Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Anton Razzhigaev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Crossroads of text, vision, and graph models: advancing multimodal systems

Supervisor: Associate Professor Alexander Panchenko
Co-supervisor: Dr. Andrey Kuznetsov

Name of the Reviewer: Aleksandr Petiushko

I confirm the absence of any conflict of interest	**Date: 17-09-2024**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications
- The summary of issues to be addressed before/during the thesis defense

Dr. Aleksandr Petiushko

Head of AI Research, Gatik AI Inc.

161 E Evelyn Ave

Mountain View, CA 94041

USA

Jury Member Report

The thesis presented by Anton Razzhigaev tackles the classical domain of Natural Language Processing (NLP), which has recently experienced significant advancements with the emergence of large language modeling (LLM) approaches and multi-modal capabilities. Through an exhaustive investigation encompassing multiple subfields, Anton has achieved remarkable results in several areas. Specifically, his theoretical analysis of transformer embeddings - a crucial component of LLMs - has yielded insightful findings on their anisotropy, dimensionality, and linearity. Building upon this foundational work, Anton has also explored the application of structural knowledge graph representations to multi-lingual question-answering systems, demonstrating impressive performance. Furthermore, he has made significant strides in multi-modality research, starting with the use of byte-pair encoding (BPE) for images and progressing to the fusion of text and image through a diffusion process (Kandinsky). Additionally, Anton has incorporated audio modality into his work using OmniFusion, culminating in the development of an evaluation benchmark - OmniDialog - for assessing multi-modality dialogue systems.

The breadth and depth of Anton's research are truly impressive, with few areas left unexplored. As a testament to the value of his ideas, three of his articles have been published in top-tier conferences (Core A*), while another has appeared in a leading conference (Core A). In conclusion, the work conducted by Anton represents the current state-of-the-art in theoretical analysis, architectural improvement, and practical assessment of NLP models.

Notwithstanding this significant progress, several open questions remain that may be addressed either during the formal defense of this work or, more likely, in future research endeavors.

Chapter 3:

- What is the correlation between anisotropy and intrinsic dimensionality (ID)? To what extent are they correlated?

- Would be interesting to see more analysis for encoders (e.g., for ID, and train time dependency), and ID dependency on the layer depth for decoders

Chapter 4:

- Why does the introduction of the contracting loss actually lead to reducing the dimensionality? Although it has been noted by Anton, it still deserves the deeper analysis.

- I guess only consecutive layers are taken into account while studying the linearity. Would be interesting to analyze the linearity for any (i,j) pair of layers

Chapter 5:

- Would love to hear some reasoning on the design choice for the model architecture (Figure 5-1). Why to use averaging and not something like Sentence-BERT or other aggregation function? Why are cosine similarities summed and not, for example, multiplied?

Chapter 6:

- While justifying for BPE vs pixel-level tokenization, the comparison can also consider the discretization via DVAE

- Some additional text would be helpful to describe the dark vertical bands in Figure 6-3.

Chapter 7:

- A little bit unclear description of "Image Prior" and its comparison to "Diffusion Mapping" - see Figure 7-1. I would suggest to add a diagram of its architecture.

Chapter 8:

- What is the reasoning behind transforming the embedding exactly into k vectors? Not 1? What is the ablation on this k?

In addition, there are numerous minor typographical errors scattered throughout the document, given the massive scope and complexity of the work. E.g., sometimes there are unnecessary commas (page 16, 26), incorrect spelling (page 17, "verifieble", page 109 - both “VLM” and “VLLM” abbreviations are used), wrong numbers (Figure 8-2: we have the same token for IMG_START and AUDIO_START, but different for IMG_END and AUDIO_END), wrong image captions (Figure 8-3: “OmniLLaMa” while in the thesis text we have “OmniMistral”).

Finally, it would be beneficial to devote more attention to properly citing relevant prior work and literature: e.g., missing word2vec reference in the section 2.1, VAE / VQ-VAE usage for generative modeling in the section 2.2, vision-language-action models in the section 2.4. And sometimes in the Bibliography list there are items with missing years or other important attributes.

Despite the open questions and minor suggestions for improvement, the work presented by Anton is of exceptional quality, offering valuable insights and contributions to both academia and industry. It richly deserves a definitive 'DEFENDED' rating.

Aleksandr Petiushko

September 17th, 2024

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*
☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*