**ISF RESEARCH PROPOSAL**

חקר הסברתיות לאלגוריתמי קבלת החלטות לפלטפורמות בלתי מאוישות המבוססים על מודלי שפה

**Title**:  **Large Language Models Explainability for Decision Making of Unmanned Platforms**

**Project coordinator**: **\_\_\_ Dr. Oren Gal, University of Haifa\_\_\_\_\_\_\_\_\_\_\_**

|  |
| --- |
| 1. Abstract |
| The integration of Large Language Models (LLMs) into unmanned platforms represents a significant technological advancement, pushing the boundaries of what autonomous systems can achieve. These platforms, equipped with advanced artificial intelligence, are increasingly deployed in a variety of complex environments, where they perform tasks ranging from environmental monitoring and infrastructure inspection to search and rescue operations. Despite their capabilities, a major challenge persists: the decision-making processes of LLMs are largely opaque, making it difficult for operators and stakeholders to understand or trust the system's autonomous actions. This lack of transparency is particularly problematic in critical applications where understanding the rationale behind decisions is essential for safety, compliance, and ethical considerations. This research proposal outlines a comprehensive approach to address these challenges through the innovative use of representation engineering, a method focused on enhancing the explainability of LLMs embedded within unmanned platforms. Representation engineering involves dissecting and manipulating the internal representations of neural networks to make the model's decisions more interpretable and aligned with human reasoning. This approach not only aims to increase the transparency of autonomous systems but also strives to enhance their operational reliability and public trustworthiness. |
| 1. Scientific background and state of the art |
| State of the Art in Explainable AI: Explainable AI (XAI) aims to make the behavior of AI systems more understandable to humans. Recent advancements have introduced various theoretical frameworks and practical tools to dissect and explain the decision-making processes of AI systems. Techniques such as feature attribution, model visualization, and probing are employed to interpret AI actions [1,2,3]. However, the unique nature of LLM decisions, particularly in text processing and generation, demands more specialized approaches for effective explainability.  Recent advancements in LLMs have shown impressive performance in processing and generating human-like text, enabling sophisticated decision-making capabilities in unmanned systems [5,6,7]. However, the "black box" nature of these models poses significant risks, including the potential for unexpected behaviors and decisions that are difficult to predict or explain. The challenge is compounded in dynamic environments where decisions must be made rapidly and with high stakes involved. The ability to explain an AI's decision-making process in human terms is crucial for several reasons: it facilitates debugging and improvement of the system, ensures compliance with regulatory standards, enhances user trust, and enables more effective human-machine collaboration. Importantly, explainability in AI is increasingly seen not only as a technical requirement but also as an ethical imperative. Over the last years, several explainability techniques for Large Language Models (LLMs) were addressed, focusing on methods for generating local and global explanations based on different training paradigms: attention mechanism analysis, Integrated Gradients, natural language explanations, and Explanation Regularization [1-5].  The first method related to *attention mechanism analysis*, where we would like to analyze the attention mechanisms within LLMs to understand which parts of the input sequence the model focuses on during prediction. That provides insights into the model's decision-making process by highlighting important tokens in the input sequence and helps identify biases or heuristics used by the model in making predictions [6]. On the other hand, we are limited to understanding local interactions within the model and we may not capture the overall reasoning process of the model. Following that, the *integrated gradients method* developed, which attributes the model's prediction to each input token by integrating the gradients along the path from a baseline input to the actual input. By that, we offers a systematic way to interpret the model's predictions by assigning importance to each input token and helps in identifying shortcut cues or biases in the model's decision-making process.The main difficult concern to computationally intensive, especially for large models, which can limit real-time applications, and also sensitivity to noise in the input data may lead to misleading interpretations. Over the years, natural language explanation capability was presented. By that, we generate natural language explanations for the model's decision-making by training a language model on original textual data and human-annotated explanations. That provide human-understandable explanations for model predictions and can also improve downstream prediction accuracy and serve as a data augmentation technique. On the other hand, the reliability of generated explanations may require further investigation and a separate model generation for explanations may introduce additional complexity. In the last years, Explanation Regularization (ER) was developed. ER methods are used to align the model's machine rationales with human rationales to improve generalization and performance. By that, we can improve model generalization by aligning machine rationales with human rationales and we can also enhances model accuracy for various tasks, even in the absence of human rationales. Although, evaluation of ER models for out-of-distribution generalization can be complex and also requires post-hoc explanation methods, adding an additional layer of complexity.  Each of these methods offers unique insights into the explainability of LLMs, with advantages such as interpretability and improved model performance. However, they also come with limitations such as computational intensity, sensitivity to noise, and potential reliability issues. In the following, we present our concept, dealing with these challenges for unmanned platform using Representation Engineering [6-14] and offering methodology that will face these challanges.  The integration of Large Language Models (LLMs) into unmanned systems such as drones and Autonomous Underwater Vehicles (AUVs) represents a transformative leap in the capabilities of autonomous platforms. These systems rely on sophisticated AI to navigate, make decisions, and perform tasks autonomously. However, the internal mechanisms of LLMs are largely opaque, presenting significant challenges for transparency and reliability in high-stakes environments.  Current Challenges: The primary challenge in deploying LLMs in unmanned platforms lies in the "black box" nature of these models. While LLMs can process and generate human-like text, enabling complex decision-making capabilities, their decision processes are not inherently transparent. This opacity can lead to issues of trust, especially in critical applications where understanding AI decisions is crucial for safety and compliance [15-20].  Innovative Approaches: Representation Engineering: This research leverages representation engineering to enhance LLM explainability in unmanned platforms. Representation engineering involves two key phases: understanding (representation reading) and actively modifying (representation control) the internal representations of neural networks. This approach not only aims to make the AI's decision-making process transparent but also enhances the AI’s performance by aligning its decision-making with ethical and practical standards.  *Representation Engineering:* Representation engineering is a method used in the field of machine learning that focuses on designing and selecting features in data that are most beneficial for training models. The primary goal of this approach is to enhance the interpretability and performance of machine learning algorithms by optimizing the input data representation. This process involves transforming raw data into a format that better exposes the underlying patterns and correlations relevant to the prediction task. Representation engineering can include techniques such as feature extraction, where new features are derived from the raw data, and feature selection, where redundant or irrelevant features are removed. This method is crucial because the quality and appropriateness of features can significantly impact the learning process and the model's eventual accuracy. By carefully engineering features, researchers can improve model robustness, reduce computational costs, and increase the generalizability of the model to new data sets. The effectiveness of representation engineering has been demonstrated in various applications, from image and speech recognition to predictive analytics in finance and healthcare. The mathematical methodology behind representation engineering in machine learning typically revolves around optimizing the input features to improve the efficacy of learning algorithms. This involves several key techniques and concepts: Feature Extraction and selection, Dimensionality Reduction and Information Theory. By integrating these mathematical tools and concepts, representation engineering systematically enhances the features used in machine learning models, aiming to provide cleaner, more informative, and less redundant data which in turn helps to build more accurate and efficient predictive models.  *Challenges in Representation Engineering:*  Complexity of Models: The complexity of LLMs, with potentially billions of parameters, makes it difficult to map and modify internal representations without oversimplifying the model's capabilities. Scalability: The techniques must effectively scale with the increasing size and complexity of state-of-the-art LLMs. Accuracy vs. Interpretability Trade-off: Increasing a model’s interpretability can sometimes reduce its decision-making performance. Dynamic Operational Environments: The methods must be robust enough to adapt to the dynamic and unpredictable environments in which unmanned platforms operate.  Focus of our research: The research focuses on developing scalable, precise representation engineering techniques applicable in both simulated and real-world settings, that was not done till today in this field. This includes: - Developing sophisticated probing methods to analyze deep feature representations within LLMs. - Enhancing algorithms for feature visualization and manipulation. - Conducting rigorous experimental validation in controlled and real environments. - Collaborating with end-users and stakeholders to ensure practical applicability and operational trust. Representation engineering offers distinct benefits over other explainability techniques by providing a deeper, active understanding of model operations. This method allows both local adjustments to specific representations and global assessments of model behavior, which are essential for aligning LLM outputs with human-understandable logic.  *Representation Engineering Advantages for Explainability for Unmanned:*  Representation Engineering offers several distinct advantages over traditional methods for enhancing the explainability of Large Language Models (LLMs), based on the following characters: 1. Deeper Insight into Model Representations – Representation Engineering provides a more granular understanding of how AI models process and represent information internally. Unlike surface-level techniques that may only offer insights into input-output relationships, this approach delves into the model's internal representation layers. It explores how different types of information are encoded at various stages of the processing pipeline, providing a deeper understanding of the model's cognitive processes and decision-making pathways. 2. Active Manipulation of Representations - A unique feature of Representation Engineering is its ability to not just interpret, but actively manipulate model representations. This means that researchers and developers can experimentally modify the internal states of a model to see how these changes affect outputs. Such manipulations can be used to directly enhance model transparency, test hypotheses about the model's functioning, and potentially correct or improve model behaviors based on desired outcomes. 3. Alignment with Human-Understandable Concepts - Representation Engineering focuses on aligning machine representations with human-understandable concepts. This is particularly valuable in applications where AI decisions need to be interpretable by non-expert users or where decisions must be justified in comprehensible terms. By mapping complex model representations to simpler, conceptually intuitive forms, it bridges the gap between high-level human cognitive processes and low-level machine operations. 4. Improved Model Trustworthiness and Reliability - Enhancing explainability through Representation Engineering can lead to greater trust and reliability in AI systems. By providing clearer insights into how models make decisions, stakeholders can better assess the fairness, bias, and potential risks associated with AI outputs. This is crucial for deploying AI in sensitive and critical domains such as healthcare, finance, and autonomous vehicles, where understanding AI decisions can impact safety and regulatory compliance.  5. Facilitation of Debugging and Model Improvement - With its ability to dissect and modify internal representations, Representation Engineering is an excellent tool for debugging and refining AI models. Developers can identify and isolate problematic behaviors, understand their root causes, and implement targeted modifications to improve performance. This proactive approach to model tuning can lead to more robust, accurate, and fair AI systems. 6. Scalability and Applicability Across Different Models - Representation Engineering is designed to be scalable and applicable across different types of neural networks and machine learning models. This universality makes it a versatile tool in the AI developer’s toolkit, capable of being adapted for use with emerging models and architectures. As AI technology evolves, the principles of Representation Engineering can be applied to new generations of models, ensuring ongoing relevance.  Representation Engineering advances the state of AI explainability by providing tools not only for better understanding but also for actively influencing how AI models process information. This approach is particularly well-suited to addressing the challenges posed by modern LLMs and other complex models, offering a path toward more transparent, trustworthy, and effective AI systems. |
| 1. Research objectives |
| The primary objective of this research is to advance the interpretability and transparency of large language models (LLMs) in autonomous platforms through innovative representation engineering. This study aims to establish a framework that enables the detailed examination, control, and visualization of LLM decision-making processes, making these models more accessible and understandable to both technical and non-technical stakeholders. By embedding ethical considerations and bias mitigation strategies, this research strives to develop tools and protocols that enhance trust in autonomous systems operating in critical, high-stakes environments.  The specific aims of the research include:   1. ****Development of Probing and Visualization Techniques****: Design and implement advanced probing methods and dynamic visualization tools to reveal internal LLM representations across layers and decision stages. These tools will facilitate real-time, user-friendly insights into LLM decision pathways, enabling stakeholders to assess and interact with model outputs effectively. 2. ****Establishment of Representation Control Mechanisms****: Develop algorithms that actively manipulate LLM representations, aligning model decision-making processes with human-understandable concepts and ethical standards. Using reinforcement learning and meta-learning techniques, the research will iteratively refine these mechanisms to balance interpretability with operational performance. 3. ****Validation of Interpretability and Bias Resilience****: Conduct comprehensive simulations and real-world testing to assess the scalability, robustness, and ethical alignment of the representation engineering techniques. This validation will involve diverse environmental settings and tasks, ensuring that the methods maintain reliability and transparency across various autonomous applications. 4. ****Standardization and Open-Source Contribution****: Create reproducible, standardized protocols and prepare open-source resources, including code, datasets, and documentation, to support broader research impact and external validation. The project’s open-source contribution will enable future researchers and developers to build on these tools, fostering innovation in explainable AI and advancing the field of LLM interpretability for autonomous systems.   This research is designed to produce not only technical advancements but also an ethical framework and reproducible resources that set new standards in the application of LLMs for autonomous, high-stakes operations.   1. Methodology  4.1 Data Acquisition and Preprocessing To ensure robust training and evaluation of our explainability techniques, data acquisition and preprocessing will follow a structured approach that aligns with real-world applications of autonomous systems. The datasets will be carefully selected to encompass diverse scenarios, simulating conditions encountered by unmanned platforms. These datasets will include:   1. **Publicly Available Datasets**: Where applicable, open-source datasets reflecting real-world conditions, such as those from environmental monitoring, infrastructure inspection, and emergency response scenarios, will be utilized. These will provide standardized, validated data that supports generalizable results. 2. **Proprietary and Simulated Data**: In collaboration with industry partners, proprietary datasets will be integrated to represent domain-specific applications. Additionally, we will generate simulated datasets tailored to critical decision-making tasks, enabling controlled variations and ensuring coverage of complex scenarios that may not be fully represented in existing datasets.   ***Data Preprocessing and Augmentation*** Preprocessing steps will include data normalization, de-noising, and standardization to align input data formats with large language model (LLM) requirements. Augmentation techniques, such as synthetic data generation and context variability, will be applied to bolster model resilience and mitigate overfitting to specific data distributions. These preprocessing steps are critical for ensuring the robustness of the proposed techniques across diverse conditions.  ***Control and Randomization Techniques*** To minimize bias, we will employ stratified sampling and randomized splitting. Controlled trials across varied environmental conditions will further enhance robustness and generalizability. For example, data will be partitioned based on factors like geographic diversity and environmental complexity, ensuring balanced representation across training, validation, and test sets. These controls allow for consistent, unbiased assessment of model interpretability across operational scenarios. 4.2 Statistical Analysis and Evaluation Framework A rigorous statistical framework will evaluate the alignment between LLM representations and human interpretability criteria, addressing the core research question on enhancing LLM explainability for autonomous systems.  ***Quantitative Metrics*** Metrics will include explainability scores, consistency indices, and alignment ratios to measure the degree to which LLM decision pathways align with human-understandable reasoning patterns. Explainability scores will quantify the transparency of LLM processes, while consistency indices assess stability in model interpretation across varying conditions. Confidence intervals will be used to estimate the reliability of interpretability enhancements, particularly in dynamic, real-time applications.  ***Statistical Testing and Significance Evaluation*** We will apply statistical tests such as Analysis of Variance (ANOVA), t-tests, and Chi-square tests to evaluate the statistical significance of enhancements in model interpretability. These tests will ascertain whether observed improvements are consistent and significant across multiple conditions, strengthening the validity of our findings. Significance testing will be supplemented by effect size calculations to measure the practical impact of representation engineering techniques.  ***Qualitative Assessment*** To complement quantitative metrics, human evaluators will assess the model’s interpretability based on predefined criteria for human-like reasoning. Evaluators will review anonymized outputs and assign interpretability scores. Inter-rater reliability, measured using Cohen’s kappa, will ensure consistency in qualitative assessments. This mixed-methods approach combines quantitative rigor with qualitative insights, providing a comprehensive evaluation of technique effectiveness. 4.3 Advanced Probing and Visualization Techniques Advanced probing and visualization will be central to exploring and illustrating the LLM’s internal representations, offering insight into the model’s decision-making process and enabling real-time interpretability adjustments.  ***Probing Mechanisms*** Advanced probes will be deployed to examine LLMs’ internal representations across multiple layers, with a focus on detecting patterns in neural activations related to decision points. Probes will interact with individual layers to capture activation patterns, allowing us to map the influence of specific data inputs on model outputs. Probes will be validated through simulation studies to ensure robustness across different model architectures.  ***Dynamic Visualization Tools*** We will develop a suite of dynamic visualization tools designed to represent LLM decision-making pathways visually. These tools will enable interactive examination of LLM processes, displaying activation patterns, decision branches, and potential outcome variations. A graphical user interface (GUI) will allow users to adjust inputs in real time, observing changes in model output dynamically. This feature will support exploratory analysis by stakeholders and facilitate engagement with the model’s cognitive processes. 4.4 Representation Control and Manipulation Representation control is pivotal for aligning LLM decision-making with human interpretability standards. Through active manipulation, we aim to enhance the transparency of LLMs in autonomous systems by developing algorithms that adjust the model’s internal representations in response to external feedback.  ***Algorithm Development for Representation Control*** Algorithms will be developed to refine LLM representations based on human-reasoning frameworks. These algorithms will incorporate reinforcement learning and meta-learning approaches, enabling dynamic adaptation of internal representations to align with established interpretative standards. Reinforcement learning will iteratively optimize interpretability by rewarding alignment with human reasoning patterns, while meta-learning will facilitate generalization across diverse tasks and environments.  ***Iterative Testing and Refinement*** The control algorithms will undergo iterative testing in simulated environments and real-world applications. Each iteration will be evaluated for interpretability enhancement and decision accuracy, particularly in time-sensitive scenarios. Feedback loops will allow for continuous refinement, ensuring that representation adjustments maintain the model’s operational reliability without sacrificing interpretability. 4.5 Simulation and Real-World Validation To evaluate robustness, scalability, and practical utility, the proposed techniques will be validated across controlled simulation environments and real-world unmanned platforms.  ***Simulation Environments*** Simulated environments will replicate real-world conditions with controlled complexity, allowing us to evaluate technique performance across varied conditions. High-performance computing resources will support these simulations, facilitating rapid testing iterations. Simulation parameters will include environmental diversity, temporal dynamics, and operational demands typical of unmanned systems.  ***Real-World Testing on Unmanned Platforms*** Upon successful simulation validation, we will deploy the refined techniques on unmanned platforms (e.g., drones, AUVs) in controlled real-world environments. Real-world validation will focus on essential tasks such as navigation, object recognition, and obstacle avoidance. These experiments will provide critical insights into scalability and adaptability, demonstrating the methods’ practical utility for enhancing the explainability of LLMs in autonomous applications. 4.6 Ethical Considerations and Bias Mitigation This research is grounded in ethical responsibility, recognizing the critical need for transparency and trust in AI-driven autonomous systems. Ethical audits and bias mitigation strategies will be implemented to safeguard data integrity and model reliability.  ***Ethical Audits*** Regular ethical audits will evaluate data privacy, security, and bias considerations. An interdisciplinary ethics committee will provide oversight, ensuring compliance with industry standards and regulatory requirements. Audits will focus on data handling, storage, and access, particularly regarding sensitive data used in simulations and real-world validations.  ***Bias Mitigation Strategies*** Bias mitigation will be integral to the representation engineering process. Techniques such as data diversification, randomization, and balanced feature selection will reduce potential biases in LLM interpretation. Post-hoc evaluations will assess biases that may arise from representation manipulation, and corrective adjustments will be made as necessary. 4.7 Reproducibility Strategy To enhance reproducibility, the research will adhere to open-source standards, allowing others to replicate and extend our findings. Comprehensive documentation, model versioning, and standardized testing protocols will be incorporated to facilitate external validation.  ***Open-Source Code and Data Repositories*** We will release code, data, and experimental protocols in public repositories, accompanied by detailed documentation. Version control will be maintained to track model adjustments and updates, providing a transparent record of iterative refinements.  ***Standardized Protocols for Testing and Validation*** All research phases will be governed by standardized testing protocols, including data splits, hyperparameter settings, and model configurations. These protocols will enable consistent benchmarking and support reproducibility efforts by the broader AI research community. |
| 1. ****Research Phases**** |
| The following sections outline the phases of this four-year research project, structured to achieve rigorous development, validation, and reproducibility of explainability techniques for large language models (LLMs) in autonomous systems. Each phase builds upon the previous, ensuring a comprehensive approach to enhance interpretability and transparency within high-stakes autonomous platforms. **Phase 1: Data Acquisition and Preprocessing (Months 1-12)** ****Objective****: Establish a robust and unbiased data foundation for training and evaluating representation engineering techniques.   1. ****Data Selection and Acquisition****: Datasets will be acquired from publicly available sources, proprietary partnerships, and simulations specifically tailored to key autonomous tasks (e.g., environmental monitoring, search and rescue). This ensures relevance and coverage of complex, real-world scenarios. 2. ****Data Preprocessing and Augmentation****: Standard preprocessing methods, such as normalization, de-noising, and format standardization, will prepare data for consistent model input. To further enhance model robustness, data augmentation techniques (e.g., synthetic data generation, contextual variation) will be applied, creating a diverse set of inputs for training and evaluation. 3. ****Control and Randomization Techniques****: Bias mitigation strategies will include stratified sampling, randomized data splitting, and controlled trials across multiple scenarios, ensuring balanced representation and reliability in testing interpretability improvements. These techniques allow for a rigorous evaluation across a broad spectrum of operational conditions, supporting the generalizability of our findings.  **Phase 2: Statistical Analysis and Evaluation Framework (Months 6-18)** **Objective**: Develop and implement a comprehensive evaluation framework to assess model interpretability and alignment with human-understandable reasoning.   1. **Metric and Evaluation Framework Design**: Quantitative metrics such as explainability scores, consistency indices, and alignment ratios will be designed to quantify model interpretability. Explainability scores will assess transparency, while consistency indices measure interpretative stability across varied conditions. 2. **Statistical Testing and Qualitative Assessment Setup**: Statistical tests, including ANOVA, t-tests, and Chi-square tests, will measure the significance of observed interpretability improvements. A qualitative assessment will complement these quantitative metrics through human evaluators who assess the model’s decision-making transparency. Inter-rater reliability, measured by Cohen’s kappa, will ensure consistency across evaluations.   This framework provides a rigorous, dual-methods approach to validate enhancements in LLM interpretability, ensuring that both numerical metrics and human assessments confirm alignment with human reasoning patterns. **Phase 3: Probing and Visualization Development (Months 12-30)** ****Objective****: Create probing mechanisms and dynamic visualization tools to explore and illustrate LLM decision-making processes.   1. ****Development of Probing Mechanisms****: Advanced probes will be designed to access and interpret the internal representations of LLMs across multiple layers, allowing us to observe activation patterns during decision-making processes. These probes will be optimized through simulation studies to ensure reliability across different architectures. 2. ****Dynamic Visualization Tools Development****: Visualization tools will display real-time internal states and decision pathways, enabling interactive examination of model behavior. A graphical interface will allow users to adjust inputs and immediately observe changes in model outputs, providing transparency into the LLM’s cognitive processes. 3. ****Pilot Testing and Validation in Simulated Environments****: Initial testing of probing and visualization tools will occur in simulated settings to validate functionality and gain insights into tool effectiveness. This early testing will guide iterative improvements, ensuring these tools meet usability and interpretability standards.  **Phase 4: Representation Control and Manipulation (Months 18-36)** ****Objective****: Develop algorithms for active representation control, enabling the alignment of LLM internal representations with human reasoning frameworks.   1. ****Development of Representation Control Algorithms****: Algorithms will be designed to adjust LLM representations based on human-reasoning principles. Using reinforcement learning and meta-learning approaches, these algorithms will iteratively optimize interpretability, rewarding alignment with human reasoning while maintaining model accuracy. 2. ****Iterative Testing and Refinement****: Control algorithms will be subjected to iterative testing in both simulated and real-world settings. Feedback loops will support continuous refinement, ensuring that interpretability improvements do not compromise model reliability. This iterative process will optimize alignment with human-understandable concepts while maintaining performance standards in operational scenarios.  **Phase 5: Simulation and Real-World Validation (Months 24-48)** ****Objective****: Validate the interpretability techniques under controlled and real-world conditions to ensure scalability and applicability to autonomous platforms.   1. ****Extensive Simulation Testing****: Simulation environments will be configured to replicate operational conditions, including environmental variability and task complexity. These extensive simulations will assess the robustness and scalability of the representation engineering techniques, allowing rapid testing iterations within a high-performance computing environment. 2. ****Real-World Validation on Unmanned Platforms****: After successful simulation validation, the techniques will be implemented on unmanned platforms such as drones and autonomous underwater vehicles (AUVs). Real-world testing will focus on core tasks like navigation, object recognition, and obstacle avoidance, providing critical insights into the model’s adaptability in real-world conditions.  **Phase 6: Ethical Audits and Bias Mitigation (Months 18-42)** ****Objective****: Ensure ethical transparency and mitigate bias in data handling and model outputs throughout the research process.   1. ****Ethical Audits****: Regular ethical audits will assess data privacy, security, and bias considerations. These audits, guided by an interdisciplinary ethics committee, will ensure adherence to regulatory standards and best practices in ethical AI development. 2. ****Bias Mitigation Strategies Implementation****: Data diversification, randomization, and feature balancing techniques will be applied to minimize biases in representation engineering. Post-hoc bias evaluations will identify and address any emergent biases, ensuring ethically sound model outputs that align with interpretability goals.  **Phase 7: Reproducibility and Open-Source Release (Months 30-48)** ****Objective****: Ensure reproducibility and facilitate broader research impact through open-source code, data, and documentation.   1. ****Preparation of Open-Source Code and Documentation****: All code, data, and methodological documentation will be prepared for open-source release, including detailed explanations of each step to support transparency and reproducibility. 2. ****Establishment of Public Repositories****: Public repositories will host the code, datasets, and documentation, with version control to track model iterations and refinements. These repositories will serve as a foundation for external validation and collaboration, ensuring that the findings and tools developed in this project can be utilized and built upon by the broader research community. |
| 1. Significance, innovation and potential benefits of the proposed research |
| This research addresses a critical gap in the transparency and interpretability of large language models (LLMs) when used in autonomous platforms, such as drones and autonomous underwater vehicles (AUVs). By developing advanced representation engineering techniques, the project aims to provide deep insights into the decision-making processes of LLMs, making them accessible and understandable for stakeholders. This is particularly important for high-stakes environments where trust, reliability, and accountability are paramount.  The project’s focus on LLMs in autonomous systems responds to the urgent need for explainable AI in domains where safety, compliance, and ethical considerations are crucial. The proposed framework will allow users to visualize and interact with model decisions, setting a standard for transparent AI operations in unmanned platforms.  The project introduces a novel combination of probing, real-time visualization, and active representation control. This approach goes beyond traditional XAI by enabling both local and global interpretability while allowing modifications that align AI behavior with human reasoning patterns. By embedding bias mitigation and ethical audits, the research promotes ethically aligned and reliable AI.  The research is expected to yield operational tools for real-time decision monitoring, leading to better-informed and safer AI deployment in critical sectors like environmental monitoring, search and rescue, and infrastructure inspection. Open-source contributions will foster broader adoption, providing the AI research community with scalable, reproducible methods that can be built upon in future explainability research. |
| 1. Applicability |
| Unmanned platforms, such as UAVs and AUVs, will benefit substantially from representation engineering, which enhances operational transparency and opens new applications by making AI decision-making processes understandable. This increased explainability is essential in sectors like healthcare, search and rescue, and agriculture, where clear, interpretable AI actions are critical for trust, safety, and regulatory approval. Furthermore, representation engineering supports real-time adaptive learning and collaborative tasks, allowing these systems to transparently communicate their choices in dynamic, high-stakes environments, thereby fostering broader acceptance and successful deployment across industries.  1. **Expected Results and Pitfalls**   This research aims to establish a comprehensive framework for enhancing the explainability of large language models (LLMs) in unmanned platforms through representation engineering. By enabling the precise mapping, analysis, and control of internal LLM representations, we seek to develop scalable techniques that will significantly improve transparency in AI-driven decision-making for high-stakes applications. Below, the expected results and potential pitfalls are outlined, along with proposed mitigation strategies. **Expected Results**  1. **Enhanced Interpretability of LLM Decision-Making**: Through advanced probing and visualization techniques, the project aims to provide insights into the internal decision-making pathways of LLMs. These tools will allow stakeholders to observe how LLMs process inputs at various stages, making complex decision processes more accessible and understandable. This transparency will be particularly beneficial for high-stakes applications, where the reliability and clarity of AI decisions are paramount. 2. **Operational Tools for Real-Time Analysis**: The development of dynamic visualization tools and interactive interfaces is expected to yield a set of operational tools that allow real-time monitoring of LLM decision-making pathways. These tools will be user-friendly, enabling both technical and non-technical stakeholders to engage with and evaluate AI behavior. This may also lead to an open-access platform or database for broader community engagement and educational purposes. 3. **Framework for Ethical and Bias-Resilient Autonomous Systems**: By embedding ethical audits and bias mitigation strategies into every research phase, this project aims to set a standard framework for LLM applications in autonomous platforms. This framework will ensure that AI decisions are not only transparent but also ethically sound and resilient against biases, aligning with the regulatory requirements and ethical standards for autonomous systems. 4. **Standardized Protocols and Open-Source Contribution**: The final stages of the project will focus on establishing standardized, reproducible protocols and preparing open-source code, data, and documentation. This transparency in development will support external validation and collaboration, enabling future researchers and industry stakeholders to build on and refine the project’s findings. The open-source release is expected to catalyze advancements in explainable AI by providing accessible tools and frameworks for the community.  **Potential Pitfalls and Mitigation Strategies**  1. **Complexity of Large Language Models**: **Pitfall**: The inherent complexity of LLMs, with billions of parameters and intricate processing pathways, may impede the ability to consistently map and interpret internal representations. Extracting clear, human-interpretable patterns could be challenging, especially when trying to align these with simplified reasoning processes.   **Mitigation Strategy**: To address this, we will use iterative probing techniques to gradually refine the interpretability of complex patterns within LLMs. Advanced data visualization techniques will be employed to represent high-dimensional data in more accessible forms, breaking down complex processes into smaller, interpretable steps. Additionally, controlled testing will focus on isolating specific aspects of LLMs to reduce complexity and improve interpretability.   1. **Trade-off between Interpretability and Performance**: **Pitfall**: Enhancing interpretability by adjusting LLM representations may risk reducing the accuracy or efficiency of the models, particularly in real-time applications. This trade-off between transparency and operational performance could impact the model’s effectiveness in high-stakes scenarios.   **Mitigation Strategy**: A balanced approach will be employed, where interpretability improvements are incrementally introduced while rigorously monitoring model performance. Each change in model representation will undergo performance testing to ensure that interpretability gains do not compromise task accuracy. Techniques such as meta-learning and reinforcement learning will enable adaptive adjustments, allowing the model to maintain optimal performance even with interpretability constraints.   1. **Scalability of Representation Engineering Techniques**: **Pitfall**: Representation engineering is computationally intensive, which may present scalability issues, especially when applied to larger LLMs or in real-time applications with limited resources.   **Mitigation Strategy**: Initial studies will focus on optimizing the techniques in high-performance computing environments, followed by incremental scalability assessments in simulated settings. Hybrid approaches that combine representation engineering with established XAI techniques will be explored to balance computational demands, making the framework adaptable to various operational constraints.   1. **Inconsistency in Interpretability across Diverse Environments**: **Pitfall**: The interpretability techniques developed may show varying performance across different environments, especially in real-time applications where conditions are dynamic and unpredictable.   **Mitigation Strategy**: Validation studies will be conducted across diverse simulated and real-world environments to iteratively refine the techniques and improve generalizability. Feedback loops from real-world testing will guide adjustments, allowing us to identify and address environment-specific limitations.   1. **Potential for Unintended Bias in Representation Adjustments**: **Pitfall**: Actively modifying LLM representations to enhance interpretability might introduce unintended biases, which could impact decision-making reliability and ethical alignment.   **Mitigation Strategy**: Bias assessment protocols will be embedded into each phase, with regular evaluations using diverse datasets to detect and address emerging biases. Ethical audits will ensure compliance with best practices, while adjustments will be made to maintain alignment with ethical standards. Post-hoc bias analyses will be conducted, and corrective strategies will be implemented if any biases are identified.   1. **Reproducibility and Open-Source Limitations**: **Pitfall**: Given the complexity of representation engineering, ensuring reproducibility and effectively transferring knowledge to an open-source format may be challenging.   **Mitigation Strategy**: Detailed documentation and standardized testing protocols will support transparency and reproducibility. Each development phase will include version control and detailed change logs to provide a clear record of model updates. The open-source release will include tutorials, examples, and technical support to facilitate understanding and usability by external researchers and practitioners. |
| 1. Work plan and Gantt |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | ****Phase**** | ****Task**** | ****1-6 Months**** | ****6-12 Months**** | ****12-18 Months**** | ****18-24 Months**** | ****24-30 Months**** | ****30-36 Months**** | ****36-42 Months**** | ****42-48 Months**** | | ****Phase 1: Data Acquisition and Preprocessing**** | 1.1 Data selection and acquisition | **X** |  |  |  |  |  |  |  | |  | 1.2 Data preprocessing and augmentation | **X** | **X** |  |  |  |  |  |  | |  | 1.3 Control and randomization implementation | **X** | **X** |  |  |  |  |  |  | | **Phase 2: Statistical Analysis and Evaluation Framework** | 2.1 Metric and evaluation framework design |  | **X** | **X** |  |  |  |  |  | |  | 2.2 Statistical testing and qualitative assessment setup |  | **X** | **X** |  |  |  |  |  | | **Phase 3: Probing and Visualization Development** | 3.1 Development of probing mechanisms |  |  | **X** | **X** |  |  |  |  | |  | 3.2 Dynamic visualization tools development |  |  | **X** | **X** | **X** |  |  |  | |  | 3.3 Pilot testing in simulated environments |  |  |  | **X** | **X** |  |  |  | | **Phase 4: Representation Control and Manipulation** | 4.1 Development of representation control algorithms |  |  |  | **X** | **X** |  |  |  | |  | 4.2 Iterative testing and refinement |  |  |  | **X** | **X** | **X** |  |  | | **Phase 5: Simulation and Real-World Validation** | 5.1 Extensive simulation testing |  |  |  |  | **X** | **X** |  |  | |  | 5.2 Real-world validation on unmanned platforms |  |  |  |  | **X** | **X** | **X** | **X** | | **Phase 6: Ethical Audits and Bias Mitigation** | 6.1 Ethical audits and review |  |  |  | **X** | **X** | **X** | **X** |  | |  | 6.2 Bias mitigation strategies implementation |  |  |  | **X** | **X** | **X** | **X** | **X** | | **Phase 7: Reproducibility and Open-Source Release** | 7.1 Preparation of open-source code and documentation |  |  |  |  |  | **X** | **X** |  | |  | 7.2 Establishment of public repositories |  |  |  |  |  | **X** | **X** | **X** |  |  | | --- | |  | |
|  |
| 1. Bibliography |
| 1. \*\* Keneni, Blen et al. Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles. IEEE Access 7: 17001-17016, 2019. 2. Chiyah-Garcia, Javier et al. Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models. International Conference on Natural Language Generation, 2018. 3. \*\* Ahn, Michael et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. Conference on Robot Learning, 2022. 4. Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. Towards interpreting and mitigating shortcut learning behavior of nlu models. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. 5. Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM (CACM)*, 2023. 6. Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*, 2023. 7. \*\* Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387. URL https://aclanthology.org/2022.naacl-main.387. 8. Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield- Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Frame- work for Transformer Circuits — transformer-circuits.pub. https://transformer-circuits.pub/2021/ framework/index.html, December 2021. [Accessed 27-11-2023]. 9. Joseph Enguehard. Sequential Integrated Gradients: a simple but effective method for explaining language models, May 2023. URL http://arxiv.org/abs/2305.15853. arXiv:2305.15853 [cs]. 10. Kawin Ethayarajh and Dan Jurafsky. Attention Flows are Shapley Value Explanations, May 2021. URL http://arxiv.org/abs/2105.14652. arXiv:2105.14652 [cs]. 11. Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Patholo- gies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1407. URL https://aclanthology.org/D18-1407. 12. Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based Adversarial Examples for Text Classi- fication. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6174–6181, 2020. doi: 10.18653/v1/2020.emnlp-main.498. URL http://arxiv.org/abs/ 2004.01970. arXiv:2004.01970 [cs]. 13. Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020. 14. Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020. 15. Luyu Qiu, Yi Yang, Caleb Chen Cao, Jing Liu, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet Hsiao, and Lei Chen. Resisting Out-of-Distribution Data Problem in Perturbation of XAI, July 2021. URL http: //arxiv.org/abs/2107.14000. arXiv:2107.14000 [cs]. 16. \*\* Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 17. Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile ̇ Lukošiu ̄te ̇, et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023. 18. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 19. Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain Yourself! Lever- aging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487. 20. \*\* Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 88–102, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology. org/2020.starsem-1.10.   Basic Science Declaration  The primary aim of this research is to advance the theoretical framework for explainability in large language models (LLMs) used in unmanned platforms, such as drones and autonomous underwater vehicles (AUVs). Our objectives focus on developing scalable methods for understanding and modifying internal model representations, which allow for a deeper, theory-driven comprehension of LLM decision-making. By dissecting these internal structures, the research will shed light on the core cognitive processes that underlie LLM functionality, enhancing their interpretability without prioritizing immediate application.  This research will contribute significantly to scientific knowledge by proposing and validating methodologies for representation engineering—approaches that expose, analyze, and potentially manipulate the inner workings of LLMs to align with human-understandable logic. The project is anchored in the fundamental exploration of AI cognition and seeks to address the gap in how complex machine learning models process and prioritize information, providing new insights into the principles governing autonomous decision-making.  Given these objectives and the emphasis on foundational theory, this proposal constitutes basic science, driven by the goal of expanding theoretical knowledge and understanding within the field of AI explainability, rather than focusing on direct applications. |