Dissertation on

"Capstone Tracker with an Integrated Evaluation System"

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology
in
Computer Science & Engineering

UE22CS320A – Capstone Project Phase - 1

Submitted by:

Dhanush S Jettipalle	**PES2UG22CS175**
Ketan Kancharla	**PES2UG22CS263**
Nitheesh Pugazhanthi	**PES2UG22CS371**
Rohan M G	**PES2UG22CS454**

Under the guidance of

Dr.Richa Sharma
Associate Professor
PES University

Aug-Dec 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'Capstone Tracker with an Integrated Evaluation System'

is a Bonafide work carried out by

Dhanush S Jattipalle	**PES2UG22CS175**
Ketan Kancharla	**PES2UG22CS263**
Nitheesh Pugazhanthi	**PES2UG22CS371**
Rohan M G	**PES2UG22CS454**

In partial fulfillment for the completion of Fifth-semester Capstone Project Phase - 1 (UE22CS320A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Aug 2024 – Dec. 2024. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 5th-semester academic requirements in respect of project work.

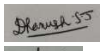
Dr. Richa Sharma Associate Professor	Dr. Sandesh B J Chairperson	Dr. B K Keshavan Dean of Faculty

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"**Capstone Tracker with an Integrated Evaluation System**"** has been carried out by us under the guidance of **Dr. Richa Sharma, Associate Professor** and submitted in partial fulfilment of the course requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester Aug. – Dec 2024. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES2UG22CS175	**Dhanush S Jattipalle**	
PES2UG22CS263	**Ketan Kancharla**	
PES2UG22CS371	**Nitheesh Pugazhanthi**	
PES2UG22CS454	**Rohan M G**	

ACKNOWLEDGEMENT

I would like to express my gratitude to **Dr. Richa Sharma,** Department of Computer Science and Engineering, PES University, for his/ her continuous guidance, assistance, and encouragement throughout the development of this UE22CS320A - Capstone Project Phase – 1.

I am grateful to all Capstone Project Coordinators, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Professor & Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro-Chancellor, PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University, and Prof. Nagarjuna Sadineni, Pro-Vice Chancellor, PES University, for providing me with various opportunities and enlightenment every step of the way. Finally, Phase 1 of the project could not have been completed without the continual support and encouragement I have received from my family and friends.

TABLE OF CONTENTS

	6.4.1 : Introduction 6.4.2 : Implementation 6.4.3 : Evaluation 6.4.4 : Conclusion 6.5 : LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing 6.5.1 : Introduction 6.5.2 : Implementation 6.5.3 : Evaluation 6.5.4 : Conclusion 6.6 : The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review 6.6.1 : Introduction 6.6.2 : Implementation 6.6.3 : Evaluation 6.6.4 : Conclusion 6.7 : Fine-Tuning LLMs for specialized use cases 6.7.1 : Introduction 6.7.2 : Implementation 6.7.3 : Evaluation 6.7.4 : Conclusion 6.8 : LLMs in Automated Essay Evaluation: A Case Study 6.8.1 : Introduction 6.8.2 : Implementation 6.8.3 : Evaluation 6.8.4 : Conclusion 6.9 : LLM-RUBRIC: A Framework for Multidimensional Evaluation of Natural Language Texts 6.9.1 : Introduction 6.5.2 : Implementation 6.5.3 : Evaluation 6.5.4 : Conclusion 6.10: Large Language Models as Tools for Textual Analysis 6.10.1: Introduction 6.10.2: Implementation	

List of Figures

Introduction

In our project "Capstone Tracker with an Integrated Evaluation System" aims at two primary things, one is to simplify the process of tracking all the capstone projects, their progresses and their deliverables and the other is to automate the evaluation of the capstone reports and final papers using LLMs.

The capstone tracking system will help address the challenges that a capstone guide might face while managing multiple teams and multiple deliverables for each team which will act as a centralized server where the mentors and panel members can track the progress of each team through every phase. By streamlining the submission and review of deliverables, the tracker allows mentors to focus on guiding students academically while our project will take care of the rest.

In addition to the tracker, we plan on solving the issues mentors face while evaluating reports which are time consuming and too long to evaluate consistently by introducing a LLM which will be fine-tuned to evaluate most of the deliverables that is expected in the capstone process as well as the evaluation of the final paper. The LLM will need to check for the formats and have to check for various parameters like relevance, factual correctness etc.

We plan on introducing a new quality index that will better fit our reports and help us better evaluate the reports accordingly. This way will also give us a fair and objective way of evaluation without any disparities.

Problem Statement

Our project's main aim is twofold: one to make the tracking of the various capstone deliverables much easier and two to automate the evaluation of the submitted capstone deliverables.

To reduce the burden of capstone mentors in keeping track of each individual team and their deliverables, we introduce the capstone tracker. This allows mentors and the capstone committee to track the progress of every team phase wise. This allows the mentors to focus on the academic part of the projects and leave the management part to our product. The teams will be able to submit all the deliverables and can be accessed easily anywhere and at any time by the mentors and the committee. The Project will also take care of evaluating the deliverables submitted by the students. First there will be a plagiarism check done, to avoid any cases of stealing information or content.

After the plagiarism check the deliverables i.e. the power point presentations and the reports. The mentors need not comb through all the reports and the documents for errors and grammatical mistakes. The submitted deliverables will be assessed on various parameters like clarity, readability and technical soundness.

The project will also be able to provide constructive feedback on the report quality and thus act like a reviewer or at most help reinforce the mentor's suggestions. The project will also be able to analyze the reports and predict whether the project can be extended and worked for a longer period of time. Thus, we propose a fine-tuned LLM which will run at the backend of the project and assess the documents on our newly defined quality index.


AI CONTENT DETECTOR

ABSTRACT AND SCOPE

Our project mainly explores the application of Large Language Model (LLMs) in automating and augmenting the systematic review process in academic or literature research. Since the volume of scientific literature research. The explosion of the volume of literature of science can pose difficulties using this labor-intensive manual search to check many things within short amounts of time: time consuming as well as intensive for human sources-this study determines if LLMs could hasten systematic reviewing-from preliminary searching or scanning for relevancy through information extractions We study both LLM technological ability as well as capability but then highlight limits related to such studies. It further emphasizes that the comparison of LLM-assisted and traditional methods for systematic reviews must be considered on the strength side, such as efficiency, accuracy, and reliability. It also goes deep with issues of ethics and methodological challenges in introduction of AI-driven review processes in the academic research frameworks.

Capstone Project Tracking Create a comprehensive digital platform to monitor and manage capstone project deliverables

Enable real-time tracking of project phases, milestones, and progress

Reduce administrative burden for capstone committees through automated tracking and reporting

Research Review Automation using Large Language Models (LLMs)

Leverage advanced AI technologies to streamline the academic paper review process

Develop a fine-tuned LLM model to assist in preliminary screening and assessment of research submissions. The system will feature a custom quality index for assessing submitted papers, generate automated reviews and scores, and enable students to enhance their project impact.

Deliverables include the software system, comprehensive documentation, and a detailed presentation deck covering technical architecture, LLM capabilities, implementation methodology, performance metrics, and future enhancement roadmap.

The research examines LLM architectures optimized for academic text processing, focusing on transformer-based model's capabilities in parsing scientific language, domain-specific terminology, and contextual understanding across disciplines. Primary emphasis lies on model's proficiency in interpreting scientific literature structure, including methodological frameworks and statistical analyses. This method centers on creating standardized protocols for the integration of LLMs in systematic reviews so that the procedure of screening articles is performed with the maximum level of precision and the mechanism of extracting data through automation processes are highly accurate. The primary methodological elements consist of frameworks of quality assessment and validation of the findings from LLMs, ensuring the reliable synthesis of data across different studies. Accuracy, completeness, efficiency, and resource utilization metrics will be evaluated through comparative analysis of reviews produced by LLMs with the traditional human-conducted reviews. The encompasses standardized validation protocols that are applicable across academic disciplines and ensure methodological robustness and generalizability of the LLM-assisted review processes.

- Integration of Automated Scoring: Developing a robust algorithm for automated scoring of deliverables while ensuring fairness and transparency can be challenging, especially when subjective evaluation is required but at the same time it is not practical to read line by line of the deliverables to evaluate them accurately.
- Contextual Scoring Systems: A gap exists in creating automated scoring algorithms capable of understanding the context and quality of deliverables beyond surface-level features.
- Auto-Evaluation Complexity: Designing an auto-evaluation system capable of assessing diverse deliverables (e.g., reports, presentations, code) accurately across different capstone phases is a significant technical challenge.
- Real-Time Evaluation Models: Research on developing models that can evaluate deliverables in real-time with minimal latency is still evolving.

RESEARCH / TECHNOLOGY GAP AND CHALLENGES

Developing an automated tracking system for capstone projects presents several challenges. Integrating a fair and transparent scoring mechanism requires robust algorithms, particularly for subjective deliverables. Incorporating a reliable plagiarism detection tool that can handle various formats, such as text, code, and diagrams, is complex and resource-intensive. Ensuring secure, real-time access to submitted files from anywhere demands a scalable and secure cloud infrastructure. The auto-evaluation system faces difficulties in assessing diverse deliverables across different capstone phases accurately, while generating specific, constructive feedback necessitates advanced natural language processing and machine learning capabilities. Scalability to manage submissions from 150 teams, along with data privacy and security, adds another layer of complexity. Furthermore, interoperability with existing academic systems, balancing automation with human oversight to mitigate biases and errors, and ensuring faculty and student adoption through adequate training are significant challenges.

In terms of research gaps, there is limited advancement in plagiarism detection techniques capable of handling multimedia deliverables effectively. The lack of contextual scoring systems that evaluate the quality and depth of submissions beyond surface-level features remains a significant gap. Personalizing feedback based on submission content and project phase is still an underdeveloped area. Research on real-time evaluation models with minimal latency and tools adaptable to diverse academic disciplines is also evolving. Additionally, addressing AI biases in evaluations and developing comprehensive systems that evaluate submissions across different formats holistically are crucial areas requiring further study. Data anonymization techniques for unbiased evaluation and frameworks for human-AI collaboration in academic assessments are also underexplored. Finally, creating institution-specific systems that adapt to varying academic policies and evaluation criteria remains a pressing need.

Objectives

1. Automated tracking system: To reduce the burden on the capstone committee, on looking out for various capstone deliverables, and following up the submission of around 150 teams.
2. Faster scoring system: Using the automated system one can easily which teams have missed their deadlines and can assign marks accordingly.
3. Any-time File access: The deliverables submitted can be accessed and viewed anytime and anywhere by the faculty.
4. Auto-evaluation of reports and ppts: The proposed project can evaluate the various phase-wise deliverables of the capstone project, thus reducing the burden on the mentor.
5. Quality of the deliverables: The project will be able to evaluate the submitted deliverables on various qualities like clarity and technical soundness.
6. Specific Suggestions: The project will be able to provide constructive suggestions about the capstone deliverables. This will act as reviewer as well and thus help the faculty focus more on the academic support.
7. Judging the scope: The project will also be able to analyze a capstone report and other deliverables and be able to predict whether the project can be extended and worked on a longer period to enhance it further.
8. Plagiarism Check: The project will have an inbuilt plagiarism checking tool which used on every deliverable submitted to the tracker. Only after the reports pass the plagiarism check, will they be assessed on the various parameters.
9. Instant results : Since the evaluation will be done by a finetuned LLM , it will publish the results of the various checks and tests instantaneously.

LITERATURE SURVEY

Reference [1]

6.1.1

Introduction:

This paper performs an extensive pipeline of tests on various SOTA LLMS, to find if we can automate the peer review process. They pose the LLM multiple types of prompts asking to assess the submitted manuscript on various parameters like soundness, novelty and impact. They use three types of prompting techniques and check which yields the better results.

6.1.2

Implementation:

The paper tries to check the usefulness of the LLMs in three major tasks: finding technical faults in papers, selecting a better abstract for the same paper and then verifying author provided checklists.

Finding Technical faults: 13 short computer science papers were drafted with deliberately placed technical errors. Among the LLMs ChatGPT-4 performed the best. It had identified 7 out of the 13 faults in the papers. The errors present range from technical, logical and mathematical.
3 targeted prompt styles were used: prompt-direct, prompt-one shot, prompt-parts.
In which prompt-parts yielded better results than the other 2.

Selecting a better abstract: The authors drafted two abstracts for each paper. Where it was easy to identify

which abstract was better. ChatGPT-4 was only able to identify 4 better abstracts out the 10 given papers.

Verifying author given checklist: A checklist was drafted from 15 NeurIPS 2022 papers. Then GPT-4 was assigned to check for all the 119 items in each paper. It had achieved 86.6 percent accuracy.

6.1.3 Evaluation:

Technical Faults: Since the LLM shouldn't be assessed on its ability to remember the ratings from the data it has been trained on, 13 new papers were drafted for the testing purpose. Each had a technical/mathematical error. GPT-4 was able to identify 7 out of the 13 errors.

Verifying Author Given Checklists: 119 items were provided to GPT-4 to check in each paper. It had achieved 86.6 percent accuracy, while showing equal performance with all prompt types.

Choosing a Better Abstract: For 10 papers, a pair of abstracts were drafted. A human could easily identify the superior abstract, but GPT-4 struggled. It was able to identify only 4 out of the 10 cases.

6.1.4 Conclusion:

This paper concluded that GPT-4 is the best for automated reviewing out of the various other SOTA LLMs. Although it shows a lot of potential to aid the panel reviews it can never completely replace them at this point of time. It shows increased performance when parameters are assessed at a time.

Reference [2]

6.2.1

Introduction:

This paper tests GPT-4 on review generation and score generation of the research papers. It also proposes a new dataset RR-MCQ. These questions have been taken from the ICLR conference. These will be prompted to the LLM to help in generating specific reviews. This method is further compared to other prompts.

6.2.2

Implementation:

Three methods were tested: Aspect Score prediction, Review Generation and then the new RR-MCQ dataset.
Aspect Score Generation: A dataset containing papers from the ICLR -2017 labelled with their human written reviews were fed to GPT-4 as input. The LLM is tasked with reading and understanding reviews and later generating the scores for the respective papers. The score generation reduces greatly when no data is fed to the LLM.

Review Generation: The ICLR-2020 subset of the ASAP dataset, with 300 papers and 902 reviews annotated by aspect. The model then is tasked with generating reviews based on this data using zero and few shot prompts.

RR-MCQ: A novel dataset of 196 multiple-choice questions derived from ICLR-2023 review-

rebuttal discussions were formed. The LLM is then asked these questions from the dataset where it can propose new changes to the submitted manuscript.

6.2.3 Evaluation:

Aspect Score Generation: To evaluate the aspect score generation Pearson, Spearman, and Kendall's Tau Correlation coefficient were used. GPT-3.5 achieved a score of 0.65 with Pearson's coefficient. It was used to measure the correlation between the human reviews and the LLM generated scores.

Review Generation: Manual correction of the reviews generated by the LLM showed it lacked in informativeness and relevance.

RR-MCQ: GPT has achieved a micro accuracy of 71% but a macro accuracy of only 27.6 %. It shows that it still has long ways to go in logical reasoning and functional thinking.

6.2.4 Conclusion:

This paper shows some promising results when we use very specific prompts to the LLM like the RR-MCQ dataset. It shows that present era LLM although cannot replace reviewers but can assist them or act like reinforcements to their reviews and scores as well. The LLMs also need to develop long-term understanding as well to understand the entire research paper.

Reference [3]

6.3.1 Introduction:

This paper tries to reduce the issues that are present currently in the process of automating peer reviews using LLMs. They try to extend the context window by also providing extra information about the conference the manuscript has been submitted to. They also propose a watermark for the reviews generated to prevent misuse of the AI generated reviews.

6.3.2 Implementation:

This paper uses contextual prompting, i.e. it provides information about the conference the author wants to publish in. The details provided are: review forms, reviewer guidelines, ethical codes, area chair guidelines, and historical review statistics. All these features are put to test in an ablation study where performance peaked only when all 5 of these features were fed along with the paper itself. Further-more, now the probability of the next words to be generated were tweaked which now performs as a watermark for the generated reviews.

6.3.3 Evaluation

Ablation studies were set up to decide which among the 5 context features were important to the

LLM on reviewing a paper. Previous year statistics seemed to have a higher impact than all the others.

Now the generated reviews along with human reviews were put up for a blind test. Many researchers along with area chairs were given 2 sets of unlabeled reviews, where they had to judge how well are these reviews. Experts correctly identified LLM-generated reviews 59.8% of the time, validating the blind evaluation process.

6.3.4

Conclusions:

This paper concludes that a LLM can augment and reinforce a reviewer but not completely replace them yet. They say that currently this space is rich for human-AI collaboration. Although LLMs lack the critical thinking power to correctly understand the context and concepts present in a research paper. However feeding the LLM with context like area chair guidelines and previous year statistics , can improve their performance slightly and introduces some variability in their responses.

6.4

Reference [4]

6.4.1 Introduction:

This paper addresses the issues of automated English article scoring with an emphasis on the need for accurate, objective evaluations in educational contexts. Traditional grading methods are labor-intensive and biased by subjective influences, especially when grading diverse genres such as narrative and explanatory essays. This proposed system captures unique genre-specific features by integrating BERT and Chat-GPT, thus delivering comprehensive evaluations and personalized feedback. This innovation aligns with the rise of online education and aims to reduce the burden of manual grading while improving the quality of automated assessments.

6.4.2 Implementation:

The proposed implementation integrates BERT for feature extraction and Chat-GPT for feedback generation. The input text is tokenized by BERT's WordPiece tokenizer, embedding word, segment, and positional information. A fully connected layer then scores the output of the BERT model as the overall semantic representation. To further evaluate, genre-specific features are identified and analyzed. Chat-GPT provides detailed, personalized feedback about the strengths and weaknesses of the article, focusing on improvement in writing. The system achieves high accuracy and adaptability, validated on the ASAP++ dataset.

6.4.3

Evaluation:

The ASAP++ dataset, that includes argumentative, narrative, and question-answering essays, scored for genre-specific feature, was used for the evaluation of the proposed system. The QWK metric, which denotes the quadratic weighted Kappa, was reported as 0.803; this was a better performance than the state-of-the-art models, Tran-BERT-MS-ML-R. The generalization across genres was also better for the proposed system, which indicated certain improvements in narrative essay scoring. Experimental results confirmed its effectiveness in delivering accurate, genre-aware evaluations and personalized feedback, making it a robust tool for automated article scoring.

6.4.4

Conclusion:

The study presents an advanced article scoring system, integrating BERT and Chat-GPT to help resolve genre-specific challenges in article evaluation. It brings the advantage of accuracy and comprehensiveness delivered by feature-based scoring blended with personalized feedback compared to traditional methods. Experimental results from ASAP++ have shown its better performance and adaptability with genres. It further alleviates burdens from manual graders and furnishes actionable advice on how a writer can correct and improve work. The significance is also highlighted based on the possibilities AI-driven technologies could bring, to transform evaluative processes and contribute to facilitating more personalized educational assessments.

6.5

Reference [5]

6.5.1

Introduction:

This study looks into the capabilities of LLMs to help NLP researchers in the task of reviewing and meta-reviewing academic papers. The work presents the ReviewCritique dataset that contains both human-written and LLM-generated reviews with fine-grained deficiency annotations. Two primary questions are explored: the capability of LLMs as reviewers compared to humans and the capability of LLMs as meta-reviewers to identify deficiencies in reviews. The findings reveal that although LLMs can be used for reviews, the results are usually superficial or incomplete, which underlines the need for human expertise in such knowledge-intensive tasks. This work gives valuable insights into the strengths and limitations of LLMs in academic peer review processes.

6.5.2

Implementation:

This research explores the capabilities of large language models in helping NLP researchers, especially in reviewing and meta-reviewing academic papers. The paper introduces the ReviewCritique dataset, which consists of human-written and LLM-generated reviews with fine-grained deficiency annotations. It answers two key questions: how well LLMs perform as reviewers compared to humans and their effectiveness as meta-reviewers in identifying deficiencies in reviews. The results show that although LLMs can provide support for reviews, the reviews made by them appear to be sometimes shallow or partial, thus needing human expertise on knowledge-intensive issues. This study offers valuable contributions toward understanding strengths and limitations of LLMs for use in an academic peer-review process.

6.5.3 Evaluation:

This evaluation used the ReviewCritique dataset, which compares human-written reviews with those from LLMs on parameters like deficiency detection, review diversity, and constructive feedback. It assessed the performance of LLMs in detecting deficient review segments through precision, recall, F1 scores, ROUGE, and BERTScore. Results show that the deficient segments are found to have a higher percentage of LLM-generated reviews lacking specificity and diversity. Closed-source models like Claude Opus and GPT-4 showed better quality and depth but not as perfect as human evaluators. That results indicate limitations in the usability of LLMs for expert tasks, which may include, as in our example, meta-review and pure review.

6.5.4 Conclusion:

It therefore explores the capabilities and limitations of LLMs for supporting NLP researchers in conducting review and meta-review activities. LLMs can actually produce reviews; however, it is usually vague or insufficient. The kind of feedback needed to evaluate academically is very deep and precise. The ReviewCritique dataset allows for an extremely fine-grained comparison of human versus LLM-generated reviews, which, in turn shows that LLMs fail when it comes to nuanced judgment and reasoning. Human expertise is still needed for high-quality peer reviews despite the progress. This research verifies the importance of proper integration of LLMs and that humans are collaborators, not competitors, in knowledge-intensive tasks.

6.6

Reference [6]

6.6.1 Introduction:

The exponential growth of scientific literature has made systematic and scoping reviews essential tools for synthesizing evidence across a wide range of disciplines. Traditional review processes are, however time-consuming, sometimes taking months or even years to complete. The potential of Large Language Models, including GPT and BERT, is that they could automate several steps in the review process, from literature screening and data extraction to evidence synthesis. This paper evaluates the current applications of LLMs in review automation, evaluates their performance, and outlines how they could change the landscape of evidence synthesis by making it less time and resource demanding without sacrificing rigor and accuracy.

6.6.2 Implementation:

This study explores the use of Large Language Models (LLMs) to automate systematic reviews. It reviewed 3,788 articles, identifying 172 eligible studies. GPT-based models were most commonly used (73.2%) and showed better performance in data extraction (precision: 83%, recall: 86%) than BERT models. Automation covered stages like data extraction, publication search, and screening, significantly reducing review time. While promising, few studies used LLMs for the full review cycle, and ethical concerns and limitations in precision remain. The findings indicate LLMs can enhance efficiency in systematic reviews, with potential for widespread adoption in research workflows.

6.6.3 Evaluation:

The evaluation shows that GPT-based models outperformed BERT in data extraction tasks, with higher precision at 83% and recall at 86%. However, they performed slightly worse in title and abstract screening compared to BERT. LLMs significantly reduced the time required for systematic reviews, saving over 300 hours in screening and extraction processes. Despite their efficiency, challenges like lower accuracy in some tasks, ethical concerns, and the need for human oversight persist. Only 15.1% of studies fully utilized LLMs in conducting reviews. Overall, LLMs show great potential for review automation but require improvements in precision and reliability for broader adoption.

6.6.4 Conclusion:

The study concludes that Large Language Models (LLMs) revolutionize systematic reviews by automating time-consuming stages such as data extraction, screening, and evidence synthesis. It was found that GPT-based models were the most effective models in reducing the workload and improving efficiency. Still, accuracy limitations, ethical concerns, and dependency on human supervision stand as challenges. While only a few studies integrated LLMs fully in the entire review process, there is clear evidence of their ability to democratize and accelerate reviews. Further advancement could transform scientific evidence synthesis to become faster and more accessible, maintaining quality.

6.7 *Reference [7]*

6.7.1 Introduction:

This paper talks about the fine tuning of large language models for specific domains and tasks. Even though pre trained LLMs like BioBert or GPT-4 can easily process and generate human – like text while understanding the context, they are often limited in their effectiveness in specialized fields for example medicine where domain specific knowledge and expertise is of utmost importance. By fine tuning these LLMs we can train the LLMs to perform well on these domain specific tasks with high accuracy and efficiency.

6.7.2 Implementation:

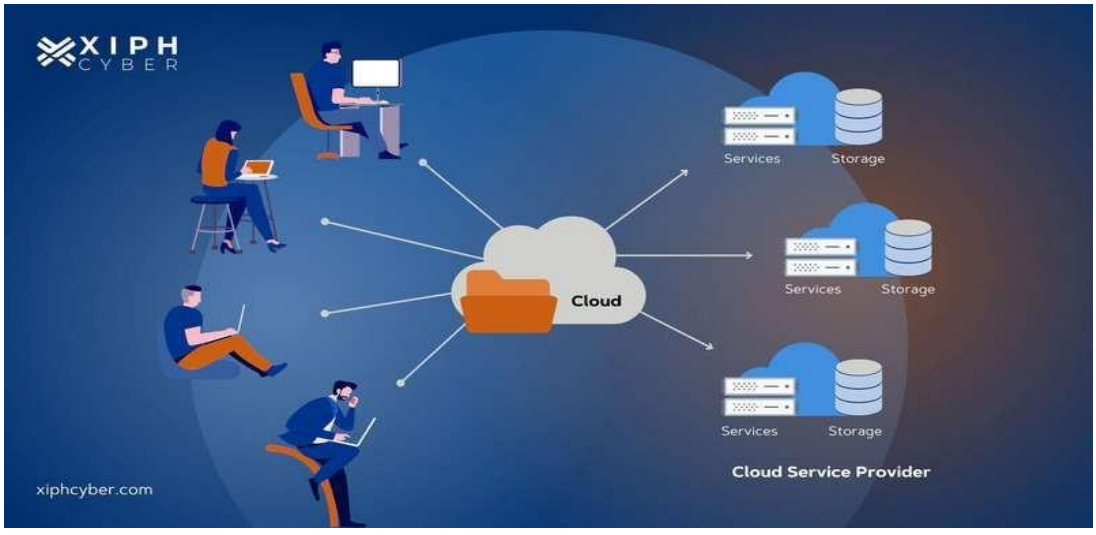
This paper discusses the various types of fine tuning which includes transfer learning, multitask learning and instruction learning, and reinforcement learning from human feedback which includes techniques like reward modelling, proximal policy optimization, preference feedback and comparative ranking.
The pipeline for fine tuning includes data preparation which is the data set preparation for fine tuning the LLM which are pre-processed, model selection which involves choosing an appropriate model as the base LLM some examples for this step can be BERT, LLaMA, GPT etc. The next step is the fine tuning which is training the model on a task specific dataset. Here the hyperparameter are tuned such as changing learning rate, batch size, number of epochs etc.

6.7.3 Evaluation:

The model after fine tuning model is validated and evaluated using metrics like accuracy, ROGUE score, and basic human evaluation. In this paper the fine-tuned model was evaluated on the medical domain and it showed that the fine-tuned models (in this case BioBERT was fine-tuned using PubMed abstracts for biomedical text mining) outperformed the base models in the domain specific task.

6.7.4 Conclusion:

This paper expresses the potential of fine-tuned LLM in specialized use cases by using specific datasets and fine-tuning techniques which helps us achieve better evaluation metrics despite challenges we face like hallucination, legal and safety concerns, biases in dataset, data leakage which may negatively affect our results.



6.8

Reference [8]

6.8.1

Introduction:

This paper looks into the use of LLMs specifically the GPT-4 model for the automated evaluation of student's essays. Since manual evaluation is quite tedious and time consuming this solution is a great alternative to the problem with the LLMs advanced text processing capabilities. This particular paper is applied in evaluating German language transfer assignments at the Swiss Institute of Business Administration (SIB), comparing LLM performance to the human lecturer's evaluation.

6.8.2

Implementation:

For the dataset, transfer assignments from the SIB which is used to test the practical application of business knowledge. These assignments were graded based on a specified rubric which had six criteria for the evaluation. GPT-4 was the model selected for its ability to work with multi modal data. Three experiments were conducted initially, the model was provided with the rubric for the evaluation, but the feedback did not alignment with the rubric. The second attempt included detailed evaluation instructions which resulted in slightly better results but still the feedback was inconsistent. In the third test, all evaluation information was embedded in the prompt which led to the overestimation of scores.

6.8.3

Evaluation:

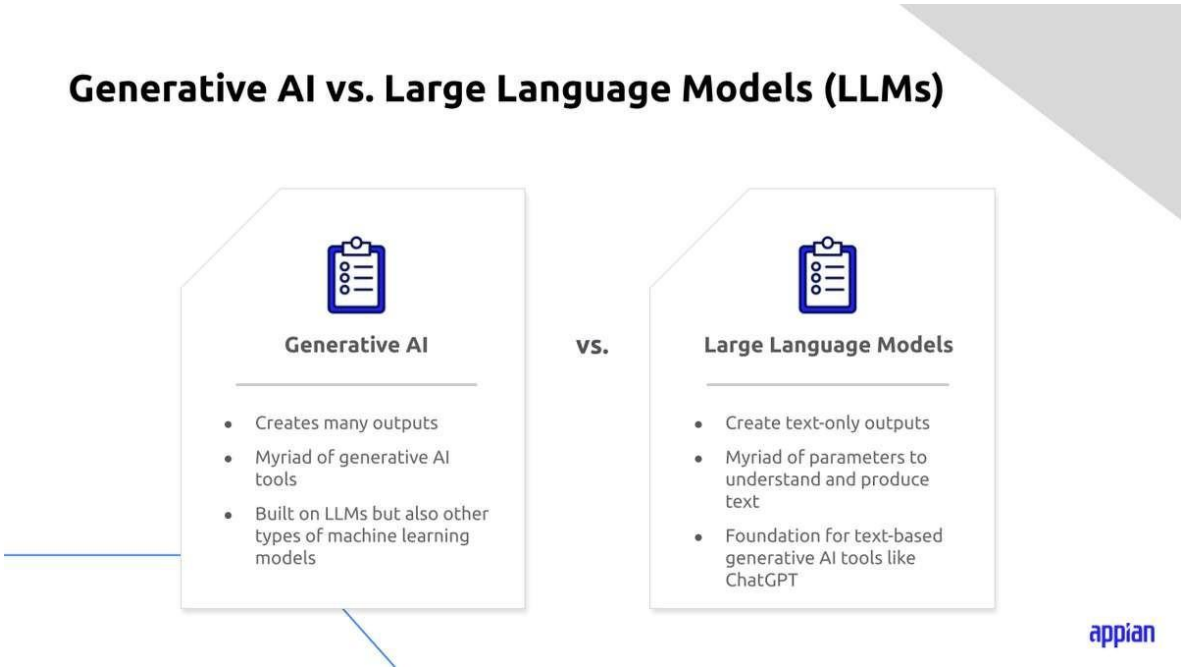
The grading was both lectures and the model and was evaluated on metrics like MAD and PCC. The finding revealed that in the first test GPT-4 assigned 52 points to a low-quality assignment. In the final attempt

GPT-4 provided maximum points to the same misaligning with the quality of the work. While the human evaluators displayed variance in grading, their assessments aligned more closely with predefined criteria compared to GPT-4.

6.8.4

Conclusion:

This paper comes to the conclusion that while LLMs like GPT-4 show potential for automated essay evaluation, their current limitations, such as hallucinations, misaligned feedback, and over-reliance on prompts, make them less reliable for nuanced academic assessments. Hence, we need to focus on improving the LLMs rubric assessment and develop domain specific datasets which will solve this problem and LLMs could be used to streamline assessment processes.



6.9
Reference [9]

6.9.1
Introduction:

The paper introduces a framework LLM-RUBRIC which is designed to automate the evaluation of natural language texts using a multidimensional and rubrics-based approach. It solves problems like manual evaluation being inconsistent and labor intensive by using a LLM and a calibrated feed forward neural network. It predicts human judgments across multiple dimensions of text quality, such as naturalness, conciseness, and citation quality, to provide an overall evaluation score.

6.9.2
Implementation:

The first step in the framework is a rubric construction which consists of all the evaluation metrics like naturalness, citation accuracy etc. We predefine the scoring criteria for each type of question. The LLM then predicts the score for each metric for the text generated which generates a distribution of responses for each dimension. A feed forward neural network was trained and is used to make the LLM predictions align with human behavior. After this network the final scores for each evaluation metric is attained.

6.9.3
Evaluation:

The framework was tested on information-seeking dialogue systems in the IT help domain. Metrics such as root mean square error and Pearsons's correlation coefficient with the human evaluation was used to evaluate the model. It is seen that the RMSE was half than of the base models and the PCC was significantly high er was the framework as compared to the base models.

6.9.4

Conclusion:

This paper shows us that LLM-RUBRIC which is a LLM trained to evaluate a rubric and has a calibrated neural network for automated text evaluation, though this approach efficiently predicts human behavior it is limited by the extensive pretraining and calibration that is required. Future work could explore adaptive rubrics, finer-grained evaluations, and broader applications in education and dialogue systems.

6.10
Reference [10]

6.10.1

Introduction:

The focus of this paper is the use of LLMs like GPT 3.5 for textual analysis to remove subjective bias. It aims to provide a systematic, scalable and a more objective alternative for deductive coding. By repeatedly applying a structured guide analysis overtime the LLM learn to identify the theme and concepts of textual data. They use Large Language Model Quotient (LLMq) as a quantitative measure of consistency and reliability in LLM generated analysis as compared to human generated analysis.

6.10.2

Implementation:

The dataset used for this model is the archival interview transcripts from a project exploring the transition of Ph.D. students into independent researchers which is already preprocessed and was analyzed for 5 codes: Autonomy, Persistence, Identity Perception, Novelty, and STEM Interests. Chat-GPT 3.5 was the model selected and prompts were crafted to evaluated the 5 predefined metrics. For each text sample, the query was iteratively run 160 times. Each iteration involved logging out and logging back in to ensure independence of results and to account for the stochastic nature of LLMs.

6.10.3

Evaluation:

Three human coders independently analyzed the same interview excerpts using the same set of codes. Results were compared to LLM-generated outputs to evaluate the results. Results showed that the output

from the LLMs converged after the iterations. LLMs effectively identified straightforward codes such as "Autonomy" and "Persistence," while harder codes like "Identity Perception" and "STEM Interests" required contextual prompts for reliable detection.

Overview of Datasets

To fine-tune the LLM on the quality index , we will need to collect various scientific research papers. We plan on using the previous capstone research papers as they would have multiple comments and score labelled with them.

To evaluate the LLM we will need to test them with unpublished papers, i.e. papers that the LLM has not seen. This is done so that the LLM doesn't remember the reviews and scores for the test paper as already published papers have them on internet, on which the LLM has been trained on.

Therefore, we plan on using the current batch's capstone papers before they publish to test out the validity of the scores and reviews generated by our product

CONCLUSION OF CAPSTONE PROJECT PHASE - 1

In this Phase of the capstone project, we achieved the following:

1) Initial Project Report:

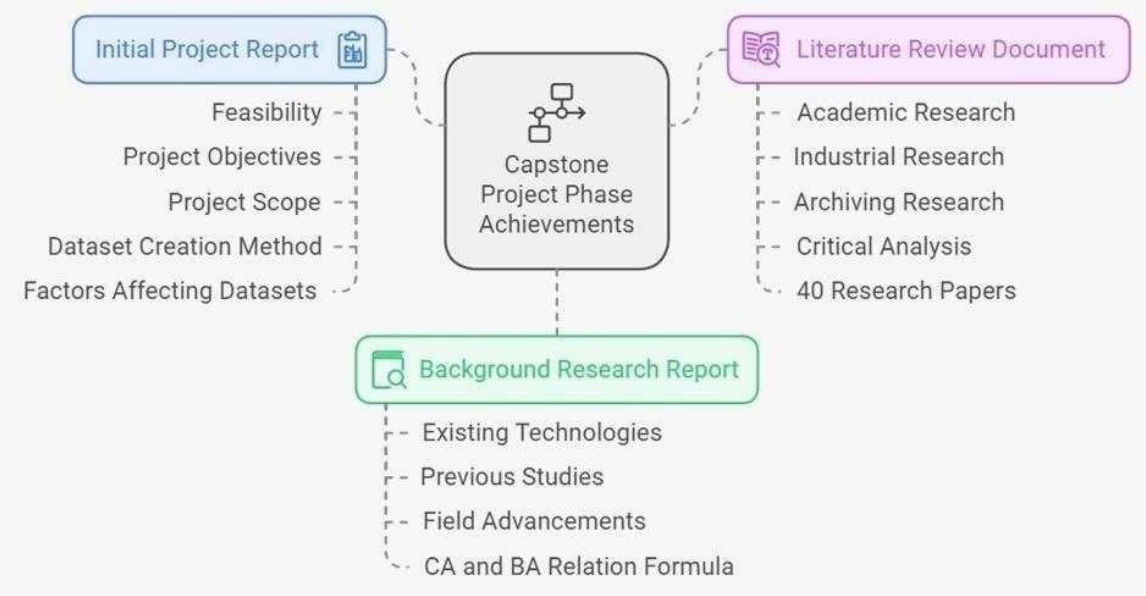
- Defining the outlines of the project such as objectives, problem statement and planned methodology for the project.
- Defining the scope and establishing project plan for future development.
- Identify the method of creating a dataset.

2) Background Research Report:

- Comprehensive understanding of existing technologies and diagnostic methods in the field.
- Finding previous studies on the topic and assessing feasibility.
- Scoping out any advancements in the field

3) Literature Review Document:

- In depth exploration of academic and industrial research.
- Archiving all known research and prior works in the field.
- Critical analysis of works and pinpointing areas of interest as well as future improvement.


Initial Project Report
Feasibility
Project Objectives
Project Scope
Dataset Creation Method
Factors Affecting Datasets
Capstone Project Phase Achievements
Literature Review Document
Academic Research
Industrial Research
Archiving Research
Critical Analysis
40 Research Papers
Background Research Report
Existing Technologies
Previous Studies
Field Advancements
CA and BA Relation Formula

PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

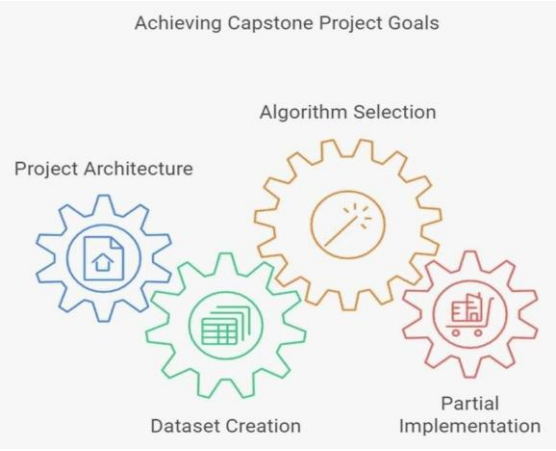
Plan for capstone phase 2

1. Project architecture: We need to design a system architecture for the automated report and paper evaluator and details about key components and data flow. We need to start developing modules for the capstone tracker and decide what the functionalities of this tracker will be.

2. Securing Dataset: We have to start collecting the deliverables of the previous year capstone projects and their deliverables.

3. Design and Model Selection: We need to decide on what LLM we will work on fine tuning to help with the evaluation of reports and design on the design for the capstone tracker

4. Partial Implementation: We will begin the partial implementation of our project by starting to make our quality index and start implementation of basic features of the tracker.



REFERENCES/BIBLIOGRAPHY

[1] https://arxiv.org/abs/2306.00622.Ryan Liu and Nihar Shah,{ryanliu,nihars}@andrew.cmu.edu, Carnegie Mellon University

[2] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 9340–9351, Torino, Italia. ELRA and ICCL.

[3] https://doi.org/10.48550/arXiv.2010.05137,David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, Tom Goldstein

[4] https://doi.org/10.48550/arXiv.2410.14165,Chihang Wang, Yuxin Dong, Zhenhong Zhang, Ruotong Wang, Shuo Wang, Jiajing Chen

[5] https://doi.org/10.48550/arXiv.2406.16253,Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Jiayang Cheng, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, Wenpeng Yin

[6] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, MSc2, Leslie A.Lenert,MUSC,South Carolina, USA.

[7] D.M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, Zachi I. Attia,Fine-Tuning Large Language Models for Specialized Use Cases,Mayo Clinic Proceedings: Digital Health,Volume 3, Issue 1,2025,100184,ISSN 2949-7612,https://doi.org/10.1016/j.mcpdig.2024.11.005.(https://www.sciencedirect.com/science/article/pii/S2949761224001147)

[8] Milan Kostic1, Hans Friedrich Witschel2, Knut Hinkelmann1, 2, Maja Spahic-Bogdanovic1, 21University of Camerino (UNICAM)
2FHNW University of Applied Sciences and Arts Northwestern Switzerland
milan.kostic@unicam.it, {hansfriedrich.witschel, knut.hinkelmann, maja.spahic}@fhnw.ch

[9] Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

[10] Ryan Liu and Nihar Shah,{ryanliu, nihars}@andrew.cmu.edu,Carnegie Mellon University .

APPENDIX A DEFINITIONS, ACRONYMS, AND ABBREVIATIONS

Acronyms and Abbreviations:

XAI: Explainable Artificial Intelligence

CNN: Convolutional Neural Network

FLIR: Forward Looking InfraRed

MATLAB: MATrix LABoratory

PyTorch: Machine learning library for Python

TensorFlow: Open-source machine learning framework

GPU: Graphics Processing Unit

ICMR: Indian Council of Medical Research

LLM: Large Language Model

SIB: Swiss Institute of Business Administration

BERT: Bidirectional Encoder Representations from Transformers

PCC: Pearson Correlation Coefficient

RMSE: Root Mean Square Error

STEM: Science, Technology, Engineering, and Mathematics

IT: Information Technology

ROGUE: Recall-Oriented Understudy for Gisting Evaluation

Key Definitions

1. Capstone Project: A comprehensive academic project that demonstrates a student's accumulated knowledge and skills in their field of study, typically completed in the final year of an academic

program.

2. Large Language Model (LLM): An advanced AI model trained on vast amounts of text data, capable of understanding and generating human-like text across various domains.
3. Systematic Review: A structured method of collecting, analyzing, and synthesizing research findings from multiple sources to provide a comprehensive overview of a specific research topic.
4. Fine-Tuning: The process of adapting a pre-trained AI model to perform better on a specific task or domain by further training it on a specialized dataset.
5. Peer Review: A critical evaluation process where experts in a field assess the quality, validity, and significance of academic research before publication.
6. Quality Index: A standardized metric used to evaluate the overall quality and effectiveness of academic work based on predefined criteria.
7. Automated Evaluation: The use of computational techniques, particularly AI and machine learning, to assess and score academic or creative work with minimal human intervention.