Project 1

In ability testing, cheating can compromise the validity of scores. Nowadays, many exams, like the TOEFL, are computerized and can record every action examinees take, such as when they move to the next question or how many times they change an answer. This generates log data, also known as process data. Traditional cheating detection methods mainly rely on rule-based approaches or simple statistical models, which struggle to effectively analyze such process data, leaving room for improvement in detection performance.

Motivated by this, I develop an anomaly detection system based on the BERT language model, leveraging its strong contextual modeling capabilities to identify complex behavioral patterns.

Here's how the model works. I treat an examinee's entire response sequence as a sentence, where each individual response acts as a token. Instead of using pre-trained embeddings like in traditional language models, I create embeddings by concatenating key features of each response, such as response time and answer accuracy. These embeddings go into Transformer encoder layers, where the attention mechanism captures contextual relationships between responses. Finally, the model outputs the probability of each response being anomalous.

This approach significantly improves anomaly detection performance compared to traditional methods by incorporating richer contextual information. More importantly, this study introduces a novel method for integrating multiple behavioral signals—such as response time, accuracy, and answer changes—into a unified analysis framework. This represents a breakthrough in educational measurement, as it enables a more comprehensive and data-driven approach to modeling examinees' behaviors.

Project 2

This project is a collaboration with the operational risk team at an investment bank. They maintain records of various anomaly events within the company, such as external fraud, execution errors, and processing failures. Each record contains a detailed description of the incident. However, these incidents are categorized into only a few broad classes using simple keyword matching, which limits the team's ability to identify meaningful risk patterns. They need a more nuanced classification approach based on the actual content of the descriptions.

To address this, I develop a topic modeling approach leveraging LLMs. I start by generating sentence embeddings for each record to capture its semantic meaning. Then, I apply clustering algorithms to group similar anomalies together. Finally, I use the LLaMA model with prompt engineering to generate a topic label for each cluster, making the classification process more interpretable and useful for risk assessment.

Compared to an existing topic modeling method, BERTopic, my approach improves classification accuracy and produces more semantically coherent topic labels, helping the risk team uncover deeper insights into operational risks.

Project 3

This project is a collaboration with a company. Employees communicate with customers daily through a chat platform, and these conversations are recorded. The company wanted to detect inappropriate or non-compliant discussions, but the existing rule-based methods—like keyword matching—were not very effective. They had low accuracy, and often flagged false positives.

To solve this, I built an anomaly detection model using the DeBERTa language model. The model processes short chat segments (around two minutes) and assigns a probability score indicating potential non-compliance. Since DeBERTa is already strong at language understanding, I applied a parameter-efficient fine-tuning approach, which allowed the model to quickly adapt to the company’s specific needs while keeping computational costs low.

This new approach significantly improved detection accuracy, reducing false positives and capturing more actual violations. It helped compliance teams focus on real risks instead of manually reviewing large amounts of irrelevant flagged conversations.

Example 1 挑战与解决问题类

In one of my research projects, we were preparing to submit our findings to a top academic conference. However, just two weeks before the deadline, we ran into two major challenges: the model's results were unstable, and each experiment took too long to run, making it difficult to iterate and finalize our findings in time.
As the project lead, I needed to address both issues quickly while ensuring we stayed on track to submit a high-quality paper. To tackle these challenges, I split the team into two groups—one focused on model optimization and the other on paper writing—so we could work in parallel. To improve efficiency, I introduced daily stand-ups and used Slack for real-time updates and quick problem-solving. On the technical side, I guided the optimization team to speed up experiment runtime by leveraging AutoML and parallelizing hyperparameter search, allowing us to iterate faster and stabilize the model.
Thanks to these efforts, we completed all necessary experiments and finalized the paper one day before the deadline. The paper was ultimately accepted at a top-tier conference, bringing valuable recognition to our team.

Example 2 超越预期与主动性类

In a research project, my role was to validate a LLaMA-based model developed by another team, ensuring it performed well on our education dataset. During testing, I noticed the model struggled with long-text inputs, leading to inconsistent accuracy. Fixing this wasn't part of my original task, but I saw an opportunity to improve it.
To address the issue, I explored two solutions. First, I used regex-based chunking to break long texts into smaller sections and feed only the relevant parts into the model, avoiding excessively long inputs. At the same time, I searched on Hugging Face and found a fine-tuned LLaMA model that could handle up to 80K tokens—far beyond the original 8K limit.
I shared these findings with the development team, and they decided to adopt the fine-tuned model as the backbone while using chunking for even longer texts. As a result, the model's performance on short texts remained stable, but its accuracy on long texts improved significantly.

Example 3 冲突与决策类

As a professor, I write grant proposals to secure funding. In one project, I collaborate with professors from the CS department to prepare an interdisciplinary research proposal. However, we disagree on the writing style—the CS team prefers highly technical language to emphasize academic rigor, while I believe the proposal needs to be more accessible for education-focused reviewers to fully understand its value.

As the lead on the proposal, I need to bridge this gap. First, I analyze successful proposals for this grant and find that most focus on explaining ideas rather than relying on formulas. I also consult senior professors with experience winning this grant, asking them to review our draft and advise on which technical details are essential.

Based on this, I propose a structured approach: in the methodology section, we keep key formulas where standard symbols are used but replace formulas that require newly defined symbols with plain-language explanations. This ensures clarity without sacrificing technical accuracy. After discussion, the team agrees on this approach. The final proposal receives positive feedback for its readability, and this strategy becomes our standard practice for future interdisciplinary collaborations.

Example 4 失败与反思类

In one of my early research projects, I lead a team developing a machine learning model to detect cheating in exams. I am confident in this approach and believe it will advance the application of deep learning models in this field.

However, after months of work, we realize that while our model performs perfectly on simulated data, it struggles with real-world data. The biggest issue is data quality—actual exam data is noisy and incomplete, and our model requires high-quality inputs, which are not feasible in practice. As a result, this approach not only fails to improve detection performance but also increases detection costs due to its complexity.

This failure teaches me two key lessons. First, a more complicated method isn't always a better solution—sometimes, a simpler, more interpretable model is more effective. Second, I learn the importance of working closely with practitioners who actually use the model. If we had involved frontline practitioners earlier in the development phase, we would have identified the data limitations sooner and adjusted our approach accordingly.

Since then, I always strive to make my research more practical and work closely with end users. This shift has helped my work gain more recognition from industry researchers and has led to more opportunities to collaborate with testing companies.