Individual plan

PROJECT INFORMATION

Preliminary title, that indicates what the degree project will be about:
Optimizing LLM-Based Retrieval of Swedish Housing Association Reports

The name and e-mail address of the student:
Pooya Rayat, rayat@kth.se / prayat91@live.com

The name of the examiner at KTH:
Jonas Beskow

The name of the supervisor at KTH:
Jose Berengueres

The name and e-mail address of the supervisor, if the thesis is performed outside KTH:
Christopher Madsen, christopher.madsen@booli.se

Current date:
2025-01-28

Keywords:
Multimodal Large Language Model, Document Understanding, Automated Data Extraction

BACKGROUND & OBJECTIVE

Scientific and Societal Context

Unstructured data is prevalent across many industries, where companies often design their own document layouts, making standardized data extraction challenging. One such example is housing association inspection reports in the Swedish real estate market. While these reports contain structured information, their format varies significantly depending on the issuing organization. As a result, simple rule-based approaches, such as regex-based extraction, are insufficient.

From a societal perspective, automating information retrieval from such documents could reduce manual labor, free up human resources for more analytical and creative tasks, and enhance cost efficiency for businesses. By improving access to structured data, potential home buyers and real estate professionals can make better-informed decisions based on extracted insights.

Automating information retrieval reduces manual labor, reallocates human resources to analytical tasks, and improves cost efficiency. As open-source models continue to advance rapidly, an empirical evaluation of their performance in real-world document extraction tasks will provide valuable insights into their feasibility as cost-effective alternatives.

Organizational Interest

Booli is interested in automating its data extraction pipelines to reduce operational costs, improve throughput, and maintain high precision. The goal is to investigate both proprietary (e.g., GPT-4o) and open-source LLMs to see whether open-source models can match or approximate GPT-4o's performance in a more cost-effective and maintainable way. If open-source solutions can reliably achieve above 0.95 precision, and ideally approach or surpass 0.99, it would constitute an attractive alternative to proprietary APIs.

High-Level Objectives

The overarching goal is to devise a methodology for automating information retrieval from PDF documents using multimodal LLMs and semi-supervised learning. Booli envisions two potential outcomes of high value:

1. **Immediate Automation with Proprietary Models**: Leverage GPT-4o to achieve precision above 0.95, or preferably exceeding 0.99, thereby minimizing the need for manual intervention.
2. **Open-Source Alternatives**: Investigate whether fine-tuned open-source LLMs can replicate GPT-4o's precision levels in a manner that is cost-effective and operationally scalable

Data Description

The dataset used in this study consists of approximately 4,000 housing association reports in PDF format. These reports vary in structure; some are text-heavy, while others contain complex tables and financial data. The average length of a report is 10–20 pages, although some extend to over 50 pages. The key challenge is extracting structured information from these reports, which varies significantly in format and complexity. The primary fields of interest include:

Key Data Fields and Extraction Challenges

- Field Name: Example Values
- Value Range: Defined numerical or categorical limits
- Challenge: Level of complexity in extraction

Manual Labeling Requirement

To support fine-tuning and evaluation, manual labeling will be necessary. A subset of reports will be annotated to create a ground truth dataset, ensuring a reliable benchmark for model performance. This manual labeling will serve three key purposes:

1. **Training:** A high-quality subset will be used for fine-tuning open-source models to enhance extraction accuracy.
2. **Validation:** Labeled data will be used to measure model performance, ensuring precision and recall metrics are met.
3. **Benchmarking:** A manually verified test set will be established to provide a consistent evaluation framework for different models.

Risk of Incorrect Extraction

- **Low (1 - Minor Issue):** Incorrect extraction does not significantly impact decision-making. Example: Slightly incorrect postal address formatting.

- **Medium (2 - Needs Verification):** Extraction errors could mislead users but are unlikely to cause severe financial or safety risks. Example: Wrong inspection date might cause confusion but not major consequences.

- **High (3 - Critical for Decision-Making):** Incorrect extraction may lead to significant financial loss or health/safety risks.

Field Name	Example Values	Value Range	Challenge	Risk of Incorrect Extraction
Cadastral Designation	"Stockholm Fastighet 123"	Text string	Different formats across municipalities	Misidentification of property
Postal Address	"Storgatan 12, 111 22 Stockholm"	Text string	Addresses may be incomplete or formatted inconsistently	Incorrect location reference
Water Leakage	"No leakage reported" / "Leakage in basement"	Categorical (Yes/No)	Ambiguity in language used to describe leakage	Missing critical damage reports
Inspection Company	"Svenska Inspekt AB"	Text string	Multiple companies may be involved, unclear naming conventions	Misattribution of inspection
Inspection Date	"2023-05-15"	Date format (YYYY-MM-DD)	Incorrect date formats, missing values	Invalid reference for report validity
Expiration Date	"2024-05-15"	Date format (YYYY-MM-DD)	Assumed to be 1-year validity, but exceptions may exist	Risk of outdated assessments
Building Description	"5-floor apartment complex, built in 1985"	Text string	Highly variable descriptions	Structural misclassification
Energy Data	"Energy rating: B, Consumption: 120 kWh/m²"	Categorical + Numeric	Missing or inconsistent formatting	Incorrect energy classification
Last Renovation	"Kitchen renovated 2018, Bathroom 2021"	Year	Different fields for different renovations	Missed critical upgrade data
Radon Levels	"Measured at 40 Bq/m³"	Numeric (Bq/m³)	Some reports lack this information	Overlooking hazardous radon levels

Reliable extraction requires handling inconsistencies in structure, terminology, and formatting across reports. Below is an example thumbnail of a typical housing association report, illustrating the document format and layout complexity.

Sida 5 av 25

Varudeklarerat Säljare

Besiktningsutlåtande

1. Insamling av upplysningar och handlingar

Fastighetsägarna tillträdde huset år 2005.

2005 - Badrum övre plan renoverad.
2009 - 3-glas isolerfönster monterade (ej källare).
2008 - Dusch/wc entréplan renoverad.

2. Besiktning, analys av risker samt rekommendationer om fördjupande undersökningar

Utvändigt / Markförhållanden

Inget att notera.

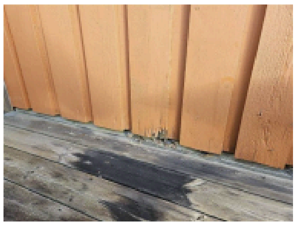
Utvändigt / Sockel

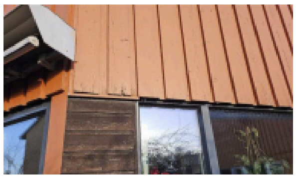
Inget att notera.

Utvändigt / Fasad

Enstaka panelbrädor har rötskador.

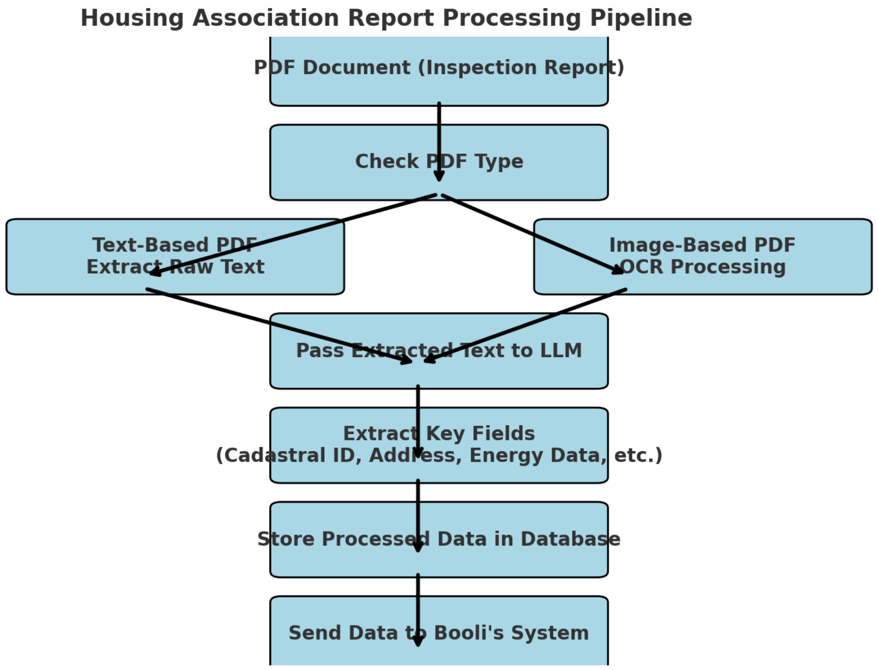
Fasadbeklädnad och träkonstruktioner där målningsbehov finns riskerar att få fukt- och rötskador på grund av sämre vattenavvisande förmåga. För att öka livslängden på fasadbeklädnaden rekommenderas målning och underhåll. I samband med ommålning rekommenderas att de rötskadade panelbrädorna byts ut.





Booli aims to automate key field extraction efficiently, maintaining high precision while reducing human verification efforts.

Data Pipeline



Evaluation of Accuracy and Precision

To assess model performance effectively, we must define precision, recall, and accuracy in the context of structured data extraction. Performance will be measured per field and across aggregated results:

1. **Field-Level Precision & Recall:**
   - Precision: Correctly extracted values / Total extracted values.
   - Recall: Correctly extracted values / Total actual values in dataset.
   - Example: If 90 out of 100 extracted postal addresses are correct, precision = 90%.
2. **Weighted Performance Metrics:**
   - Some fields, such as water leakage and radon levels, are more critical than others, like postal code formatting.
   - A weighted evaluation method will be considered, assigning higher importance to fields that impact homeowner decision-making.
3. **Handling Multiclass Predictions:**
   - Some fields (e.g., energy data, building description) involve classification rather than simple extraction.
   - In such cases, log-loss will be used as an additional metric where a probability distribution over classes is predicted.

This approach ensures that evaluation aligns with both technical performance metrics and business needs.

Required Background Knowledge

- **Multimodal Large Language Models (MLLMs):** Understanding how models like GPT-4o process and extract structured data from unstructured PDFs.
- **Information Retrieval & Document Understanding:** Applying NLP techniques to extract relevant content from real estate reports with varying layouts.
- **Optimization Techniques:** Exploring optimization techniques such as pseudo-labeling, Low-Rank Adaptation Techniques (LoRA), and more to iteratively improve model performance with minimal human annotation.
- **Performance Benchmarking:** Evaluating different models based on precision, computational cost, and practical deployment feasibility.
- **Data Pipeline Integration (low priority):** Design a workflow to extract, structure, and store retrieved data efficiently in a database or other structured format.

RESEARCH QUESTION & METHOD

Research Questions, Hypotheses, Objectives and Challenges

Research Question 1: How do proprietary models (e.g., GPT-4o-mini, GPT-o1) compare with open-source models (e.g., LLaMa 3.3) in extracting structured data from Swedish housing association reports, based on precision, recall, and extraction quality across key metrics?

Hypothesis 1: Proprietary models (e.g., GPT-4o-mini, GPT-o1) will achieve higher precision and recall in extracting structured data compared to open-source models (e.g., LLaMa 3.3) due to their extensive pre-training and resource advantages.

Research Question 2: What optimization techniques (prompt engineering, fine-tuning, semi-supervised learning) yield the highest precision improvements for MLLM-based document extraction? Given the lack of manually annotated datasets, pseudo-labeling will be explored as a method to generate training data for fine-tuning.

Hypothesis 2: Optimization techniques such as prompt engineering, fine-tuning with pseudo-labeling, and semi-supervised learning will improve extraction accuracy, with fine-tuning yielding the highest gains in precision.

Research Question 3: What are the cost-performance trade-offs between proprietary and open-source MLLMs for document extraction, and how do different deployment strategies (e.g., self-hosted, cloud-based APIs) affect scalability and feasibility?

Hypothesis 3: Proprietary models may offer superior accuracy, but open-source MLLMs optimized using semi-supervised learning could provide a more cost-effective alternative, considering infrastructure and computational expenses.

Research Questions Broken Down into Objectives and Challenges

RQ1 Objective 1: Establish a performance baseline for MLLMs

- **RQ1 O1 Task 1:** Select representative inspection reports and define key fields for extraction based on Booli's needs.
  - *Challenge:* Housing association reports come in varied formats, making it difficult to decide which reports should be representative without changing the performance evaluated.
- **RQ1 O1 Task 2:** Test GPT-4o and selected open-source MLLMs using zero-shot learning to establish baseline precision.
  - *Challenge:* Different models may interpret the same prompt differently, leading to inconsistent results.
- **RQ1 O1 Task 3:** Evaluate baseline performance using precision, recall, F1-score, and log-loss.
  - *Challenge:* Could be time-consuming to process and validate model outputs.

RQ2 Objective 1: Improve precision using optimization techniques

- **RQ2 O1 Task 1:** Experiment with prompt engineering (zero-shot vs. few-shot learning) to assess precision improvements.
  - *Challenge:* Small changes in wording can significantly impact results, making it difficult to determine the optimal prompt formulation.
- **RQ2 O1 Task 2:** Fine-tune open-source models using domain-specific reports and compare performance, leveraging pseudo-labeling to generate training data.
  - *Challenge:* The data consisting of housing association reports is not labeled, making fine-tuning less effective than with manually labeled data.
- **RQ2 O1 Task 3:** Implement semi-supervised learning (pseudo-labeling) and analyze its impact on model precision.
  - *Challenge:* If the model generates incorrect pseudo-labels, these errors can be reinforced in training, degrading overall performance.

RQ3 Objective 1: Identify and evaluate cost-performance trade-offs

- **RQ3 O1 Task 1:** Select a subset of alternative models to compare with GPT-4o based on precision, cost, and feasibility, balancing diversity and feasibility within the project timeframe.
  - *Challenge:* Finding the most promising models requires extensive literature review and testing.
- **RQ3 O1 Task 2:** Test the selected models in zero-shot and fine-tuned settings to establish their baseline precision and cost-effectiveness.
  - *Challenge:* Some models are non-deterministic, meaning comparisons may differ with the exact same input.
- **RQ3 O1 Task 3:** Compare performance vs. cost trade-offs across models by analyzing precision, recall, computational cost per inference (API pricing for GPT models vs. hosting cost for open-source models), and infrastructure feasibility (ease of integration into Booli's workflow).

- *Challenge:* While open-source models avoid API costs, they require expensive GPUs, which Booli doesn't provide.

RQ3 Objective 2: Optimize cost-performance trade-offs

- **RQ3 O2 Task 1:** Apply optimization techniques (prompt tuning, fine-tuning, LoRA) and evaluate impact on both precision and cost.
    - *Challenge:* Different models might have different methods to become optimized for our task.
- **RQ3 O2 Task 2:** Implement semi-supervised learning (pseudo-labeling) and measure its effect on extraction accuracy relative to cost.
    - *Challenge:* If the model generates incorrect pseudo-labels, these errors can be reinforced in training, degrading overall performance.

METHODOLOGY

Building on the recent advancements in multimodal document understanding and optimization techniques (Huang et al., 2022; Jin et al., 2024), this project proposes an integrated methodology to automate information retrieval from Swedish housing association reports using multimodal MLLMs. The approach is organized into several interrelated phases addressing data preparation, baseline evaluation, model optimization, comparative benchmarking, and real-world feasibility assessment.

Phase 1: Label data and data wrangling

A comprehensive dataset of housing association reports will be assembled from Booli's internal repositories and thousands of publicly available Swedish reports. Given the complexity introduced by heterogeneous document layouts, automated tools will be developed to convert PDF files into structured formats that preserve both text and layout features. An initial manual labeling pass will establish a reliable ground truth, ensuring subsequent evaluations are based on high-quality reference data.

Phase 2: Build a Baseline Model

Baseline performance will be established by evaluating a range of MLLMs, including both proprietary models (such as GPT-4o and GPT-o15) and open-source alternatives (e.g., LLaMA 3 and Mistral). Zero-shot and few-shot prompting strategies will be used as starting points, with model outputs assessed using precision, recall, and semantic similarity measures (Huang et al., 2022; Jin et al., 2024). This phase will provide an initial performance benchmark and identify areas requiring further optimization.

Phase 3: Optimize the Model

To enhance performance, targeted optimization strategies will be employed. DVarious prompt engineering techniques will be systematically explored in both zero-shot and few-shot contexts. Low-rank adaptation methods (LoRA), as proposed by Hu et al. (2021), will be applied to fine-tune selected open-source models on domain-specific data, improving extraction precision while managing computational costs. Complementarily, a pseudo-labeling pipeline following the guidelines of Xie et al. (2021) will iteratively incorporate the model's high-confidence predictions into the training set, with strict thresholding to mitigate error propagation.

Phase 4: Evaluate the Model

All models, both baseline and optimized, will be evaluated on a standardized test set under consistent conditions. Performance improvements from the optimization techniques will be quantified, and the financial viability of each approach will be analyzed using a flexible cost analysis framework tailored to our specific use case, with further refinement pending additional research.. This dual focus ensures that the methodology addresses both high precision in information extraction and the economic constraints and scalability challenges inherent in industrial deployment.

Phase 5: Create Pipeline

The optimized models will be integrated into a prototype pipeline tailored to Booli's operational environment. This phase includes designing interfaces that allow human oversight for low-confidence extractions, thereby balancing full automation with necessary manual intervention. Iterative feedback loops with Booli's technical team will enable continuous refinement of the system based on real-world performance data and operational requirements.

This holistic methodology, informed by the latest literature, aims to achieve a precision level exceeding 0.95 in automated document processing while ensuring that the solution is both cost-effective and scalable. Through advanced prompt engineering, efficient fine-tuning via LoRA, and robust pseudo-labeling techniques, the project aspires to set a new benchmark for automated information retrieval in the domain of document understanding.

ETHICS

Data Privacy and Compliance

This project processes housing association reports that are publicly available, meaning no personal or sensitive data is handled. However, data leakage risks and compliance with GDPR and the EU AI Act are still considered to ensure responsible AI use.

- **Data Leakage Considerations**: While the data is public, precautions are taken to prevent unintended memorization or inference-based disclosures that could aggregate structured financial trends beyond their original context. The model is not fine-tuned on full documents, only on extracted fields, reducing the risk of reconstruction. Additionally, query-based constraints ensure that outputs remain within predefined, structured data fields rather than generating unrestricted inferences.
- **GDPR and AI Act Compliance**: Since the data is non-personal, the project falls outside the high-risk AI system category under the EU AI Act. However, best practices in transparency, accountability, and bias mitigation will still be followed, as AI models used for information extraction may fall under minimal-risk AI applications,

which are currently unregulated but should still adhere to ethical standards (Future of Life Institute, 2024).
- **Booli's Internal Policies**: The project will align with Booli's data governance framework, ensuring that all data handling remains within the company's security and compliance guidelines. Regular reviews with Booli's technical team will confirm adherence to internal policies.

Impact of Automation on Employment

Automating document processing reduces the need for manual data entry but also creates new roles in AI system validation, oversight, and fine-tuning. Research (David & Reynolds, 2023) suggests that while some routine tasks may be displaced, new roles emerge in AI maintenance, compliance, and quality assurance.

LIMITATIONS

This study focuses specifically on evaluating multimodal large language models (MLLMs) for automated information retrieval from Swedish housing association inspection reports. The study evaluates both proprietary models (GPT-4o, GPT-4, GPT-3.5) and open-source models (e.g., LLaMA 3, Mistral) primarily on precision, recall, and cost-effectiveness. However, it does not provide an exhaustive comparison of all available MLLMs, as new models are continually emerging. Furthermore, the project does not cover end-to-end automation of data integration into Booli's systems but rather focuses on the extraction accuracy of relevant data.

Optimization techniques such as prompt engineering, fine-tuning, and semi-supervised learning will be explored to improve model precision. However, full-scale fine-tuning of large proprietary models is not feasible due to access restrictions and computational constraints; fine-tuning experiments will therefore be limited to open-source models with manageable resource requirements.

The cost analysis will focus on comparing costs between proprietary API-based solutions and self-hosted open-source alternatives. Long-term maintenance costs, infrastructure scaling, and real-world integration complexities are not deeply analyzed, as these extend beyond the scope of this academic study and would require long-term deployment data from Booli.

The study is further constrained by the availability and quality of housing association reports. Variations in report formatting and completeness may introduce challenges. Additionally, reliance on manual and semi-automated labeling to establish ground truth data means that large-scale annotation efforts will be limited by time constraints.

RISKS

Data Limitations

Variations in report formats and completeness may challenge model performance. Mitigation strategies include thorough preprocessing and synthetic data augmentation if necessary.

Computational Constraints

Fine-tuning large open-source models requires significant computational resources, which may be limited. To address this, priority will be given to methods that require lower computational cost than fine-tuning an entire model; such as prompt engineering and LoRA-based fine-tuning to optimize model performance without excessive resource consumption. Additionally, cloud-based compute resources will be explored when needed.

Time Constraints

The project timeline may restrict the number of models and techniques that can be tested. Prioritization of the most promising approaches will be essential to ensure timely completion.

Public Data and Data Leakage Considerations

Since the data used in this project is publicly available, there is no risk of private data leakage. However, the model's use will be monitored to prevent the unintended aggregation of structured information in ways that could generate sensitive insights beyond their original context.

EVALUATION & NEWS VALUE

Evaluation

This study demonstrates how multimodal large language models (MLLMs) can be leveraged to automate data extraction from Swedish housing association reports while balancing precision, cost-effectiveness, and scalability. The project evaluates GPT-4o against open-source models, comparing their accuracy, computational costs, and feasibility for real-world deployment. A cost analysis examines API expenses, infrastructure costs, and long-term maintenance, providing a clear assessment of whether open-source models can serve as viable alternatives to proprietary AI solutions.

Expected Scientific Results

For the first time, this study provides a comprehensive evaluation of how state-of-the-art MLLMs handle structured document extraction in the Swedish real estate sector. While previous research has assessed general document understanding, this project uniquely focuses on housing association reports, where tables, financial data, and complex layouts challenge traditional NLP models. The results also shed light on how open-source AI solutions can approach, or even rival, commercial models in real-world business applications.

Innovation/News Value

Together with Booli, we will explore how AI can streamline data extraction from publicly available housing association reports, a process that today involves significant manual work.

By evaluating both commercial and open-source AI models, the study investigates whether cost-effective AI solutions can perform at a level comparable to high-end proprietary models like GPT-4o.

The research provides valuable insights for industries that process large volumes of structured documents, such as real estate, finance, and legal services. By comparing precision, cost, and deployment feasibility, this study helps businesses determine whether they can reduce reliance on expensive AI providers and instead adopt self-hosted, open-source models.

This project not only demonstrates the potential of AI automation in real-world applications but also raises important discussions about data privacy, sustainability, and the future of cost-efficient AI adoption.

PRE-STUDY

This project builds on three key areas of research:

Document Understanding for AI-Driven Data Extraction

The study examines how multimodal large language models (MLLMs) process complex structured documents, such as Swedish housing association reports, which contain a mix of text, tables, and financial data. Prior research on document intelligence (Huang et al., 2022; Gu et al., 2023) explores how pre-trained transformers can improve information extraction accuracy in structured documents. Techniques like LayoutLMv3 and Donut (Kim et al., 2022) show promise for understanding text-layout relationships in PDFs.

Optimization Techniques for Performance Improvement

To enhance model performance while managing computational efficiency, the study explores prompt engineering (Jin et al., 2024), fine-tuning via LoRA (Hu et al., 2021), and semi-supervised learning approaches like pseudo-labeling (Xie et al., 2020). Research on adaptive fine-tuning methods (Touvron et al., 2023) suggests that domain-specific optimizations can significantly improve precision in structured data extraction while minimizing training overhead.

Evaluation Frameworks for Measuring AI Success

The study defines an evaluation pipeline that benchmarks models based on precision, recall, and cost-effectiveness. Existing frameworks, such as BERTScore (Zhang et al., 2020) for semantic similarity and entity-matching approaches (Nguyen et al., 2023) for structured field extraction, provide robust methodologies for assessing document understanding accuracy. Additionally, cost-effectiveness metrics (Brown et al., 2023) will help compare API-based inference vs. self-hosted deployment in real-world settings.

CONDITIONS & SCHEDULE

Resources

The project requires access to both proprietary and open-source MLLMs, as well as housing association reports for evaluation. Proprietary models (GPT-4o, GPT-4, GPT-3.5) will be accessed via OpenAI's API, while open-source models (e.g., LLaMA 3, Mistral) will be tested using Hugging Face's Transformers library. Cloud-based GPUs (provided by Booli) will be used for fine-tuning.

For evaluation, labeled datasets will be created through manual annotation and pseudo-labeling, with key fields in selected reports annotated. Additional test documents may be sourced from publicly available datasets. Evaluation will rely on metrics such as precision, recall, and semantic similarity measures, with input from Booli's data engineers to ensure alignment with real-world requirements.

External Supervisor Involvement

Christopher Madsen will provide external supervision and expertise. Weekly check ins on top of continual communication will serve as checkpoints for progress updates, scope adjustments, and problem-solving. His familiarity with Booli's current workflows will help ensure that any solution developed can be seamlessly integrated and maintained.

Timeline

The project is structured into five major phases over a total of 18 weeks, the rest of the time between week 18 and the final due date will be used as a buffer. Given that the first week is nearly complete, the individual plan will be finalized by the end of this weekend.

Milestone	Activities	Weeks	Completion Week
M1: Project Setup	Finalize individual plan	3	Week 8
M2: Literature Review & Data Preparation	Complete pre-study: literature review & methodology refinement	2	Week 10
M3: Baseline Model Evaluation	Data structuring & baseline model evaluation	3	Weeks 13
M4: Optimization & Fine-Tuning	Model training & optimization: fine-tuning, prompt tuning	3	Weeks 16
M5: Model Benchmarking & Cost Analysis	Final evaluation: cost-performance trade-offs, improvements	2	Weeks 18
M6: Thesis Writing & Final Evaluation	Thesis writing: analysis, discussion, conclusions	4	Weeks 22

REFERENCES

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774. (https://cdn.openai.com/papers/gpt-4.pdf, retrieved 2025-02-03)
2. David, D. A., & Reynolds, E. (2023). The work of the future: Building better jobs in an age of intelligent machines. Mit Press.
3. Future of Life Institute. (2024). High-level summary of the AI Act. (https://artificialintelligenceact.eu/high-level-summary/, retrieved 2025-02-05)
4. Huang, Z., Xu, Y., Yu, H., Chen, K., Xu, R., Huang, F., & Dai, W. (2022). LayoutLMv3: Pre-training for Document Understanding with Unified Text and Image Masking. arXiv preprint arXiv:2204.08387.
5. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.
6. Jin, C., Peng, H., Zhao, S., Wang, Z., Xu, W., Han, L., Zhao, J., Zhong, K., Rajasekaran, S., & Metaxas, D. N. (2024). APEER: Automatic Prompt Engineering Enhances Large Language Model Reranking. arXiv preprint arXiv: 2406.14449v1.
7. Touvron, H., Martin, J., Stone, K., Albert, P., & Jégou, H. (2023). Efficient Fine-Tuning of Large Language Models Using LoRA-based Techniques. arXiv preprint arXiv:2301.05650.
8. Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (202). Self-Training with Noisy Student Improves ImageNet Classification. Proceedings of CVPR.