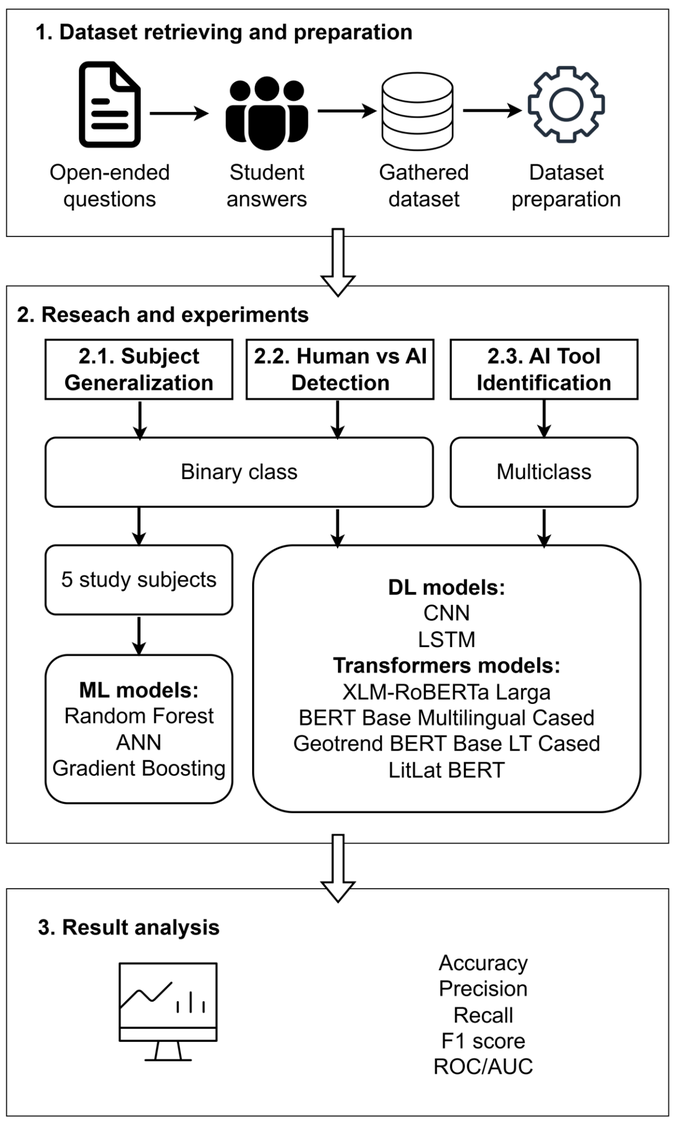
Detecting GenAI-generated Open-Ended Exam Answers across Multiple Computer Science Subjects

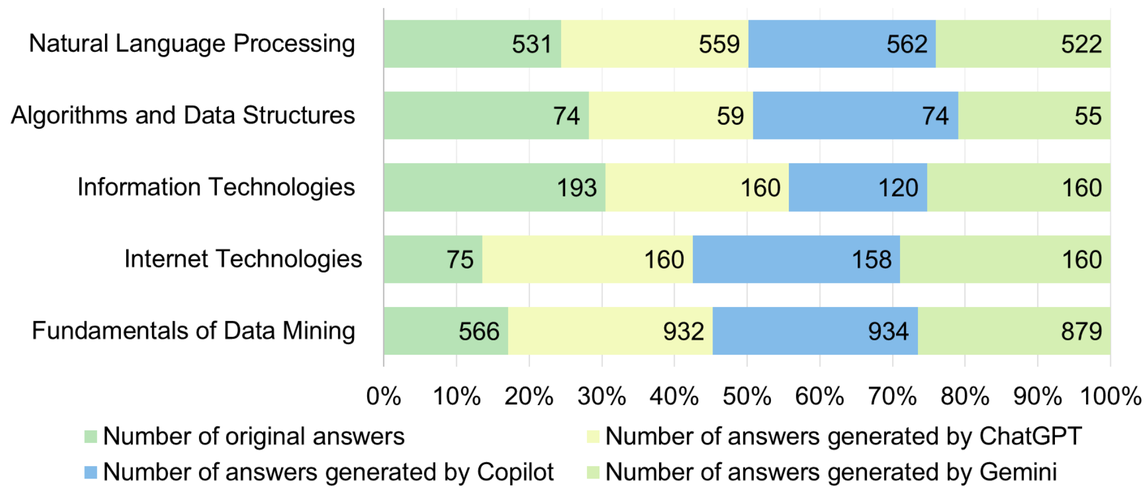
Urtė Radvilaitė, Birutė Pliuskuvienė, Simona Ramanauskaitė, Pavel Stefanovič



Research design and experimental workflow

Generative artificial intelligence (GenAI) tools are widespread adopted among students in their learning process and create challenges for verifying the authenticity of open-ended exam answers.

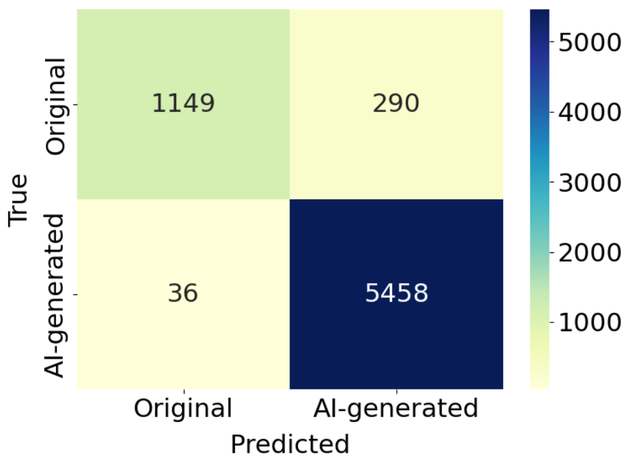
This study evaluates whether deep learning and transformer-based models can distinguish student-written from GenAI-generated answers in Lithuanian across five computer science courses.



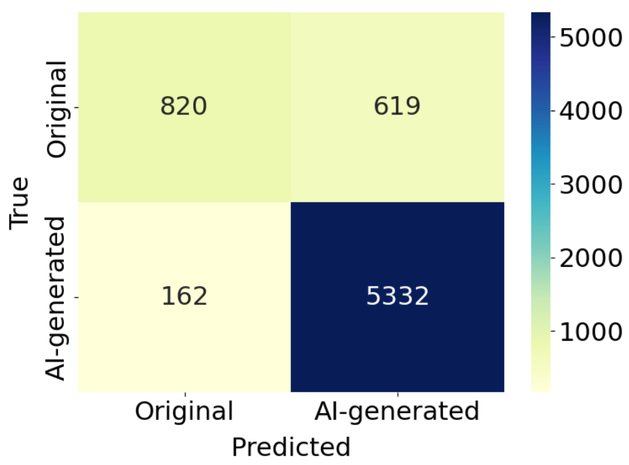
Distribution of classes across the study subject

Machine learning (ML) models were trained on data from the *Fundamentals of Data Mining* subject and data from other 4 subjects were used as test data.

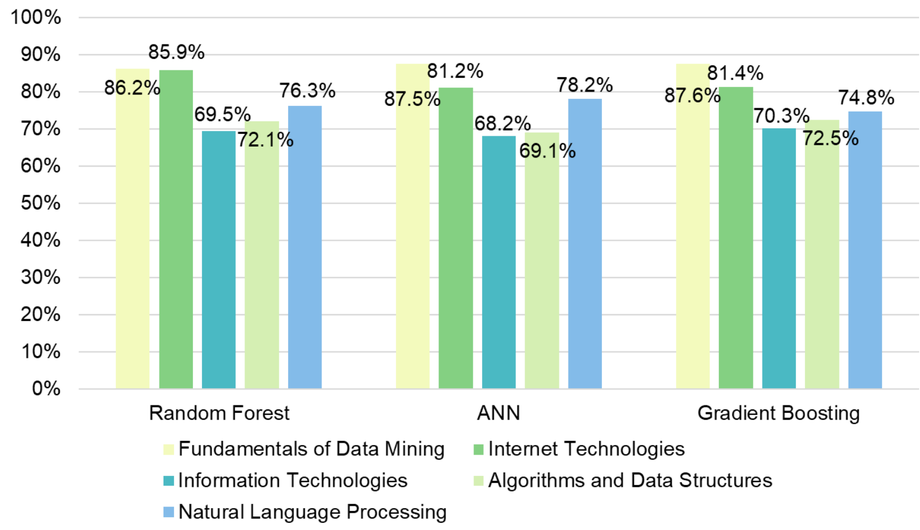
Deep learning models and transformer-based models were used to detect student-written and GenAI-generated answers. Furthermore, multiclass problem, determining which GenAI tool was used to generate answers, was solved.



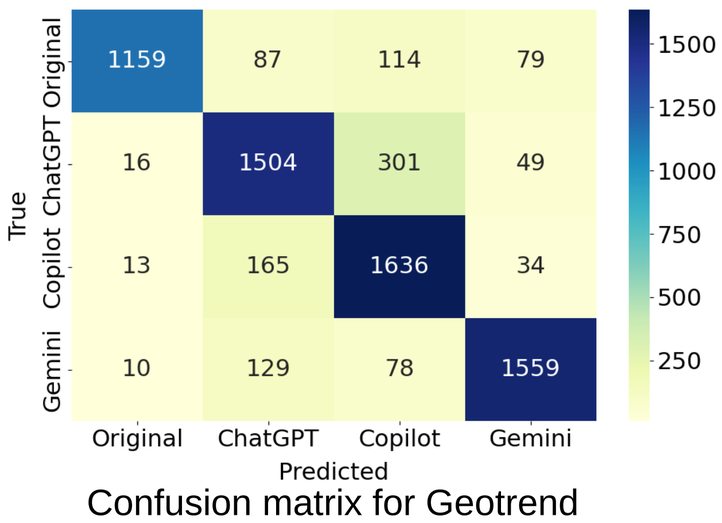
Confusion matrix for BERT Base Multilingual Cased model



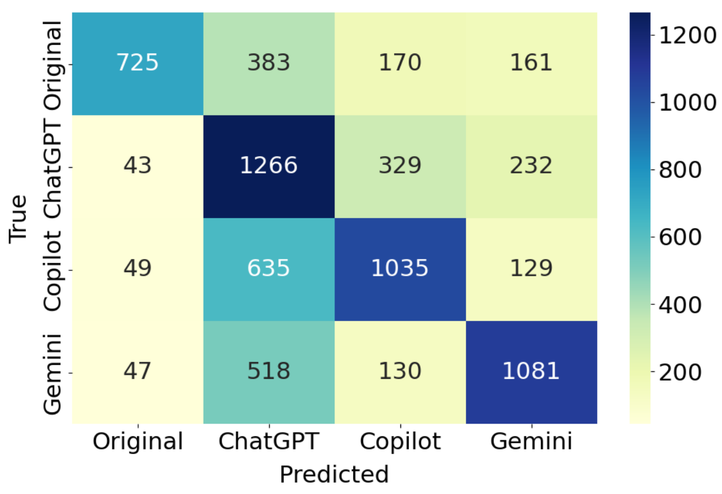
Confusion matrix LitLat BERT model



Accuracy for each study subject



Confusion matrix for Geotrend BERT Base LT Cased model



Confusion matrix for LitLat BERT model

ML models trained on data from a single study subject show limited cross-subject generalization, highlighting the need for more diverse training data.

Accuracy in distinguishing student-written from GenAI-generated answers reached 89-95%, while identifying the specific GenAI tool was more challenging, with accuracy ranging from 68% to 85%.