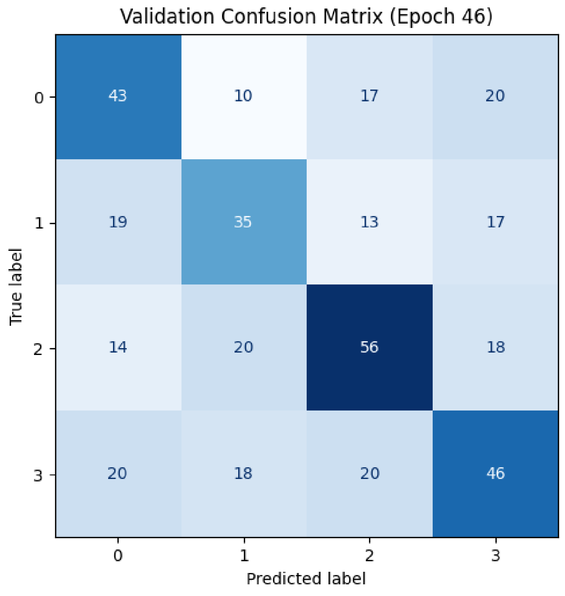
Model	Prc-0	Prc-1	Prc-2	Rec-0	Rec-1	Rec-2	F1-0	F1-1	F1-2
Llmic 3B [1]	0.91	0.0	0.97	0.98	0.0	0.96	0.95	0.0	**0.97**
mGPT 1.3B [4]	0.99	0.96	0.96	0.90	0.96	0.96	0.94	0.96	0.96
Ro Llama2 7B [3]	0.97	0.95	0.94	0.98	0.99	0.98	**0.97**	0.97	0.96
Llama 3.1 8B [2]	0.98	0.95	0.96	0.97	0.98	0.96	0.97	**0.97**	0.96
Gemma 3 1B [6]	0.97	0.89	0.96	0.88	0.96	0.96	0.92	0.92	0.96

Table 2: Performance across pooling strategies for RoARC

Model	Prc-0	Prc-1	Prc-2	Rec-0	Rec-1	Rec-2	F1-0	F1-1	F1-2
Llmic 3B [1]	0.37	0.32	0.33	0.37	0.32	0.32	0.37	0.32	**0.32**
mGPT 1.3B [4]	0.32	0.31	0.27	0.32	0.30	0.27	0.32	0.31	0.27
Ro Llama2 7B [3]	0.41	0.45	0.24	0.40	0.44	0.24	**0.40**	**0.44**	0.24
Llama 3.1 8B [2]	0.34	0.35	0.25	0.34	0.34	0.25	0.33	0.34	0.25
Gemma 3 1B [6]	0.23	0.23	0.21	0.23	0.23	0.21	0.23	0.23	0.21

Note: 0 = mean pooling, 1 = last token, 2 = echo token.

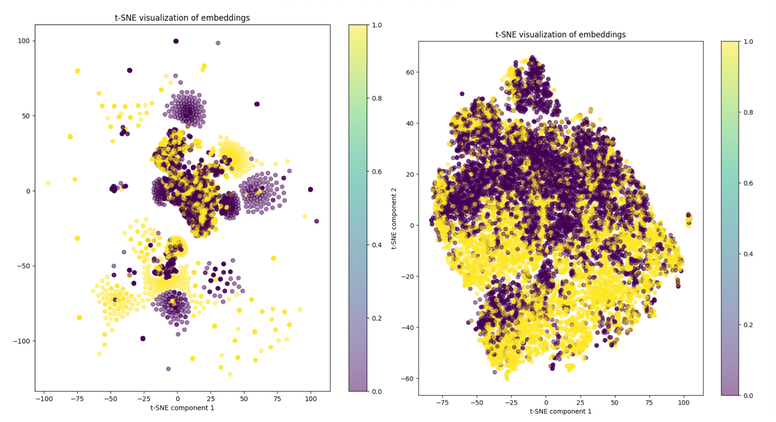




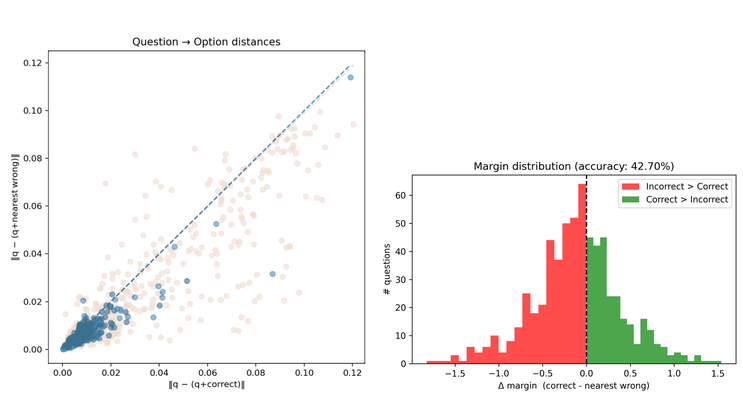
5. Qualitative Analysis

The left visualizations show t-SNE plots of title and content embeddings. Titles form distinct clusters, indicating strong sentiment cues. Content embeddings are more scattered, suggesting deeper context is needed.
We compare two strategies: (1) independent embeddings, emb($option_i$), and (2) combined, emb($question + option_i$). The scatter plot (right) contrasts the distance from the query to the correct vs. nearest wrong option. Most blue points below the diagonal show the model ranks correct answers closer.
The margin histogram (far right) shows that `RoLLaMA2-7b-Instruct` with last token pooling achieves **42.70% accuracy**. Green bars = correct; red = incorrect.

LAROSeDa



RoARC



6. Conclusions

Our findings highlight that even lightweight neural models, when paired with Romanian LLM embeddings, can capture sentiments effectively, but they still have some problems with reasoning. Title embeddings proved especially informative, while contrastive losses enhanced performance in reasoning tasks.