

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Суворов Антон Дмитриевич
Должность: Ректор
Дата подписания: 31.08.2026 16:35:32
Уникальный программный ключ:
a39bdb15d680d3b0adbfced0af5c1efb14747dc0

Skoltech

Course Syllabus

Course Title

Modern Methods and Algorithms of Generative Artificial Intelligence

Course Title (in Russian)

Современные методы и алгоритмы генеративного искусственного интеллекта

Lead Instructor

Andrey Kuznetsov

Co-Instructor

First Name	Last Name
Konstantin	Sobolev
Elizaveta	Goncharova
Vladislav	Shakhuro
Irina	Abdullaeva

1. Annotation

Course Description

In this course we will discuss modern approaches and challenges in generative AI. First, we go deep in foundational models and multimodality, which includes large language models as foundation knowledge representation and different approaches for modalities fusion. Further we will dive in multimodal architectures like Fromage, LLaVA, OmniFusion, Qwen-VL, Qwen-Omni, DeepSeek, etc. to provide a full understanding of how SoTA models operate with multimodal data simultaneously. With regard to different approaches we will discuss transformer-based architectures with adapters, end-to-end multimodal transformers, diffusion transformers and different ways of extending the capabilities of data processing with RL, reasoning, evolution and memory techniques. When speaking about modalities we mean images, videos, audio as major data non-text sources nowadays. Then we proceed to text-to-X generative models for images, videos and 3D generation and editing. This will include transformer (DALL-E, eDiff-I) and diffusion models (Imagen, Stable Diffusion, Kandinsky, etc.), as well as state-of-the-art diffusion transformer architectures (Wan, Hailuo, etc.). We will also go deeper in what is called world models to show the applicability of current models in real life scenarios simulation. Such applications as robotics, self-driving, and generative design will be discussed additionally. Overall you get comprehensive knowledge about current state in GenAI and be ready to go deeper in any of the studied generative AI fields.

Course Description (in Russian)

В этом курсе мы обсудим современные подходы и проблемы в области генеративного ИИ. Сначала мы углубимся в фундаментальные модели и мультимодальность, которая включает в себя большие языковые модели как базовые концентраторы знаний и различные подходы к смешиванию модальностей. Далее мы углубимся в мультимодальные архитектуры, такие как Fromage, LLaVA, OmniFusion, Qwen-VL, Qwen-Omni, DeepSeek и т. д., чтобы сформировать полную картину того, как SoTA модели одновременно работают с разными типами данными. В отношении различных подходов мы обсудим архитектуры на основе трансформеров с адаптерами, end-to-end мультимодальные трансформеры, диффузионные трансформеры и различные способы

расширения возможностей обработки данных с помощью RL, рассуждений, эволюции и памяти. Говоря о модальностях, мы подразумеваем изображения, видео, аудио как основные нетекстовые источники данных в настоящее время. Затем мы перейдем к генеративным text-to-X моделям для генерации и редактирования изображений, видео и 3D. В курсе обсудим трансформерные модели (DALL-E, eDiff-I) и диффузионные модели (Imagen, Kandinsky и др.), а также современные архитектуры диффузионных трансформеров (Wan, Hailuo, Kosmos и др.). Мы также углубимся в так называемые модели мира (WorldDreamer, Genie 3, etc.), чтобы продемонстрировать возможности и ограничения современных моделей при моделировании реальных физических процессов. Дополнительно будут рассмотрены такие области, как робототехника, беспилотное вождение и генеративное проектирование. В целом, вы получите исчерпывающие знания о текущем состоянии GenAI и будете готовы к дальнейшему развитию любой из изучаемых областей генеративного ИИ.

2. Basic Information

Course Academic Level

MSc

PhD

Number of ECTS credits

3

Course Prerequisites / Recommendations

Students should be familiar with transformer architecture. Diffusion for data generation would be beneficial.

Type of Assessment

Graded

Mapping from grades to percentage:

A:	86
B:	76
C:	66
D:	56
E:	46
F:	0

Term

Term 5

Students of Which Programs do You Recommend to Consider this Course as an Elective?

BSc Programs	Masters Programs	PhD Programs
	Advanced Computational Science Data Science	Computational and Data Science and Engineering

Maximum Number of Students

	Maximum Number of Students
Overall:	75
Per Group (for seminars and labs):	15

3. Course Content

Topic	Summary of Topic	Contact Hours: Lectures	Contact Hours: Seminars	Contact Hours: Labs	Non-contact Hours: Student's Independent Study
1. Generative AI	Intro to generative AI. Existing fields of study. Research trends	1.5	1.5	0	6
3. Foundation models	Foundation model definition. Language model as a foundation model. Tasks and challenges.	1.5	1.5	0	6
4. Multimodality . Perception. Part I	Types of fusion: early, middle and late. Visual language models. Types of image and video encoders. Fusion of encoders and LLMs. Overview of transformer-based approaches and state-of-the-art architectures. Tasks, challenges and benchmarks.	1.5	1.5	0	6
5. Multimodality . Perception. Part II	Omni models to operate with several modalities simultaneously for perception and generation. Comparison with separate perception and generation approaches. GRPO, reasoning and memory techniques for quality improvement. Fine-tuning approaches.	1.5	1.5	0	6
6. Long context analysis	Ways of video perception: from short videos to streaming understanding. Approaches for frame encoding and selection, training-free approaches and benchmarking of long-context perception tasks.	1.5	1.5	0	6
7. Domain-specific applications	Multimodality in robotics, self-driving cars, medicine and other fields of research and industry.	1.5	0	0	3
8. Image generation and editing	Transformer-based image generation models in general and specific-domains. Image personalization and editing approaches. Fine-tuning mechanics.	1.5	1.5	0	6
9. Video generation and editing	Diffusion architectures for content generation. Existing SoTA models and approaches. Diffusion transformers, flow matching techniques for video generation and editing.	1.5	1.5	0	6
2. Large language models and applications	Transformers and attention mechanism intro. Existing approaches and solutions. Mechanical interpretation techniques in terms of LLM explainability. Vanilla transformers, retrieval-augmented generation, mixture-of-experts approach.	1.5	1.5	0	6

Topic	Summary of Topic	Contact Hours: Lectures	Contact Hours: Seminars	Contact Hours: Labs	Non-contact Hours: Student's Independent Study
10. Generative Design	Task description of domain-specific data perception and generation in construction, architecture and industry. AI-assistants, multimodal models for point clouds, images and texts for CAD generation. Text2BIM models for building information modeling data generation.	1.5	0	0	3

4. Learning Outcomes

Skoltech Learning Outcomes are indicated as per [Skoltech Learning Outcomes Framework](#).

1. FUNDAMENTAL KNOWLEDGE

1.2. KNOWLEDGE OF APPLIED SCIENCE AND ENGINEERING SCIENCE INCLUDING CONTEMPORARY METHODS AND TOOLS

2.1. COGNITION AND MODES OF REASONING

2.1.1. Analytical reasoning and problem solving

2.1.2. System thinking

2.1.3. Creative thinking

3.1. COMMUNICATIONS IN INTERNATIONAL ENVIRONMENTS

3.1.3. Oral presentation and discussion

3.1.4. Inquiry, listening and dialogue

3.1.5. Communications in English in scientific, business and social settings

3.2. TEAMWORK AND LEADERSHIP

3.2.1. Forming effective teams

4.2. VISIONING – INVENTING NEW TECHNOLOGIES THROUGH RESEARCH

4.2.1. The research process – hypothesis, evidence and defense

4.2.3. Research aimed at developing new technologies

4.2.4. Imagining utility of new science and technology

5. Assignments and Grading

Physical Attendance Requirement 80
(% of classes)

In-person Attendance Requirement Comment

A medical report or taking part in international conferences or contests may be a reason to miss a class

Assignment Type	Assignment Summary	% of Final Course Grade
Homework Assignments	Homework 1 — A task for multimodal model fine tuning in perception task. Homework 2 — A task on text-to-image model LoRA-based fine tuning.	50
Final Project	During the course a student should provide a presentation and report about the project on generative AI according to the list of course topics: optimization, specific domain fine-tuning, architectural updates for quality improvement	50

6. Assessment Criteria

Assignment 1 Type

Homework Assignments

Sample of Assignment 1

Given a pretrained visual language model develop its suitable fine tuned version to improve the model's skills in handwritten text recognition.

Assessment Criteria for Assignment 1

- 1) a student should provide a full pipeline of the model's development process beginning with dataset collection and processing, following with the fine tuning method for the task and finally providing the chosen metric value
- 2) the source code of the model should not be a simple fork of an open source solution, but should provide a separate GitHub repo, organized in a well-structured form
- 3) the assignment should be accompanied with a presentation of the method and all the details of its implementation

Assignment 2 Type

Final Project

Sample of Assignment 2

Design a project proposal on using multimodal models in remote sensing data analysis

Assessment Criteria for Assignment 2

- 1) the idea for implementation of the project should be realistic and easy to follow
- 2) SoTA architectures should be used as baselines and the choice should be explained
- 3) training data should be described or there should be an instructing for data collection
- 4) implemented PoC of the project on github will be a benefit
- 5) the project proposal should include description of all the research stages for implementation (a document and/or presentation)
- 6) the team for project development may include from 1 to 3 members

7. Textbooks and Internet Resources

You can request at most two required textbooks. Additionally, you can suggest up to nine recommended textbooks.

Required Textbooks	ISBN-13 (or ISBN-10)
Mastering PyTorch: Create and deploy deep learning models from CNNs to multimodal models, LLMs, and beyond, Ashish Ranjan Jha, 2024	978-1801074308
Generative Artificial Intelligence. Exploring the Power and Potential of Generative AI, Shivam Solanki, Drupad Khublani, 2024	979-8868804038

Papers	DOI or URL
Y. Jin, et al. Efficient Multimodal Large Language Models: A Survey, 2024.	https://arxiv.org/abs/2405.10739v2
J. Cho, et al. Sora as an AGI World Model? A Complete Survey on Text-to-Video Generation, 2024	https://arxiv.org/abs/2403.05131v2
C. Zhang, et al. Text-to-image Diffusion Models in Generative AI: A Survey, 2023	https://arxiv.org/abs/2303.07909v2

Web-resources (links)	Description
https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models	
https://github.com/jianzhnie/awesome-text-to-video	
https://github.com/Yutong-Zhou-cv/Awesome-Text-to-Image	

8. Facilities

Software
Google Colab

9. Additional Notes

Comments

The contents of the lectures may be updated due to some significant technical releases in generative AI

The proposed course 1) has explicit academic content and requirements for receiving credits, 2) is in alignment with the program's learning outcomes, 3) adheres to policies and Skoltech regulations.

Lead Instructor confirms