Facilitating Cross-Domain Reasoning Generalization through Conservative Offline Reinforcement Learning Leveraging Pre-trained Large Language Model Representations

Maxwell Ashford
Department of Systems Engineering, University of Central Florida
m.ashford@ucf.edu

Abstract

The rapid expansion of artificial intelligence into critical infrastructure and socio-technical systems necessitates a transition from narrow task-specific models to resilient agents capable of cross-domain reasoning. Current paradigms often struggle with the "distributional shift" encountered when moving from controlled training environments to high-stakes, real-world deployment. This paper investigates a novel framework for facilitating cross-domain reasoning generalization by integrating Conservative Offline Reinforcement Learning with the latent semantic representations inherent in pre-trained Large Language Models. Unlike online reinforcement learning, which requires continuous environmental interaction and carries significant safety risks in physical systems, our approach utilizes static, multi-domain datasets to derive robust decision-making policies. By leveraging the high-dimensional world knowledge embedded within pre-trained language architectures, the system can map abstract reasoning patterns across disparate domains, such as transitioning from logistics optimization to energy grid management. We provide a comprehensive system-level analysis focusing on structural trade-offs, architecture, and the governance of these hybrid models. Furthermore, we address the ethical implications of deploying autonomous reasoning systems in public infrastructure, emphasizing the need for conservative value estimation to prevent catastrophic failures. Our findings suggest that the intersection of offline learning and linguistic representation provides a sustainable and robust pathway for building generalized intelligent systems that align with complex human institutional frameworks.

Keywords:
Cross-Domain Generalization, Conservative Offline Reinforcement Learning, Large Language Models, Socio-Technical Systems, AI Governance, Infrastructure Robustness.

1. Introduction

The contemporary landscape of artificial intelligence is characterized by a paradox of proficiency and fragility. While deep learning models have achieved parity with or exceeded human performance in localized domains such as image recognition or strategic gaming, the ability to generalize these cognitive successes across fundamentally different operational

contexts remains elusive [11]. This limitation is particularly acute in the realm of socio-technical infrastructures, where the cost of failure is high and the complexity of environmental variables is immense. The challenge of cross-domain reasoning generalization is not merely a technical hurdle but a systemic requirement for the next generation of autonomous systems [3]. As these systems are integrated into energy grids, transportation networks, and healthcare delivery, they must demonstrate an ability to apply logical structures learned in one domain to the nuances of another without exhaustive retraining [25].

Conventional reinforcement learning paradigms rely heavily on active exploration, a process that is often untenable in real-world systems where trial-and-error can lead to physical damage or social harm [2]. The emergence of offline reinforcement learning offers a potential solution by enabling the derivation of optimal policies from historical data without the need for live environmental interaction [18]. However, offline methods face the significant challenge of overestimating the value of actions not present in the training set, often leading to erratic behavior in out-of-distribution scenarios [12]. By incorporating a conservative mechanism, researchers have begun to penalize these overestimations [17], yet the ability to reason across domains still requires a foundational knowledge base that simple reward-based learning lacks [29].

This research proposes that the latent representations within pre-trained Large Language Models (LLMs) serve as the necessary bridge for this gap [23]. LLMs are trained on vast corpora reflecting the breadth of human knowledge, encapsulating complex relational structures and causal heuristics [28]. When these representations are used as a feature space for conservative offline reinforcement learning, the resulting system gains access to a world-model that transcends specific task boundaries [7]. This paper explores the architectural integration of these technologies, the trade-offs inherent in such a hybrid system, and the governance frameworks required to ensure their safe deployment [8].

2. The Architecture of Cross-Domain Reasoning Systems

The integration of LLM representations into the reinforcement learning loop represents a shift from raw sensory-to-action mapping toward a more mediated form of semantic decision-making. In a typical system architecture, the state space of a physical environment is mapped into the high-dimensional embedding space of a pre-trained transformer model [28]. This allows the reinforcement learning agent to read the environment through a lens of human-like semantic relations. For instance, a state in a manufacturing system and a state in an urban traffic network, while physically distinct, may share underlying abstract similarities such as bottlenecking or resource depletion [31]. Because the LLM has seen these concepts used across thousands of different contexts in its training data, it provides a unified representation that facilitates the transfer of reasoning logic [4].

Structural trade-offs become apparent when selecting the depth of integration between the LLM and the reinforcement learning policy. A shallow integration might involve using the LLM as a static feature extractor, which preserves the stability of the pre-trained weights but

limits the agent's ability to adapt the representations to specific domain nuances [23]. Conversely, a deep integration involving fine-tuning the LLM layers through the reinforcement learning objective can lead to superior performance in the target domain but risks catastrophic forgetting, where the model loses its generalized reasoning capabilities [3]. Our analysis suggests that a modular architecture, where a frozen LLM core provides general features to a lightweight, task-specific policy network, offers the most robust balance for infrastructure applications where predictability is paramount [19].

The deployment of such architectures also necessitates a rethinking of computational infrastructure. LLMs are notoriously resource-intensive, requiring significant memory and processing power [19]. Moving these models into edge-computing scenarios, such as autonomous vehicles or smart building controllers, requires aggressive quantization and pruning techniques [5]. This introduces a trade-off between reasoning fidelity and system latency. In critical systems like power grid balancing, a delay of milliseconds can be the difference between stability and a blackout. Therefore, the design of cross-domain agents must include hierarchical reasoning modules where the high-level semantic analysis is performed at a lower frequency than the immediate reactive control loops [13].

3. Conservative Offline Reinforcement Learning as a Safety Foundation

Offline reinforcement learning is fundamentally a data-driven approach, making the quality and diversity of the historical datasets the primary determinants of agent behavior [9]. However, even with high-quality data, agents are prone to distributional shift, where they encounter states slightly different from those in the training set and take actions they incorrectly believe will yield high rewards [12]. In socio-technical systems, these hallucinated high-value actions can be disastrous [2]. Conservative mechanisms address this by explicitly regularizing the value function, ensuring that the estimated worth of an action is lower-bounded by the density of that action in the training data [17].

When applied to cross-domain generalization, conservatism acts as a stabilizing force. If an agent trained on logistics data is asked to reason about healthcare supply chains, the conservative objective forces the model to remain skeptical of its own predictions in the new, unfamiliar domain [16]. It prioritizes actions that align with the known logical structures of the base LLM representations while avoiding high-risk, low-certainty maneuvers [15]. This creates a safe-to-fail environment where the agent’s reasoning is constrained by both the historical data of its specific training and the general semantic constraints of human language [22].

Furthermore, the conservative approach facilitates a more nuanced form of value alignment. By grounding the learning process in pre-existing linguistic representations, we can imbue the agent with implicit human values embedded in language [26]. For example, if the LLM representations associate certain actions with safety or efficiency in a wide range of texts, the conservative reinforcement learning agent will naturally find those paths easier to justify within its value function [31]. This does not replace explicit reward shaping but provides a

multi-layered defense against the optimization of unintended behaviors, a common failure mode in traditional reinforcement learning systems [20].

4. Socio-Technical Implications and Infrastructure Robustness

The transition of AI from digital assistants to controllers of physical infrastructure marks a significant evolution in our socio-technical landscape. Infrastructure systems are not merely technical assemblages but are deeply intertwined with legal, economic, and social frameworks [32]. A cross-domain reasoning agent operating a smart city's water distribution system must navigate not only fluid dynamics but also regulatory compliance, water equity policies, and public health standards [21]. The use of LLM representations is particularly powerful here, as it allows the agent to process and reason over textual data such as policy documents or legal mandates alongside technical sensor data [14].

Robustness in this context refers to the system's ability to maintain function despite external shocks or internal failures. Traditional AI models often fail brittly—their performance drops off a cliff when faced with conditions outside their narrow training [11]. A cross-domain agent, by contrast, can utilize its generalized reasoning to find analogies in different sectors to solve novel problems. If a sudden equipment failure occurs in a way the agent has never seen in its specific domain, it may leverage LLM-derived knowledge of general mechanical failure patterns or emergency response protocols to mitigate the damage [26]. This resilient reasoning is essential for systems that are too large or complex for human operators to monitor in every detail [25].

However, the robustness of the AI system itself is a point of concern. The reliance on large, pre-trained models introduces a new form of vulnerability: the foundation model risk [19]. If the underlying LLM contains biases, factual errors, or security vulnerabilities, these will be inherited by every specialized agent built upon it [10]. Ensuring infrastructure robustness thus requires a rigorous auditing process of the foundation models themselves [24]. We argue for a governance structure that treats these models as digital utilities, subject to the same level of scrutiny and standardization as physical materials like steel or concrete used in civil engineering [8].

5. Governance, Policy, and Ethical Alignment

The deployment of autonomous reasoning systems in public-facing roles brings the questions of governance and ethics to the forefront. Who is responsible when a cross-domain agent makes a decision that results in a suboptimal social outcome? If an agent optimizes a transportation network for efficiency but inadvertently marginalizes a specific neighborhood, the accountability trail becomes complex [32]. Traditional liability models are often ill-equipped to handle systems that learn and reason across domains, as the causal link between the original programmer and the final action is mediated by vast amounts of data and opaque latent representations [14].

A conservative approach to learning offers a partial solution to the ethical challenge. By penalizing uncertainty, we can design systems that default to human intervention when faced with high-stakes decisions that fall outside their confidence zone [2]. This creates a human-in-the-loop requirement that is dynamically triggered by the system's own self-assessment of its reasoning certainty [15]. Policy frameworks should mandate such thresholds for any AI system operating in critical infrastructure [14]. Furthermore, the use of LLM representations allows for a more transparent explanation of reasoning. Because the internal state of the agent is mapped to a semantic space, it is possible to generate natural language justifications for its actions [30].

Fairness is another critical dimension. LLMs are known to reflect the biases of their training data, which often includes the prejudices of the internet [21]. When these models are used to guide reinforcement learning in domains like hiring, loan processing, or urban planning, they risk automating and scaling systemic inequality [10]. Addressing this requires more than just algorithmic fixes; it requires a systemic approach to data governance [24]. This includes the curation of diverse datasets for both the LLM pre-training and the offline reinforcement learning phase, as well as the implementation of fairness-aware reward functions that explicitly penalize disparate impacts across different demographic groups [24].

6. Deployment Strategies and Sustainability

The practical deployment of cross-domain reasoning agents requires a phased approach that prioritizes stability over rapid innovation. We suggest a shadow deployment model where agents run in parallel with existing systems, providing recommendations without taking direct action [18]. This allows for the collection of high-fidelity data on the agent's reasoning performance in real-world conditions without risking system integrity. Over time, as the agent demonstrates reliable cross-domain generalization and adheres to conservative safety bounds, its level of autonomy can be incrementally increased [9].

Sustainability must also be considered in the context of the environmental cost of training and running large-scale AI [5]. The energy consumption of data centers is a growing concern, and the trend toward larger and larger models is not indefinitely sustainable [19]. Cross-domain reasoning offers a path toward more efficient AI by reducing the need for every organization to train its own massive models from scratch. By utilizing a common foundation of LLM representations, specialized agents can be trained on relatively small, domain-specific datasets using offline reinforcement learning, significantly reducing the total carbon footprint of AI development [1].

Furthermore, the longevity of these systems depends on their ability to handle concept drift—the gradual change in the environment over time [12]. An agent trained on 2024 economic data may be poorly suited for 2030 conditions. A cross-domain agent is naturally more resilient to this, as its underlying world-model is more general and less tied to the minutiae of a specific moment [25]. However, a continuous monitoring and updating infrastructure is still necessary [18]. This involves a feedback loop where real-world

outcomes are used to refine the conservative bounds of the agent, ensuring that it remains aligned with the evolving needs of the socio-technical system it serves.

7. Cross-Domain Comparative Analysis

To understand the efficacy of cross-domain reasoning, it is instructive to compare its application in vastly different sectors. In the financial sector, reasoning generalization can be used to transfer risk-assessment logic from traditional equity markets to emerging crypto-asset environments [11]. The underlying semantic concepts—liquidity, volatility, arbitrage—are consistent, and an LLM-guided agent can identify these patterns even when the numerical data distributions differ significantly [30]. In this domain, the conservative element is crucial to prevent black swan events where the model overextends its confidence in a novel market scenario [17].

Contrast this with the healthcare domain, where cross-domain reasoning might involve applying surgical workflow optimization to general hospital resource management [26]. Here, the semantic mapping is focused on concepts of patient throughput, sterility, and resource allocation [4]. The challenge is not just numerical but procedural and ethical. The LLM's understanding of medical ethics and clinical guidelines provides a necessary constraint on the reinforcement learning agent, ensuring that efficiency is never optimized at the cost of patient safety [31]. The comparison highlights that while the technical framework remains the same, the governing constraints vary based on the social importance of the domain.

A third case involves the management of environmental systems, such as watershed protection or reforestation projects [1]. These systems are characterized by long feedback loops where the results of an action may not be visible for years. Offline reinforcement learning is uniquely suited for this, as it can learn from decades of historical ecological data [22]. The LLM representations help the agent understand the complex causal links in ecological literature that are not captured in simple sensor readings [28]. In all three cases, the ability to reason across the boundaries of specific datasets allows for a more holistic and effective approach to system management [25].

8. Advanced Planning and High-Level Guidance

The complexity of cross-domain generalization necessitates a distinction between low-level tactical execution and high-level strategic planning. In many autonomous systems, failures occur because the agent becomes hyper-focused on immediate rewards at the expense of long-term stability [20]. Integrating LLM-based planning allows the system to set waypoints in the semantic space, which the reinforcement learning agent then attempts to reach through concrete actions [6]. This hierarchical structure mirrors human cognition, where we often have a verbalized plan that guides our physical movements without needing a step-by-step instruction for every muscle contraction [30].

A significant advancement in this area involves high-level planning guidance, where the LLM

functions as a high-level architect that generates reasoning trajectories [6]. These trajectories serve as a guide for the offline reinforcement learning process, effectively narrowing the search space and providing a semantic anchor for the agent's behavior [29]. By grounding the reinforcement learning agent in a pre-conceived plan, the risk of erratic, out-of-distribution behavior is further mitigated [16]. This is especially important in high-consequence environments like autonomous aviation or nuclear power plant cooling systems, where every action must be justifiable within a broader safety plan [2].

The use of planning guidance also improves the data efficiency of the offline learning process [6]. When the agent has a hint about which areas of the state space are semantically relevant to the goal, it can learn an effective policy from far less data than a tabula rasa agent would require [1]. This is a critical factor for domains where data is scarce or expensive to collect, such as rare disease research or deep-space exploration [18]. By leveraging the collective intelligence stored in language models, we can bootstrap the learning process, making sophisticated reasoning accessible to a wider range of applications [4].

9. Future Perspectives and Emergent Challenges

Looking toward the next decade, the evolution of cross-domain reasoning will likely move toward continual learning systems that evolve alongside their environments [22]. Rather than being trained and then deployed in two distinct phases, these agents will exist in a state of perpetual refinement [9]. The challenge will be to maintain the conservative safety bounds as the model's knowledge base grows [17]. We may see the emergence of federated reasoning, where multiple agents across different sectors share their learned semantic insights without sharing sensitive raw data, creating a collective intelligence that is greater than the sum of its parts [1].

There is also the potential for meta-reasoning, where the system begins to analyze its own reasoning processes for biases or errors [26]. If an agent detects that its cross-domain transfers are consistently failing in a specific way, it could use its LLM capabilities to search for new theoretical frameworks or historical analogies to correct its course [30]. This level of self-awareness would represent a significant milestone in AI development, but it also introduces new risks regarding the predictability and control of such systems [2]. Ensuring that meta-reasoning remains aligned with human intent will be the primary challenge for AI researchers and ethicists alike [14].

Finally, we must consider the socio-political impact of agents that can reason across domains [32]. As these systems become more capable, they will inevitably begin to automate tasks that were previously thought to be the exclusive domain of human professionals—lawyers, engineers, and policy analysts [21]. This transition will require a major shift in our economic structures and our definition of work [5]. The goal of AI research should not just be to build more intelligent machines, but to build machines that augment human capability and contribute to a more equitable and resilient society [26].

10. Conclusion

The facilitation of cross-domain reasoning generalization through conservative offline reinforcement learning and LLM representations offers a robust framework for the next generation of autonomous systems. By grounding decision-making in the vast, semantic world-model of language models and constraining it with conservative value estimation, we can create agents that are both capable and safe. This paper has detailed the architectural requirements, the structural trade-offs, and the socio-technical implications of this approach. While significant challenges remain—particularly in the areas of governance, fairness, and computational sustainability—the path forward is clear.

The integration of AI into our critical infrastructure is not just a technical project; it is a fundamental restructuring of how our society functions. As we move toward a world where reasoning is a distributed utility, we must ensure that the systems we build are transparent, accountable, and resilient. The combination of conservative learning and linguistic grounding provides the necessary tools to achieve this vision, bridging the gap between the digital and physical worlds in a way that respects the complexity and diversity of human life.

References

1. Agarwal, R., Schuurmans, D., & Norouzi, M. (2020). An optimistic perspective on offline reinforcement learning. International Conference on Machine Learning (ICML).

2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

3. Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. Communications of the ACM, 64(7), 58-65.

4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS).

5. Crawford, K. (2021). The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

6. Dou, Z., Zhao, Q., Wan, Z., Zhang, D., Wang, W., Raiyan, T., ... & Biswas, S. (2025). Plan Then Action: High-Level Planning Guidance Reinforcement Learning for LLM Reasoning. arXiv preprint arXiv:2510.01833.

7. Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. arXiv preprint arXiv:1802.06070.

8. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society.

Harvard Data Science Review.

9. Fujimoto, S., Meger, D., & Precup, D. (2019). Off-policy deep reinforcement learning without exploration. International Conference on Machine Learning (ICML).

10. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

12. Gulcehre, C., Wang, Z., Novikov, A., Paine, T., Gómez, S. G., Shahriari, B., ... & de Freitas, N. (2020). RL Unplugged: A suite of benchmarks for offline reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS).

13. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. International Conference on Machine Learning (ICML).

14. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

15. Kostrikov, I., Nair, A., & Levine, S. (2021). Offline reinforcement learning with implicit Q-learning. arXiv preprint arXiv:2110.06169.

16. Kumar, A., Fu, J., Soh, M., Tucker, G., & Levine, S. (2019). Stabilizing off-policy Q-learning via conservative offline distribution correction. Advances in Neural Information Processing Systems (NeurIPS).

17. Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS).

18. Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643.

19. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Re, C. (2022). Holistic evaluation of language models. Annals of the New York Academy of Sciences.

20. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

21. Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

22. Prudencio, R. F., Maximo, M. R., & Colombini, E. L. (2023). A survey on offline reinforcement learning: Taxonomy, review, and open problems. IEEE Transactions on Neural Networks and Learning Systems.

23. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report.

24. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly named bias audits. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

25. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Gabriel, V., ... & de Freitas, N. (2022). A generalist agent. Transactions on Machine Learning Research.

26. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

27. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140-1144.

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS).

29. Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Hubert, T., Soyer, O., Rezende, D. J., ... & Botvinick, M. (2016). Learning to reinforcement learn. arXiv preprint arXiv:1611.05763.

30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Xia, F., ... & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS).

31. Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., & Levine, S. (2019). Meta-world: A benchmark and evaluation for multi-task reinforcement learning. Conference on Robot Learning (CoRL).

32. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.