Thesis Proposal for MPhil Degree

Student Name: Yiming Huang

ID Number: 50023052

Group Project: Connected Transportation Information System

Project Manager: Wenxun Hu

Project Supervisor: Wenxun Hu

Individual Project: Secure Artificial Intelligence System with Mechanistic Interpretability

Thesis Supervisor(s): Xinlei He

Student's Thrust & Hub: Data Science and Analytics & Infomation

Dec. 2024

The Hong Kong University of Science and Technology (Guangzhou)

Content

Part I Introduction to the Group Project

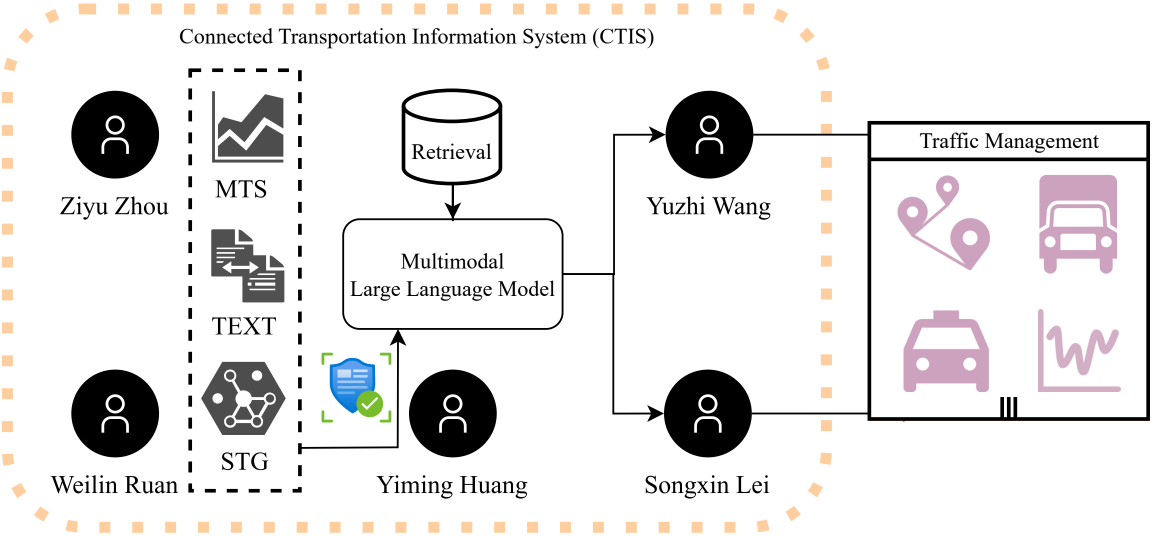

Figure 1: The outline of the composition and structure of our group project.

1.1 Background and Objective

The swift urbanization of modern cities has introduced many significant challenges for transportation infrastructures. With growing population density and rising vehicle usage (**which can be abstracted as united need to be allocated**), urban areas often struggle with issues like intense traffic congestion, extended delays, safety risks, and growing environmental hazards. These problems not only impair transportation efficiency but also threaten public safety and obstruct the achievement of urban development. Innovative solutions leveraging intelligent technologies are crucial to tackle these complex challenges.

The field of Intelligent Transportation Systems (ITS) has emerged as a powerful approach to managing better transportation. ITS combines sensing, communication, and data analysis technologies to create more flexible transportation ecosystems. By utilizing real-time traffic monitoring sensor data, predictive analytics, and connected vehicle technologies, ITS focuses on enhancing traffic efficiency, increasing safety standards, and supporting environmental sustainability. ITS encompasses various areas, such as traffic management, support for autonomous vehicles, infrastructure surveillance, and multimodal data integration to guide urban planning.

In line with the goals of ITS, our group project, the Connected Transportation Information System (CTIS), aims to develop an integrated framework that incorporates real-time traffic data, Vehicle-to-Everything (V2X) communication, and predictive analytics. This framework is designed to address critical challenges in urban transportation systems, offering novel solutions to optimize traffic flow, enhance road safety, and reduce environmental impacts.

Transportation networks face pressing challenges. Urban traffic demand has skyrocketed, leading to widespread congestion and delays, which disrupt daily commutes and incur significant economic costs. Safety concerns remain a pressing issue, with the lack of real-time hazard detection and preventive systems contributing to frequent accidents. Furthermore, suboptimal route planning and traffic management increase energy consumption and carbon emissions, worsening environmental damage. High-profile incidents, such as the catastrophic landslide on the Chaoyang section of the Meizhou Meida Expressway in May 2024 and the fatal Lixinsha Bridge collapse due to a container ship collision in February 2024, underscore the urgent need for intelligent systems to prevent such disasters and enhance emergency response mechanisms. Importantly, these events also highlight the necessity for advanced safety measures integrated into CTIS to ensure reliable performance across various scenarios.

The CTIS project employs a robust strategy to address these challenges. At the core of its approach is the integration of real-time communication and data-sharing systems, enabling seamless interactions between vehicles, infrastructure, and management centers. Predictive analytics and machine learning techniques such as large language models and model-based agent systems will be employed to analyze dynamic traffic conditions, identify potential hazards, and provide actionable insights to optimize transportation systems. V2X communication technologies will further enhance connectivity and support autonomous vehicle operations, while multimodal data fusion empowers urban planners to make strategic, data-driven decisions for long-term mobility improvements.

The CTIS initiative is structured to achieve the following key objectives:

1. **Real-time Traffic Optimization:** Develop cutting-edge algorithms for intelligent traffic management, aiming to alleviate congestion and reduce delays.

2. **Safety Enhancement:** Utilize V2X communication technologies along with real-time monitoring to improve hazard detection, accident prevention, and system reliability.

3. **Sustainability Promotion:** Facilitate optimized route planning and traffic management to significantly decrease energy consumption and lower carbon emissions.

4. **Urban Mobility Support:** Harness multimodal data fusion to generate actionable insights, empowering urban planners to tackle mobility challenges more effectively.

1.2 Significance

The CTIS project is a pioneering attempt designed to tackle urgent issues in urban transportation systems and to make a solid foundation for the advancement of Intelligent Transportation Systems (ITS). It stems from the application of seamless blending of sensing, communication, and data analytics technologies to create both intelligent and secure transportation networks. By combining these technologies, CTIS not only handles urban mobility challenges but also prepares for the needs of future, technology-driven, adaptable ubran computing paradigm.

The project's significance can be summarized as follows:

- **Reducing Traffic and Enhancing Efficiency:** Leveraging real-time traffic data and corresponding algorithms, CTIS boosts traffic flow optimization, cuts delays, and increases the operational efficiency of urban transportation systems.
- **Enhancing Road Safety:** By assembling connected vehicle technologies and proactive hazard detection, CTIS lowers accident risks and improves performance for all road users, including pedestrians and cyclists.
- **Promoting Environmental Sustainability:** Through energy-efficient traffic management and optimized route planning, the project decreases fuel consumption and carbon emissions, aiding global environmental conservation efforts.
- **Advancing Urban Planning:** CTIS offers urban planners with actionable, multimodal, data-driven insights, enabling wiser city designs and more resilient infrastructure planning for long-term urban mobility.
- **Driving Technological Innovation:** By advancing ITS research, CTIS encourages the development of scalable, adaptable solutions suitable for various urban environments, ensuring it stays at the forefront of transportation innovation.

Beyond addressing immediate concerns such as congestion and safety, CTIS introduces a scalable framework capable of adapting to changing urban demands. This progressive approach places CTIS as a transformative force in intelligent transportation management, tackling both current and future transportation challenges. By focusing on smarter traffic systems and sustainable urban mobility, CTIS contributes to a future characterized by increased adaptability, environmental responsibility, and technological progress in urban transportation.

1.3 Project Composition

The project consists of a series of interconnected individual projects, with each project contributing its own unique insights and technological advancements.

1. Weilin Ruan: **Retrieval-Augmented Universal Models for Spatio-Temporal Data**
   Weilin Ruan's project focuses on Retrieval-Augmented Universal Models for Spatio-Temporal Data, which aims to design a highly efficient framework for integrating and processing large-scale spatio-temporal datasets. With the rapid growth of urban data generated from diverse sources, such as satellite imagery, traffic sensors, and public transit records, analyzing such datasets in a unified and scalable way has become a significant challenge. This project addresses these challenges by leveraging retrieval-augmented techniques to enhance the performance and interpretability of spatio-temporal data models.

2. Ziyu Zhou: **Frequency-Enhanced Lightweight Framework for Multivariate Time Series Forecasting**
   Ziyu Zhou's individual project focuses on developing WaveTS, a lightweight, wavelet-based time series forecasting model tailored for traffic prediction within the CTIS. By decomposing multivariate traffic data into multiple frequency scales, WaveTS captures both global patterns and localized fluctuations, ensuring accurate and efficient forecasts. This adaptable approach enables seamless edge deployment, supports real-time congestion forecasting, and enhances safety measures, ultimately strengthening the CTIS framework and advancing time series forecasting methodologies in complex urban settings.

3. Yiming Huang: **Hallucination Detection and Mitigation, Robustness Evaluation, and Multi-Source Information Debiasing**
   Yiming Huang's individual research mainly concentrated on utilizing underlying mechanistic interpretability to develop a more comprehensive model securing technologies and evaluation methodologies. This personal research outcome indeed matches the safety securing need in the CTIS framework. Due to textual inputs and intermediate information in the CTIS system, the sub-project of Yiming's research will help more fine-grained hallucination detection and alleviation. Besides that, Yiming's research will provide novel AI system evaluation with a unified theoretical guarantee, which is important to eval the

CTIS system's performance. Yiming will also work on an AI debiasing project, which will help CTIS framework get rid of biased single-source information dependency.

4. Yuzhi Wang: **Large Language Model Enhanced Urban Agent Simulation and Application**
   Yuzhi's project aims to leverage large language models (LLMs) to enhance individual decision-making behaviors in city simulation tasks, especially in agent-based models (ABMs). Through LLM technologies like prompt, retrieval, generation and fine-tuning, the human-centered agent behaviors in the urban environment, e.g. mobility, economic and social interaction are generated, to provide micro-level research perspective to various urban evolution and optimization, such as point of interest (POI), land-use, and transportation infrastructures.

5. Songxin Lei: **Collaborative Public Resource Allocation: A Spatio-temporal Feature Extraction and Potential Game-Based Reinforcement Learning Framework**
   Songxin's work ensures that mobile public resources, such as intelligent trash bins or delivery stations, are effectively deployed to maximize coverage and meet dynamic demand. Additionally, the research investigates the collaborative relationships among resources, providing theoretical guarantees through the use of potential game theory. By integrating with spatio-temporal neural networks for feature extraction and constructing a robust reward function, this project delivers actionable strategies for intelligent decision-making. The outcomes contribute to the development of a scalable, adaptive, and intelligent transportation system, offering valuable insights for urban planners and policymakers aiming to enhance urban mobility and resource efficiency.

These interconnected projects collectively create a holistic and integrated approach to tackling the obstacles of incorporating electric vehicles into urban settings, all the while promoting sustainability, efficiency, and safety in urban mobility.

1.4 Project Connections

Each individual project within the Connected Transportation Information System interlinks to create a cohesive and efficient system:

- **Weilin's Predictive Analysis:** Weilin Ruan's project plays a pivotal role in the overall synergy of the Connected Transportation Information System (CTIS). The retrieval-augmented framework developed in this project not only enhances the efficiency of integrating and processing large-scale spatio-temporal datasets but also serves as a key enabler for many other sub-projects within CTIS. Specifically, Weilin's work dynamically retrieves and integrates relevant historical and real-time data, providing a solid foundation for downstream tasks. The retrieval-augmented framework seamlessly integrates with Ziyu Zhou's temporal models by supplying contextually relevant historical data, such as traffic patterns under similar conditions. This significantly improves the accuracy and robustness of multivariate temporal predictions while reducing uncertainty in dynamic urban environments. Moreover, the spatio-temporal graph representations generated by Weilin's project act as critical inputs for Songxin Lei's intelligent infrastructure planning. These representations enable more effective optimization of connected infrastructure layouts, such as smart traffic lights and vehicle-to-infrastructure (V2I) communication systems. Additionally, the project supports Yuzhi Wang's research on autonomous vehicle trajectory predictions by providing pre-trained universal models with strong cross-city generalization capabilities. Using these models, Yuzhi's system can better predict interactions between vehicles and pedestrians, improving the safety and reliability of autonomous navigation.

- **Ziyu's Strategic Deployment:** Ziyu's individual project focuses on modeling multivariate time series (MTS in Fig. 1) data within the CTIS ecosystem. Ziyu's predictions enrich the overall data pool that includes textual insights, spatial-temporal graphs (STG), and other sensor-derived metrics. As a result, the model's refined time series forecasts serve as a vital input stream for subsequent tasks—such as route optimization, demand forecasting for shared mobility services, or resource allocation for charging stations—handled by Yuzhi, Songxin, and other team members. In this manner, Ziyu's contributions ensure that the entire CTIS framework benefits from accurate, timely MTS forecasts, enhancing the decision-making capabilities and responsiveness of the interconnected urban transportation network.

- **Yiming's Securing Measures:** To be specific, there are at least three aspects of safety guarantee in the project: 1) By passing LLM-extracted text information to different modules in CTIS, it unavoidably brings LLM inherent hallucination problem. One of Yiming's individual research aligns with this problem. It will help Weilin, Ziyu, and Yuzhi to reduce the hallucination of their modules' input or output. 2) To verify whether the framework is robust to noise in any kind of input data or middle intermediate information, Yiming Huang's individual research will provide the method. It aims to check the system output of Songxin's and Yuzhi's modules. 3) Due to multi-source input information and intermediate information works in the CTIS framework, it is important to ensure the framework does not rely too much on specific parts of information which results in a biased CTIS framework and wasted other kinds of information. In this aspect, Yiming will coordinate with the whole team.

- **Yuzhi's Simulation Insights:** The LLM-based urban agent simulation proposed by Yuzhi constructs a digital twin environment for CTIS from a new generative agent-based model (GABM) perspective. With the help of scalable and interactive dynamic urban simulation environments generated by LLMs, Yuzhi's work can make good use of teammates' spatio-temporal data, especially trajectory data to build agents. Besides, the simulator can be used to further provide guidance for the optimization and future evolution of transportation information systems in urban environments.
- **Songxin's Allocation Strategies:** Songxin's research serves as a critical fine-grained decision-making component within the Connected Transportation Information System. By leveraging the predictive results provided by Ziyu and Weilin, i.e., time-series population flow predictions and spatio-temporal graph-based forecasts, my work develops deployment strategies for mobile public resources. These strategies are optimized to interact with the environment, translating predictive insights into actionable decisions that maximize real-world impact.

1.5 Project Milestones

The development of the Connected Transportation Information System (CTIS) is divided into four key phases, each designed to ensure the seamless integration of individual sub-projects into a unified, scalable system. These milestones focus on data collection, model development, collaborative integration, platform testing, and deployment.

1. **Individual Data Collection and Model Development (Months 0-3)**
    - **Focus:** At this stage, each team member focuses on collecting and processing data relevant to their specific sub-project within CTIS.

- **Key Activities:**
  - Developing initial models and algorithms for tasks such as:
    - Predictive traffic flow modeling and congestion analysis.
    - Optimizing charging infrastructure for electric vehicles.
    - Scheduling algorithms for shared-mobility services.
    - Trajectory predictions for autonomous vehicle behavior.
    - Spatio-temporal urban analysis for transportation planning.
  - Refining datasets to ensure consistency and compatibility across sub-projects.

2. **Collaborative Data Integration and Model Alignment (Months 3-6)**
   - **Focus:** This phase emphasizes collaboration among team members to integrate individual datasets and align models within the CTIS framework.
   - **Key Activities:**
     - Sharing datasets and combining insights from individual sub-projects, such as traffic flow modeling, urban analysis, and autonomous behavior predictions.
     - Synchronizing predictive models, optimization techniques, and scheduling algorithms to ensure smooth data flow and compatibility within the system.
     - Establishing a unified data pipeline to handle multimodal spatio-temporal data efficiently.

3. **Holistic Platform Integration and Initial Testing (Months 6-9)**
   - **Full Platform Integration:**
     - Combining components developed during earlier phases into a unified CTIS platform.
     - Consolidating features such as traffic flow prediction, charging infrastructure optimization, shared mobility scheduling, autonomous behavior analysis, and urban transportation insights into a cohesive system.
   - **Initial Testing and Evaluation:**
     - Conducting initial testing in simulated environments to ensure the integrated platform functions cohesively.

    - Testing the platform in diverse urban scenarios to evaluate its adaptability and robustness.

4. **Comprehensive Deployment and Future Planning (Months 9-12)**

   - **Deployment in Selected Areas:**
     - Rolling out the integrated CTIS platform in selected urban areas as pilot projects.
     - Collecting feedback and performance metrics to refine system components and improve operational efficiency.
   - **Citywide or Multi-Area Deployment:**
     - Scaling up the system for full deployment across cities or regions based on the success of pilot implementations.
     - Ensuring the platform's scalability to support diverse urban mobility challenges.
   - **Project Summary and Future Development:**
     - Compiling a comprehensive project report to document key achievements, challenges encountered, and lessons learned.
     - Proposing future development plans to enhance the CTIS platform, incorporating emerging technologies and addressing evolving urban mobility needs.

Part II Proposal of the Individual Project

2.1 Significance and Relevance of the Individual Project to the Group Project

2.1.1 Complementary Role

Yiming Huang's individual project mainly addresses the safety and robustness challenges of the group project. By focusing on hallucination detection and mitigation, robustness evaluation, and multi-source information debiasing, this work fills the gap between securing reliable and unbiased information propagation within the CTIS framework of the group project. These contributions ensure that the textual and intermediate information, pivotal to the group's objectives, remain accurate, effective, and unbiased.

2.1.2 Value Addition

The individual project adds concrete value to the group effort in several ways:

- **Hallucination Detection and Mitigation:** Provides fine-grained mechanistic perspective methods to reduce input and output hallucinations, enhancing the reliability of modules developed by Weilin, Ziyu, Songxin, and Yuzhi.
- **Robustness Evaluation:** Introduces novel evaluation methodologies to test the system's resilience against noisy or inaccurate data, benefiting modules by Songxin and Yuzhi.
- **Multi-Source Information Debiasing:** Develops methods to avoid over-dependency on single-source information, enhancing the efficiency and security of the CTIS framework.

2.1.3 Interdependence

The success of the group project is closely related to the contributions of Yiming Huang's research. Collaboration and coordination are necessary, as the proposed securing measures will directly support other members by:

- Reducing hallucination in both intermediate input and output for other modules.
- Ensuring robust system performance over data noises, which is critical for the entire CTIS framework.
- Enabling security and efficient use of multi-source data, enhancing all sub-modules, and ensuring a cohesive and unbiased framework.

However, Yiming Huang's research can start and develop without any progress of other members' research, that is, the proposal plans to produce effects in general scenarios of AI safety and then apply the research findings to the group project. This individual plan allows parallelism in the group project.

2.2 Statement of the Individual Project in Details

2.2.1 Literature/Market Review and Problem Definition

With the rapid development of advanced AI systems, such as ChatGPT [1], safety issues draw much more concern day by day. Many scenarios urge researchers and developers to alleviate these problems, such as jailbreaking prompts hijacking the Large Language Model (LLM), adversarial inputs inducing diffusion models for harmful image generation, decision-making algorithms giving unfair judgments, etc. These scenarios vary in the context of thousands of AI-based, especially LLM-based applications. Meanwhile, from the defense side, there are many valid protection methods as adversaries to defend AI systems or applications from attacks. However, these technologies do not achieve unified effectiveness toward a more generalized defense form with a certain convincing safety guarantee [2]. In other words, current work in the AI safety domain is more focused on specific categories of attack and lacks the theory to reveal the mechanism behind the attack and the defense that inherently results in more or less trustworthiness. To dive deeper into the mystery behind these both diversified and contextualized "swords" and "shields", mechanistic interpretability works in the explainable AI (XAI) field point out the way. These findings model the internal working system within transformers, diffusion models, and even agentic AI systems. Based on these analytics, the research will be able to gain the underlying principles for controlling the model acting toward safety. Therefore, it comes to the goal of my M.Phil. research: ***secure artificial intelligence systems with mechanistic interpretability.***

For a more comprehensive understanding of my research goal, the proposal does ample literature review as following demonstrations. On the one hand, it is significant to review the methodologies in the AI safety field. As for taxonomies, this proposal regards four main categorial challenges in the current AI safety domain according to existing surveys [3, 2]: 1. **Data & Input Safety**. 2. **Model Safety**. 3. **Output Safety**. 4. **Human-In-Loop Safety**. As Figure 2 shows, this proposal gives the basic demonstration as statements below:

1. **Data & Input Safety**. At this level, attack and defense technologies mainly focus on

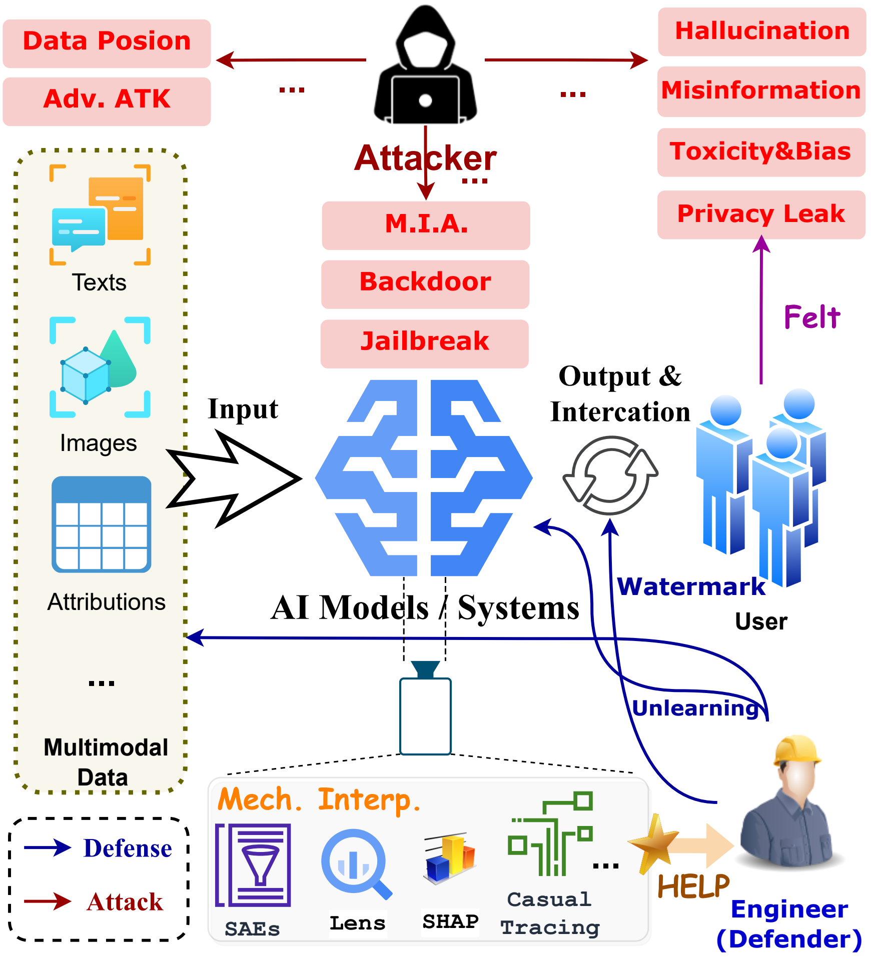

Figure 2: Illustration of Overall Proposal Research Question.

manipulating the subsets of datasets and perturbating the inputs. For detailed division, we listed these 2 kinds of safety topics (which may have overlaps between each other, but they own their unique features, the following proposal is the same): (1) **Adversarial Examples**. Adversarial examples are classical machine learning attacking methods [4], this method mainly adds diverse noise to normal input to mislead machine learning models giving wrong classification or regression results. Classical works FGSM [5] and PGD [6] start the attempt, and the afterward works develop diversely in different scenes. (2) **Data Poisoning & Backdoors**. Data poisoning and backdoors are realized by selecting a ratio of poisoned data subsets among formal datasets [7, 8]. The poisoning means are various, including but not only restricted to adding noises to data, rotating image inputs, and inserting weird prefixes into texts, these means poison benign samples as triggers. When training is done, benign inputs belonging to the formal distribution just work as expected however the triggered inputs in the poisoned distribution will trigger the model to collapse such as giving wrong statements or classification results. The most recent works like VL-Trojan [9] and BadAgent [10] successfully utilize the complexity in the application context thus making the more covert composited triggers.

2. **Model Safety**. In this aspect, the AI safety issues lie in the learned parameters which cause the ill pattern transformation inside the model and finally produce unsafe effects. These issues can be further divided into: (1) **Alignment.** Alignment is a key terminology in the wide LLM community, representing consistency with human corpus even human reaction [11, 12]. It requires a high-level synchronization of reasoning ability and core value between humans and LLM. If it fails to make it, misalignment will let the model react to depart from users' expectations. It is usually carried out post-training or extra test time enhancement. (2) **Privacy Leaking**. Due to the failed filter in pretraining-used data, sensitive privacy information is leaked and learned by the parameters of the model. A common solution is **Unlearning**. Unlearning is a typical problem in the wide machine-learning community, it is highly related to solving the model's privacy leaking risk [13]. As the "un-" prefix represents, unlearning eliminates privacy information encoded in learned parameters.

3. **Output Safety**. Discussions about safety issues over the model's output contents are widespread. These discussions are basically sorted into varieties below: (1) **Hallucination**. Just like human hallucination, hallucination is the phenomenon in which AI outputs untruthful information that conflicts with the world's knowledge or context [14, 15, 16]. Current hallucination mitigation technology mainly focuses on reducing the uncertainty of content generation and leading model generation toward truthful subspace, like DoLa [17] and ITI [18]. (2) **Fake Information**. Fake information is a general topic that involves wide practical application of AI, such as AI-generated fake news, AI-face-changing-video, and unreal factual statements from AI. The detection and mitigation methods are also diversified. (3) **Toxicity & Biases**. Harmful contents are often the consequences of insecure model inference. They are inherent properties of LLM's cognitive biases [19], which need to be eliminated. (4) **Algorithm Fairness**. Usually, many algorithms are used in our daily life decision-making, and the significance of ensuring its fairness is always the first consideration. Current work gives many solid mathmatical guarantees, however, these research outcomes are mainly applicable to the classification or regression paradigm model, not the generation paradigm model [20]. (5) **AI-Generated Contents Detection**. AI-generated content (AIGC) Detection is a wide-acknowledged AI safety issue in current applications, the most two common problems are **DeepFake** image/video identification & **LLM**

generated text detection. These problems escalated by their powerful application which offers sufficiently real outcomes fused human discriminative system. To tackle problems represented by these two issues, **Watermark** technology achieved certain effects by adding and detecting fine-grained imperceptible differences [21, 22]. Such as comparing human-eye-invisible perturbation or logit-based sequence distribution distinguishing.

4. **Human-In-Loop Safety**. Another common stage of AI safety is the danger generated in the human-computer-interaction loop. This problem does not simply occur in the once single-direction input-inference-output journey of the AI model, but the interaction between user and models. (1) **Jailbreak**. Jailbreak attacks are one of the most severe threats to LLMs, it hijacks the LLM by carefully crafted malicious prompts in multi-turn conversations. The basis of those tactics is the instruction following the ability of modern LLM, as sophisticated prompts and instruction find the sycophancy area of the model then attackers are able to elicit models to generate harmful responses [23]. AutoDAN [24] and GCG [25] are recent widely recognized methods. (2) **Membership Inference Attacks**. Membership Inference Attack (MIA) focused on maliciously speculating related also sensitive attributes of input data by the model's output. This information leak often is involved in privacy issues [26]. (3) **Model Extraction Attack**. Model extraction attack (MEA) aims to clone the model in a black-box manner, usually in cloud computing scenarios. Through querying the model's response in acceptable complexities, the attacker approximately gets the valuable capabilities of the model that only can be achieved by training in private data [27]. The successful attack will destroy the formal owner's exclusiveness in market competition, and also provide a reference model for MIA. (4) **Data Containmation**. Data contamination refers to the training datasets contaminated by leaked test data that bring fake high performance in evaluation [28, 29] and unexpected output contents. The phenomenon happened both intentionally and unintentionally but hurt a lot to rightfully evaluate the model, resulting in instability in calibrating the trustworthiness of the model or system.

On the other hand, this proposal stems from interpretability thus the interpretable/explainable AI literature is also demonstrated as follows. Overall, these works gradually step into the deeper area of interpretability which achieves a more transparent AI system. To summarize, there are four types of interpretability works, respectively, they are 1.**Behavioral Explainability**, 2.**Attributional Explainability & Interpretability**, 3.**Concept-based Interpretability**, and 4.**Mechanistic Interpretability**. Here, this proposal briefly introduces these four types of interpretability work.

1. **Behaivoural Explainability**. From the behavioral level, it is better to use the terminology "Explainability" to summarize current works at this level. The reason behind this taxonomy is that these methods lack the revealing of intrinsic features inside the models. Famous methods like LIME [30] and SHAP [31] show the mutual traits of this type of method — Model Agnostic. Usually, these methods develop sensitivity and perturbation analysis by examining input-output relations to assess the model's specific or general data dependency, which gives a clear explanation of the model's behavior. However, its model-agnostic nature coincidentally turns to its drawbacks — lack of the insight to rediscover the model's internal decision process and more fine-grained causal findings that not only explain why certain inputs influence certain outputs.

2. **Attributional Explainability/Interpretability**. As for this kind of explainability, studies mostly provide tracing means to unravel the mystery of the different contributions to the final output of different parts or units. There are a lot of famous methods in line with the direction of this paradigm, like LRP [32], Grad-CAM [33], DeepLIFT [34], Attention Rollout [35]. These methods are developed based on certain inherent properties like a gradient or neuron activation value of model architecture which means they are usually model-specific. Thus, transparency is enhanced by showing how features interact with each other and finally, interactions form the contributions. However, these methods mainly collect hand-crafted saliency like the activation value multiplying the gradients which have no underlying theory to back its efficiency, not surprisingly, the contributions accumulated by them are always distorted to some extent.

3. **Concept-based Interpretability**. In this paradigm, interpretability is a top-down approach that starts with finding representations, structures, and components of models consistent with the high-level concept of humans. The well-known works like concept bottlenecks [36] and concept activation vectors [37] successfully convert internal representation within the model to comprehensive human concepts. however, these concepts are too sparse and ignore the polysemantic and semantic composition [38, 39]in models thus their usage is limited.

4. **Mechanistic Interpretability** (MI). Compared to concept-based, mechanistic interpretability works as the bottom-up reverse engineering measure to thoroughly investigate fundamental components of models. These granular analyses of internal features, activations, layers, and different functional components offer an intrinsic view to rethink the operational mechanics of AI models. Currently, there are several styles that gain acknowledgment in the XAI community, like **sparse autoencoders (SAEs)** [40] reconstruct inner respresentations by sparse and

interpretable dictionary learning, causal tracing style **activation patching** [41, 42] methods using sufficient runs to calculate direct effects and indirect effects and **logit lens** [43, 44] early unembed the logits in middle layers. Studies related to mechanistic interpretability offer abundant insights into the inner workings of AI models, which are also the potential principles of solutions for AI safety.

As the interpretability work offers the views to reveal the intrinsic features of AI models, nowadays community has carried out some meaningful attempts at interpretability-guided securing or attacking technologies. For instance, using SAEs to control models [45], representation engineering (RepE) [46], eliciting control theory to steer the decoding [47] and adjusting attention map to make better instruction-following [48]. These works primarily attain a certain degree of success for better and safer AI applications, which indicates the promising perspective of this direction.

Over reviewing current works, existing studies point out that safety issues are gradually severe, and mechanistic interpretability is a reliable theoretical basis to be utilized for concretizing the cure for hazards in AI systems. To be direct, that is the **problem definition of this proposal:** ***How to use interpretability findings to both theoretically and generally solve certain kinds of AI safety problems***.

2.2.2 Objective and Scope of the Project

After a detailed literature review of AI safety issues, AI interpretability research progress, and current novel interpretability-guided securing technologies, the above section further stresses the urge and benefits of securing AI with mechanistic interpretability. Subsequently, this proposal clarifies the objective and scope of the research plan here. Generally, there are three research sub-domains and related goals: 1. **Pure Mechanistic Interpretability**, 2. **Genearal Mechanistic Interpretability for AI safety**, 3. **Domain-Specific MI & Safe AI**.

1. **Pure Mechanistic Interpretability**. In the first place, this proposal attaches great importance to studying the general interpretability of most advanced AI models and systems such as LLMs and Agents. The research scope and related objective involved:

- Research will be carried out exploring subtly different roles about how diverse elements of AI form and continue their inner working toward intelligent actions.
- This research also studies the alignment and misalignment between humans and humans'

best mimickers — AI, we will find whether the so-called "semantics" is a kind of co-occurrence of corpus brought reciting ability or abstract abilities over reasoning path.

- The research will refresh the theoretical basis of interpretability by cross-disciplined combination. the proposal here admits mechanistic interpretability can only go far with referring viewpoints from other disciplines, such as differential equations showing the learning dynamics of trustworthiness, control theory guaranteeing the feedback, experimental cognitive science inspiring the design of empirical methods, and psychological knowledge illustrating some traits of human's spiritual perception.
- The research will also design the faithfulness evaluations to judge whether the interpretability outcomes are consistent with the original results.

In all, pure mechanistic interpretability concentrates on finding novel and holistic views to unravel the inner workings of AI models and systems.

2. **Genearal Mechanistic Interpretability for AI safety**. As mechanistic interpretability findings of our research and others' research offer meaningful guidance to rediscover internal safety, it undoubtedly becomes the central theme of future research planned in this proposal. This core research goal ranges in different AI safety stages abovementioned, the research will be related to:

- Using mechanistic insights to redesign current AI models and systems, like adding light but efficient modules or collecting common features through a few forward passes, to enhance the tenacity of the model about various attacks.
- Finding weaknesses of the current model from the mechanistic perspective and correspondingly creating targeted attack technologies, both white-box and black-box.
- This proposal plans to seek fair metrics and benchmarks to comprehensively compare our method to others.

the proposal here foresees that future research will mainly focus on the defense side, however it will also involve the attack side.

3. **Domain-Specific Mechanistic Interpretability & Safe AI**. Execpts general research on trustworthy AI, the research will also pay attention to some specific applications. The possible research areas are:

- AI safety in intelligent transportation systems and urban computing. Since the group project mainly studies the connected intelligent transportation system (CTIS), it provides context to apply potential general findings in the context. Firstly, with the wide usage of LLM in current CTIS, it is necessary to tackle the hallucination in transportation descriptive text especially some key factors like direction and velocity. Meanwhile, due to multi-source information fusion applied in CTIS, judging whether the system is dependent on single-source information and evaluating the robustness of the system are also significant.

- Societal AI. The most impactful AI safety issues mainly occur in the Human-Computer-Interaction stage, societal influence-related safety issues will be a good point to make general securing methods more fine-grained and down-to-earth. Many research questions are worthwhile to study, for instance, human & LLM's cognitive bias, cultural differences in LLMs, etc.

- Other vertical areas need the precision and stability of AI. For instance, law judgment, medical diagnosis, and auto-driving, these fields yearn for the guarantee of AI safety.

2.2.3 Research Method and Justification

The proposed research method includes but is not limited to the following methods:

- **Literature reviewing**. The study of AI safety needs to catch up with the general AI research and application. There are 4 types of literature that need to pay attention to: 1) Cutting-edge general AI methods and applications, for the moment of writing this proposal, LLM reasoning capability is the focus of the research community. 2) Mechanistic Interpretability research, the XAI community recently noticed this sub-area, and it shows a promising future for understanding safer AI. 3) AI safety topic, many typical AI safety research still active today, and many new technologies are introduced by researchers. 4) Mechanistic Interpretability for AI safety, this paradigm is growing, as this proposal demonstrates.

- **Threat Analytics and Modeling**. As continuous new AI applications of specific domains, this proposal calls for the same continuous concern about analyzing and modeling appeared or potential threats. As for potential threats, the analysis will be carried out from a user perspective and a general prediction perspective, the former speculates risks closely related to every procedure, and the latter estimates risks from a general form and known attacks in similar scenes. Then, the threats will be modeled, the research will assume the role of the attacker, the success rate of attacks, and possible user reactions, and the modeling outcome will help the solution be more targeted.

- **Mechanistic Modeling**. Besides threats toward AI systems, it is also important to translate how these threats become the input of the model and the process it changes inside the model. Many pilot studies will be used to probe certain traits of these abnormal reactions. Based on that, the mechanistic view of these unsafe inputs is established.

- **Create novel algorithm, strategy, or framework for attack/defense**. Given detailed threat analytics and the mechanistic understanding of the inner workings of the model, the research aims to create novel and tailored attack/defense technologies as the solution.

- **Test & Eavluation**. Evaluation will involve many targets: 1) The research will develop red-teaming technology to test both our enhanced models and common models. 2) The research plans to collect data and make related metrics to benchmark the trustworthiness of different models under different attack/defense technologies 3) As for our attack/defense method, new evaluations will be introduced, and research will also use them to test other baselines for fair comparison.

- **User study**. In order to evaluate the explanation/attack/defense technology more accurately, the means utilmately comes to user study. The research plans to crowdsource the evaluation with a carefully designed questionnaire. The questions are about the subject of direct or indirect quality or truthfulness of output from anonymized models. The result will also be strictly filtered and computed.

- **Cross-Disciplined Thinking and Collaboration**. As above mentioned, the mechanistic understanding of the inner workings of AI models and systems will be rebuilt on reliable theories and these theories are mainly referred from other disciplines. Therefore, the research will be executed on the condition of cross-disciplined thinking and collaboration.

2.2.4 Execution Plan

Phase 1: 2024.9 $\sim$ 2025.4 – Primary Research about MI for Safer In this phase, also the writing moment of this proposal, projects mainly aim to build the basis for future research

Model	Factual				Negotiation			
	Different	Related	**Casual**	Valid	Different	Related	**Casual**	Valid
Llama-3.2-1B-Instruct	14	17	**9**	27	3	4	**3**	6
Llama-3.1-8B-Instruct	9	7	**3**	27	1	0	**0**	5
Qwen2.5-0.5B-Instruct	12	6	**2**	27	2	3	**1**	6
Qwen2.5-7B-Instruct	11	6	**2**	26	2	2	**1**	6
Mistral-7B-Instruct-v0.3	11	6	**2**	27	2	1	**0**	6
Phi-3.5-mini-Instruct	11	6	**2**	27	1	3	**1**	6
	Open				Legal			
	Different	Related	**Casual**	Valid	Different	Related	**Casual**	Valid
Llama-3.2-1B-Instruct	6	7	**3**	9	5	6	**3**	7
Llama-3.1-8B-Instruct	3	4	**1**	8	4	2	**0**	8
Qwen2.5-0.5B-Instruct	4	3	**2**	8	4	5	**1**	7
Qwen2.5-7B-Instruct	4	4	**1**	8	5	3	**0**	7
Mistral-7B-Instruct-v0.3	7	5	**1**	8	3	2	**0**	8
Phi-3.5-mini-Instruct	5	7	**0**	9	3	3	**2**	7

Table 2: Anchoring bias test results of different models across four datasets: Factual, Negotiation, Open, and Legal. "Causal" represents the amount of strong anchoring questions.

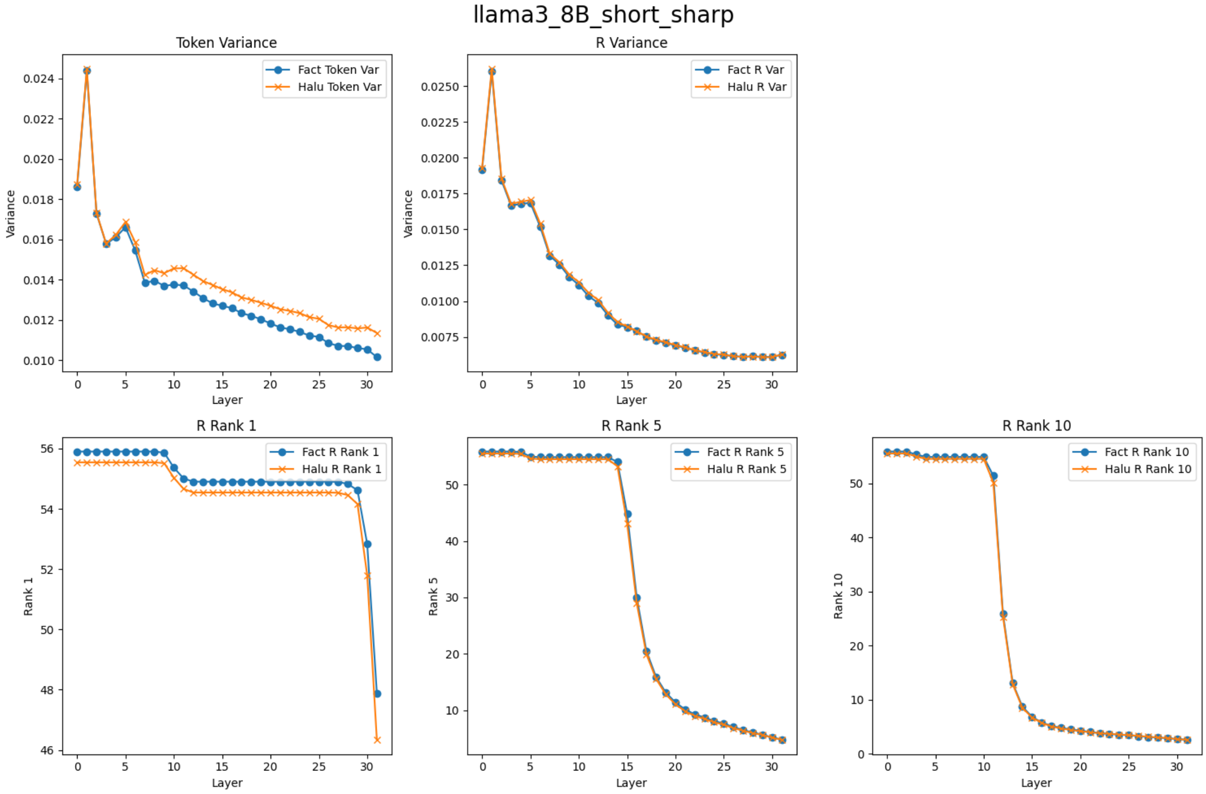

Figure 3: Specific statics difference Attention Rollout outcomes between hallucinated and factual samples.

and provide the most needed technology for group projects. Therefore, there are two concrete research projects in this phase. The first is **LLM hallucination detection and mitigation with**

mechanistic interpretability. This project aims to propose a novel hallucination indicator that helps both detection and hallucination of LLM. Some pilot studies about the information propagation difference of hallucinated samples (by altering Attention Rollout [35]) primarily proved the rightness of this direction, as Figure 3 shows. The planned attempt will concentrate on balancing the effects of token propagation and enrichment. The Second is **Anchor bias in LLMs**. This project aims to explore three research questions: 1) Does the modern LLM have an anchoring bias? 2) If so, what is the mechanism behind it? 3) The possible mitigations. As far as the current research outcome is concerned, it can answer the first question: It does, but few. Table 2, Figure 4, and Figure 5 demonstrates periodical research outcomes.

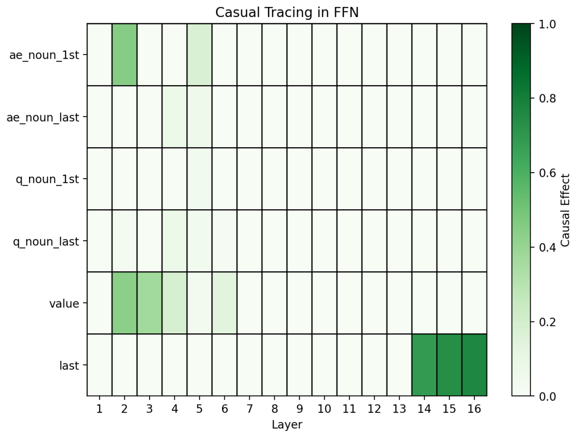

Figure 4: Anchoring bias causal tracing of FFN layers, Llama3.2-1B-Instruct.

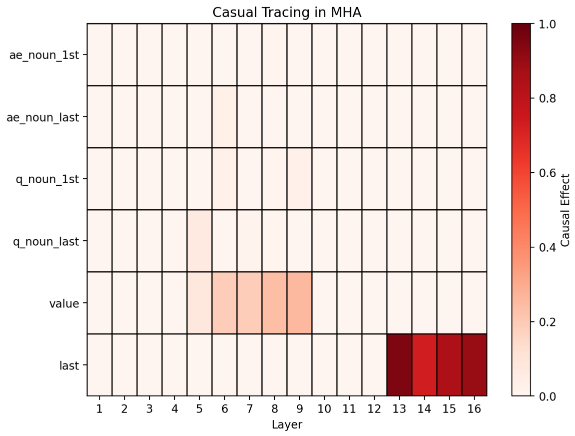

Figure 5: Anchoring bias causal tracing of MHA layers, Llama3.2-1B-Instruct.

Phase 2: 2025.5 ∼ 2025.9 – Application to Group Project and Further Research The phase two is anticipated to go further in the exploration. Although the concrete research topic can not be determined, the proposal here provides a narrow range. From a safety issue angle, it will be jailbreaking, backdoor, or red-teaming technologies, for the reason that these issue draws the most concerns. From a theoretical view, the recent quantitive conclusion between modalities [49] or classical symbolic rule-based framework [50] will be referred to. At the same time, considering the primary research outcome will be achieved, the research plan is to apply them in the CTIS context and test whether the specific version of these research outcomes works in our system.

Phase 3: 2025.10 ∼ 2026.3 – Practice in Industrial and Further Integration into Group Project Phase three concerns the extension of the research. On the one hand, the proposal highly anticipates extending the research to the industrial level during possible industrial research internships, realizing practical alignment between users and AI. On the other hand, research will focus on consolidating the integration of general AI safety research outcomes into group projects, making them practical for transportation scenarios.

Phase 4: 2026.4 ∼ 2026.7 – Further Research and Final Integration in Group Project Phase four continues the research and expects more solid results for the end of the M.Phil. research. Furthermore, the final integration of proposed methods is also expected to be useful in group projects, as the outcome of the RBM project.

2.2.5 Intended Outcomes

The proposal anticipates the following research outcomes:

1. **Novel Technology & Methodologies for AI Safety** This research focuses on developing novel technologies and methodologies to enhance AI safety, including improved detection and mitigation strategies for hallucination, anchoring bias in LLMs, etc. These advancements will directly help the CTIS system by ensuring that decision-making processes are robust against unsafe outputs and biased information, thereby improving system reliability in urban environments.

2. **Enhanced Understanding of Mechanistic Interpretability** Through this study, a deeper understanding of the inner workings of AI systems will be established, especially in the context of interactions between different elements to produce intelligent behaviors. This understanding will help refine the CTIS system by providing a transparent mechanism to analyze and verify model predictions, facilitating safer integration of AI into transportation systems.

3. **Comprehensive Evaluation Approaches** This research will establish all-around evaluation metrics and benchmarks for assessing AI performance and safety. These evaluation methods will be applied to ensure whether the CTIS framework is robust to attacks and capable of maintaining performance under diverse conditions.

2.3 Project Milestones

The alignment between the milestones of the individual project and the group project is detailed in the following Table 3, highlighting how each phase of the individual research contributes to and integrates with the group project milestones.

Group Project Milestone	Individual Project Milestone	Relation and Alignment
Months 0-3: Individual Data Collection and Model Development: Collecting and processing data for tasks such as traffic flow modeling, EV charging optimization, and scheduling algorithms.	**2024.9 ∼ 2025.4**: Primary Research about MI for Safer: Focus on LLM hallucination detection, mitigation, and anchoring bias studies.	Foundational AI safety research supports group subprojects by providing essential technologies like hallucination indicators and anchoring bias mitigation for CTIS.
Months 3-6: Collaborative Data Integration and Model Alignment: Integrating datasets, aligning models, and establishing a unified data pipeline.	**2025.5 ∼ 2025.9**: Application to Group Project and Further Research: Applying research findings to CTIS and exploring safety issues like backdoor vulnerabilities.	Mechanistic interpretability and mitigation methods are integrated into the CTIS framework, aligning with group-level model synchronization and data pipelines.
Months 6-9: Holistic Platform Integration and Initial Testing: Combining components into a unified CTIS platform and conducting initial tests.	**2025.10 ∼ 2026.3**: Practice in Industrial and Further Integration into Group Project: Extending research to industrial scenarios and consolidating AI safety tools.	Research findings on safety mechanisms and transparency tools are incorporated into the CTIS platform to ensure reliability and adaptability in urban applications.
Months 9-12: Comprehensive Deployment and Future Planning: Deploying CTIS in selected areas, scaling citywide, and proposing future enhancements.	**2026.4 ∼ 2026.7**: Further Research and Final Integration in Group Project: Finalizing AI safety methods and integrating them into CTIS.	Advanced safety measures ensure scalability and robustness for group deployment goals, aligning individual research with broader CTIS deployment strategies.

Table 3: Alignment of Individual and Group Project Milestones

2.4 Budget Plan

2.4.1 Estimated Budget

The estimated total budget for the project is anticipated to be between **10,000 to 20,000 CNY**, which will be allocated across several key categories to support the research and development efforts effectively.

2.4.2 Budget Breakdown

The detailed breakdown of the budget is provided in Table 4.

Category	**Purpose**	**Estimated Cost (CNY)**
Crowd Sourcing	Gathering high-quality datasets for model training and evaluation	3,000-5,000
API Fees	Licensing special APIs are unavailable through institutional resources, such as the Claude pro version.	2,000-4,000
Cloud Computing & Storage Fees	Accessing additional GPU clusters, cloud storage, and computational power for large-scale experiments, which hold higher demands than the HPC group of HKUST-GZ (A single A800 GPU is always in line compared to 8 A100 GPUs may be sufficient)	3,000-6,000
Total Estimated Budget		**10,000-20,000**

Table 4: Budget Breakdown for the Project

2.4.3 Cost-Effectiveness

This budget has been carefully estimated to ensure cost-effectiveness while maximizing research output. Several measures have been considered:

- Leveraging institutional resources such as free access to standard computational clusters and OpenAI APIs, which reduces reliance on external services.
- Targeted use of crowd-sourcing, focusing on high-impact datasets/benchmarks that are essential to the research objectives.
- Prioritizing additional computing resources only when large-scale experiments surpass internal capacities.

By optimizing these needs, the project seeks to balance cost and performance, ensuring that every expenditure contributes meaningfully to the research goals.

2.5 Risk Analysis and Mitigation

2.5.1 Potential Risks or Challenges

Several potential risks or challenges may occur during the execution of the project:

- **Timeliness:** Considering the rapid developments in AI and related applications, some methods mentioned in the project may become outdated or less effective by the time of application to our group project.
- **Generalization and Safety in the Wild:** Ensuring that the proposed safety means and models generalize well to complicated real-world scenes remains a challenge. Models trained in controlled conditions may fail to capture the complexity of random, dynamic practical situations.

2.5.2 Impact of the Risks

The foreseeable impact of these risks on the project is as follows:

- **Timeliness:** If the proposed methods fail to be strongly connected to timeliness due to technological changes, it could weaken an individual project's applicability and contribution to the group's projects.
- **Generalization:** Being lack of robustness could reduce the effectiveness of the proposed solutions in real-world transportation systems, impacting their practicality and deployment success.

2.5.3 Mitigation Strategies

To address these risks, the following strategies will be implemented:

- **Timeliness Mitigation:** Regularly reviewing the latest findings in AI safety and incorporating both flexible and adaptable methodologies that can be updated with minimal overhead.
- **Generalization Mitigation:** Enhancing the diversity and scale of application scope to include a wider range of practical scenarios, along with testing in both simulated and real-world environments to identify and address potential failures as soon as possible.

References

[1] OpenAI, “Chatgpt,” 2023, version GPT-4, Large language model. [Online]. Available: https://openai.com/chatgpt

[2] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, “Trustworthy artificial intelligence: a review,” *ACM computing surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.

[3] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, “Trustworthy ai: From principles to practices,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.

[4] S. Qiu, Q. Liu, S. Zhou, and C. Wu, “Review of artificial intelligence adversarial attack and defense technologies,” *Applied Sciences*, vol. 9, no. 5, p. 909, 2019.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[6] A. Madry, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[7] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, “Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, 2022.

[8] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, “Backdoor learning: A survey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5–22, 2022.

[9] J. Liang, S. Liang, M. Luo, A. Liu, D. Han, E.-C. Chang, and X. Cao, “Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models,” *arXiv preprint arXiv:2402.13851*, 2024.

[10] Y. Wang, D. Xue, S. Zhang, and S. Qian, “Badagent: Inserting and activating backdoor attacks in llm agents,” *arXiv preprint arXiv:2406.03007*, 2024.

[11] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang *et al.*, “Ai alignment: A comprehensive survey,” *arXiv preprint arXiv:2310.19852*, 2023.

[12] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong, "Large language model alignment: A survey," *arXiv preprint arXiv:2309.15025*, 2023.

[13] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 141–159.

[14] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of multimodal large language models: A survey," *arXiv preprint arXiv:2404.18930*, 2024.

[15] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922*, 2023.

[16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, 2023.

[17] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," *arXiv preprint arXiv:2309.03883*, 2023.

[18] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] J. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He, "Cognitive bias in decision-making with llms," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 12 640–12 653.

[20] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[21] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE access*, vol. 10, pp. 25 494–25 513, 2022.

[22] A. Liu, L. Pan, Y. Lu, J. Li, X. Hu, X. Zhang, L. Wen, I. King, H. Xiong, and P. Yu, "A survey of text watermarking in the era of large language models," *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–36, 2024.

[23] S. Yi, Y. Liu, Z. Sun, T. Cong, X. He, J. Song, K. Xu, and Q. Li, "Jailbreak attacks and defenses against large language models: A survey," *arXiv preprint arXiv:2407.04295*, 2024.

[24] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," *arXiv preprint arXiv:2310.04451*, 2023.

[25] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[26] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[27] X. Gong, Q. Wang, Y. Chen, W. Yang, and X. Jiang, "Model extraction attacks and defenses on cloud-based machine learning models," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 83–89, 2020.

[28] Y. Li, Y. Guo, F. Guerin, and C. Lin, "An open-source data contamination report for large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 528–541.

[29] I. Magar and R. Schwartz, "Data contamination: From memorization to exploitation," *arXiv preprint arXiv:2203.08242*, 2022.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[31] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[32] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[34] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning*. PMlR, 2017, pp. 3145–3153.

[35] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” *arXiv preprint arXiv:2005.00928*, 2020.

[36] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept bottleneck models,” in *International conference on machine learning*. PMLR, 2020, pp. 5338–5348.

[37] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[38] R. S. Zimmermann, T. Klein, and W. Brendel, “Scale alone does not improve mechanistic interpretability in vision models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[39] L. Bereska and E. Gavves, “Mechanistic interpretability for ai safety–a review,” *arXiv preprint arXiv:2404.14082*, 2024.

[40] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu, “Scaling and evaluating sparse autoencoders,” *arXiv preprint arXiv:2406.04093*, 2024.

[41] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, “Locating and editing factual associations in gpt,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 359–17 372, 2022.

[42] F. Zhang and N. Nanda, “Towards best practices of activation patching in language models: Metrics and methods,” *arXiv preprint arXiv:2309.16042*, 2023.

[43] N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt, “Eliciting latent predictions from transformers with the tuned lens,” *arXiv preprint arXiv:2303.08112*, 2023.

[44] K. Pal, J. Sun, A. Yuan, B. C. Wallace, and D. Bau, “Future lens: Anticipating subsequent tokens from a single hidden state,” *arXiv preprint arXiv:2311.04897*, 2023.

[45] S. Chalnev, M. Siu, and A. Conmy, “Improving steering vectors by targeting sparse autoencoder features,” *arXiv preprint arXiv:2411.02193*, 2024.

[46] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski *et al.*, “Representation engineering: A top-down approach to ai transparency,” *arXiv preprint arXiv:2310.01405*, 2023.

[47] L. Kong, H. Wang, W. Mu, Y. Du, Y. Zhuang, Y. Zhou, Y. Song, R. Zhang, K. Wang, and C. Zhang, “Aligning large language models with representation editing: A control perspective,” *arXiv preprint arXiv:2406.05954*, 2024.

[48] P. L. Silva, A. de Domenico, A. Maatouk, and F. Ayed, “Pay attention to what matters,” *arXiv preprint arXiv:2409.19001*, 2024.

[49] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng, “Provable dynamic fusion for low-quality multimodal data,” in *International conference on machine learning*. PMLR, 2023, pp. 41 753–41 769.

[50] T. Mu, A. Helyar, J. Heidecke, J. Achiam, A. Vallone, I. Kivlichan, M. Lin, A. Beutel, J. Schulman, and L. Weng, “Rule based rewards for language model safety,” *arXiv preprint arXiv:2411.01111*, 2024.