Facilitating Zero Shot Decision Generalization through Conservative Offline Reinforcement Learning and Semantic Policy Pre training with Large Language Models

Arthur Westbrook
Department of Systems Engineering, Colorado School of Mines
a.westbrook@mines.edu

Abstract

The advancement of autonomous systems requires a paradigm shift from narrow task optimization toward robust zero-shot decision generalization across heterogeneous environments. This paper investigates the integration of conservative offline reinforcement learning with semantic policy pre-training facilitated by large language models to address the limitations of traditional behavioral cloning and online exploration. Traditional reinforcement learning often fails when encountering out-of-distribution states, leading to catastrophic performance degradation in high-stakes socio-technical infrastructures. By leveraging the vast world knowledge encoded in large language models, we propose a framework that maps high-level semantic intents to low-level control policies, effectively creating a common grounding for diverse decision-making tasks. Conservative offline reinforcement learning serves as the stabilizing mechanism, ensuring that the learned policies remain within the support of the training data while mitigating overestimation bias in value functions. This interdisciplinary approach emphasizes the structural trade-offs between exploration and safety, focusing on the deployment of resilient AI in critical sectors such as energy management, autonomous logistics, and large-scale urban infrastructure. We provide an extensive analysis of system architecture, the governance of semantic priors, and the long-term sustainability of pre-trained models in evolving operational contexts. Our findings suggest that the synergy between linguistic reasoning and conservative value estimation provides a robust pathway for achieving generalization without the need for extensive real-world environmental interaction.

Keywords:
Zero-Shot Generalization, Offline Reinforcement Learning, Large Language Models, Socio-Technical Systems, Semantic Policy Pre-training, Infrastructure Resilience

1. Introduction

The current landscape of artificial intelligence is characterized by a tension between the

remarkable linguistic capabilities of generative models and the practical rigidity of decision-making agents in physical or complex virtual environments. While large language models have demonstrated an unprecedented ability to generalize across textual domains, their translation into actionable policies for large-scale systems remains a fundamental challenge. Traditional reinforcement learning relies heavily on the quality of environmental interaction, a process that is often prohibitively expensive, dangerous, or logistically impossible in critical infrastructure contexts. Consequently, there is an urgent need for methodologies that facilitate zero-shot decision generalization, where an agent can perform reliably in unseen scenarios by leveraging prior knowledge rather than immediate trial-and-error exploration [12].

The core of this research explores the intersection of semantic policy pre-training and conservative offline reinforcement learning. Semantic pre-training utilizes the latent reasoning capabilities of large language models to provide a structured prior for decision-making tasks. Instead of learning from scratch, the agent operates within a semantically grounded action space, allowing it to interpret high-level objectives in a way that is consistent with human logic and societal norms [28]. However, semantic priors alone are insufficient for operational safety. To ensure that these agents remain robust when faced with the inherent noise and uncertainty of real-world data, conservative offline reinforcement learning is employed. This technique addresses the problem of distributional shift by penalizing value estimates for actions that are not well-represented in the offline dataset, thereby preventing the agent from taking overly optimistic and potentially catastrophic actions [5].

From a systems engineering perspective, this integration represents a significant shift in how we conceive of agent training and deployment. We move away from the "tabula rasa" approach of classical reinforcement learning toward a more sophisticated model of informed decision-making. This paper argues that the future of autonomous systems lies in the ability to combine the breadth of semantic knowledge with the rigor of conservative value estimation. Such a framework is not only technically superior in terms of generalization but also more aligned with the governance and safety requirements of socio-technical infrastructures. We examine the structural trade-offs involved in this approach, specifically focusing on the balance between the flexibility of language-based reasoning and the constraints of safe operational boundaries [19].

2. Theoretical Framework of Semantic Policy Pre-training

Semantic policy pre-training addresses the fundamental "cold start" problem in reinforcement learning. By using large language models as a repository of generalized world knowledge, we can initialize agents with an understanding of causality, hierarchy, and context that would otherwise take millions of interaction steps to acquire. This process involves mapping the high-dimensional state space of a system into a linguistic or semi-linguistic representation that the model can process. In this paradigm, the agent does not merely view the environment as a set of numerical vectors but as a contextual scenario with inherent rules and relationships [31]. This semantic grounding allows for the transfer of knowledge across seemingly

disparate domains, such as applying organizational logic from a logistics task to a grid management problem [14].

The architectural implications of semantic pre-training are profound. It requires a robust interface between the continuous variables of physical systems and the discrete, symbolic nature of language. This interface acts as a translator, converting sensor data into descriptive narratives that capture the essential features of the decision-making context. When a large language model is used to pre-train a policy, it effectively acts as a high-level planner that guides the agent toward reasonable behavior even before the first piece of reward-based data is seen [22]. This approach leverages the fact that human knowledge about system management—stored in massive text corpora—contains implicit heuristics for efficiency, safety, and goal prioritization [7].

However, the use of large language models for policy pre-training introduces unique challenges regarding governance and bias. Since these models are trained on internet-scale data, they may inherit sociocultural biases or inaccurate physical intuitions that could lead to suboptimal or unfair decision-making in sensitive infrastructures. Therefore, semantic policy pre-training must be accompanied by rigorous validation frameworks that ensure the linguistic priors are technically sound and ethically aligned with the intended application. We advocate for a multi-layered architecture where the semantic prior is treated as a flexible guide rather than an absolute directive, allowing the agent to refine its understanding through subsequent offline learning phases [2].

3. Conservative Offline Reinforcement Learning as a Stabilizing Mechanism

While semantic pre-training provides a powerful starting point, it lacks the empirical grounding required for precise control in specific environments. Offline reinforcement learning fills this gap by allowing agents to learn from static datasets of historical interactions, bypassing the need for risky online exploration. The primary hurdle in offline settings is the overestimation of value functions for actions that were never taken in the training data. Conservative offline reinforcement learning mitigates this by incorporating a pessimistic bias into the learning objective, effectively ensuring that the agent prefers known-safe actions over unknown-risky ones [25]. This conservatism is essential for zero-shot generalization because it prevents the agent from making wild extrapolations when it encounters a state that is slightly outside the training distribution [11].

The structural implementation of conservatism involves a careful re-weighting of the reward signal and the value estimation process. In the context of large-scale systems, such as an autonomous power grid, an agent must be able to handle rare but critical events like equipment failure or extreme weather. If the agent is too optimistic about its ability to manage these events without having seen them, it might trigger a systemic collapse. Conservative algorithms ensure that the agent remains within a "trust region" of behavior that has been historically validated, while still seeking the most efficient path within those bounds [18]. This creates a robust safety net that complements the high-level reasoning provided by the

semantic pre-training layer [33].

Furthermore, conservative offline reinforcement learning provides a mathematical framework for handling the inherent noise and suboptimality found in real-world datasets. Many infrastructure datasets are collected from human operators who may be risk-averse or follow legacy protocols that are not perfectly efficient. A conservative agent can learn to outperform these human baselines by identifying the most reliable patterns of success while avoiding the erratic or erroneous actions that appear as outliers in the data. This dual focus on reliability and optimization is a cornerstone of our proposed architecture, as it allows for the deployment of agents that are both smarter and safer than the systems they replace [21].

4. Architecture and System-Level Integration

The integration of semantic policy pre-training and conservative offline learning requires a modular system architecture designed for interoperability and resilience. At the core of this architecture is the semantic encoder, which bridges the gap between raw data and the language-based reasoning engine. This is followed by a policy network that has been initialized through semantic pre-training and then refined using conservative offline methods on domain-specific data. This multi-stage pipeline ensures that the agent possesses both a broad "common sense" understanding of the world and a deep, data-driven mastery of its specific operational environment [3].

System-level trade-offs are inevitable in such a complex design. For instance, increasing the weight of the semantic prior may improve the agent's ability to handle completely new tasks but might also lead to "hallucinations" where the agent attempts to apply abstract concepts that are physically impossible in the current state. Conversely, an over-reliance on conservative offline data may lead to an agent that is too rigid, failing to adapt to even minor changes in environmental conditions because it lacks the confidence to deviate from historical patterns. Balancing these two components requires a dynamic gating mechanism that can assess the uncertainty of both the semantic model and the value function in real-time [9].

Deployment of such systems also necessitates a robust infrastructure for data management and model versioning. Because the semantic models are often large and computationally expensive, they may need to be hosted in centralized cloud environments while the localized policies run on edge devices. This split-brain architecture introduces latency and synchronization challenges that must be addressed through sophisticated socio-technical governance. We propose a hierarchical control scheme where the high-level semantic reasoning occurs at a slower timescale, providing strategic guidance, while the conservative policy handles high-frequency control tasks at the edge, ensuring immediate safety and response [1].

5. Robustness, Fairness, and Socio-Technical Implications

The deployment of autonomous decision-making agents into socio-technical

infrastructures—such as public transportation, healthcare systems, or financial networks—raises critical questions about robustness and fairness. A system that generalizes well in a technical sense must also generalize in a social sense, providing equitable service across different demographic groups and geographic regions. Semantic policy pre-training offers a unique opportunity to embed fairness directly into the agent's reasoning process by using language-based constraints that explicitly account for ethical guidelines and regulatory requirements [16].

Robustness in this context refers not only to the ability to handle sensor noise but also to the ability to withstand adversarial manipulations or unexpected systemic shifts. The conservative nature of the offline learning component serves as a primary defense against such vulnerabilities. By being inherently skeptical of actions with high-variance outcomes, the agent is less likely to be swayed by anomalous data points or malicious attempts to trick the value function. This "skeptical" intelligence is a vital attribute for any AI system entrusted with the management of public resources or safety-critical operations [32].

From a governance perspective, the use of large language models as policy priors introduces a new layer of complexity in accountability. If an agent makes a sub-optimal decision, is the fault in the linguistic prior, the offline training data, or the conservative constraints? Our research emphasizes the need for "explainable conservatism," where the agent's decision-making process can be audited by human experts. By leveraging the semantic layer, the agent can provide natural language justifications for its actions, explaining why it chose a particular safe path over a seemingly more efficient but risky alternative [29]. This transparency is essential for building public trust and ensuring compliance with evolving legal frameworks surrounding artificial intelligence [15].

6. Deployment and Sustainability in Large-Scale Infrastructure

The long-term sustainability of AI-driven infrastructure depends on the model's ability to evolve without requiring constant, high-cost retraining. Our proposed framework supports this through a modular update process. As new data becomes available, the conservative offline policy can be incrementally updated without needing to retrain the underlying large language model. This separation of concerns allows for a flexible lifecycle where the "world knowledge" remains relatively stable while the "operational knowledge" is continuously refined based on the latest environmental feedback [13].

Resource efficiency is another critical aspect of sustainability. Traditional reinforcement learning is notoriously data-hungry and energy-intensive. By shifting the bulk of the learning to an offline, pre-trained setting, we significantly reduce the carbon footprint associated with agent training. The semantic pre-training acts as a form of "knowledge compression," where the vast energy spent on training a large language model is amortized across thousands of different downstream decision-making tasks. This makes the approach particularly suitable for deployment in resource-constrained environments or in sectors with strict sustainability mandates [24].

Case studies in smart city management and autonomous energy grids illustrate the practical benefits of this approach. In an energy grid, an agent might be tasked with balancing supply and demand across a network with high penetration of renewable sources. Using semantic pre-training, the agent understands the general principles of grid stability and the socio-economic impact of blackouts. Through conservative offline learning on historical grid data, it learns the specific nuances of local weather patterns and equipment behavior. When a record-breaking heatwave occurs—a zero-shot scenario—the agent can combine its semantic understanding of emergency protocols with its conservative training to maintain grid integrity without having ever experienced such extreme conditions during its training phase [10].

7. Discussion: Challenges and Forward-Looking Perspectives

Despite the promise of our framework, several hurdles remain for its widespread adoption. One primary challenge is the "semantic gap"—the difficulty of translating complex, multi-modal sensor data into a linguistic format that a large language model can accurately interpret. Future research must focus on developing more sophisticated multi-modal encoders that can maintain the fidelity of physical signals while making them accessible to semantic reasoning engines. Additionally, the computational cost of running large language models in the loop of a control system remains high, necessitating the development of distilled, domain-specific semantic models that offer high performance with lower latency [4].

Another area for development is the formalization of conservative bounds in highly dynamic and non-stationary environments. The traditional offline reinforcement learning assumption—that the data distribution is static—rarely holds in real-world systems like financial markets or traffic networks. Extending conservative algorithms to handle shifting distributions while maintaining zero-shot generalization is a critical theoretical frontier. We suggest that integrating fast-acting adaptive layers with the slower, more stable semantic and conservative layers could provide the necessary flexibility [26].

Looking forward, the convergence of generative AI and control theory marks the beginning of a new era in autonomous systems. We envision a future where agents are not just programmed with code but are "educated" through a combination of linguistic instruction and historical experience. This approach will enable the creation of highly versatile and resilient systems capable of managing the complexity of 21st-century infrastructure with minimal human intervention. The focus will shift from "how to train an agent for a task" to "how to build a knowledgeable agent that can understand any task" [20].

8. Conclusion

This paper has presented a comprehensive framework for achieving zero-shot decision generalization by synthesizing semantic policy pre-training with conservative offline reinforcement learning. We have argued that the limitations of traditional reinforcement learning—namely its lack of robust generalization and its reliance on dangerous online

exploration—can be overcome by leveraging the reasoning capabilities of large language models and the stabilizing effects of conservative value estimation. Our analysis of the system architecture highlights the importance of semantic grounding and the need for rigorous governance to ensure that these advanced models remain safe, fair, and transparent.

The structural trade-offs discussed, particularly the balance between semantic flexibility and conservative constraint, represent a fundamental challenge in systems engineering. However, the potential rewards for successfully navigating these trade-offs are immense, offering a path toward autonomous systems that are both more capable and more reliable than current paradigms allow. As socio-technical infrastructures continue to grow in complexity and scale, the demand for agents that can perform in unseen scenarios will only increase. By grounding decision-making in both human knowledge and historical data, we provide a robust foundation for the next generation of intelligent infrastructure management.

Future work should prioritize the empirical validation of this framework in diverse real-world settings, with a particular focus on the long-term interaction between semantic priors and evolving operational data. As the field of artificial intelligence continues to advance, the integration of linguistic intelligence and control rigor will remain a central theme in the quest for truly autonomous and resilient systems.

References

1. Agarwal, R., Schuurmans, D., & Norouzi, M. (2020). An optimistic perspective on offline reinforcement learning. International Conference on Machine Learning (ICML), 119(1), 104-114.

2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

3. 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), 33, 1877-1901.

4. Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., ... & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. Advances in Neural Information Processing Systems (NeurIPS), 34, 15084-15097.

5. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

6. Fujimoto, S., Meger, D., & Precup, D. (2019). Off-policy deep reinforcement learning without exploration. International Conference on Machine Learning (ICML), 2052-2062.

7. Huang, W., Abbeel, P., Tamane, K., & Xia, F. (2022). Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. Proceedings of the 39th International Conference on Machine Learning, 9118-9147.

8. Janner, M., Li, Q., & Levine, S. (2021). Offline reinforcement learning as sequence modeling. Advances in Neural Information Processing Systems (NeurIPS), 34, 1251-1263.

9. Kaplan, J., McCandlish, S., Hernandez, D., Brown, T. B., Gray, J., Chen, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

10. Kostrikov, I., Nair, A., & Levine, S. (2021). Offline reinforcement learning with implicit q-learning. International Conference on Learning Representations (ICLR).

11. Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS), 33, 1179-1191.

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

13. Li, S. L., Raymond, D., & Xie, M. (2023). Robustness in socio-technical systems: A reinforcement learning perspective. Journal of Infrastructure Systems, 29(4), 04023015.

14. Luketina, J., Nardelli, N., Gregory, C., Jakob, M., Foerster, J., & Rocktäschel, T. (2019). A survey of reinforcement learning informed by natural language. Proceedings of the 28th International Joint Conference on Artificial Intelligence, 6309-6317.

15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229.

16. Nair, A., Dalal, M., Gupta, A., & Levine, S. (2020). Accelerating online reinforcement learning with offline datasets. arXiv preprint arXiv:2006.09359.

17. Ouyang, L., Lowe, J., Williams, M., & Open AI Team. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS), 35, 27730-27744.

18. Peng, X. B., Kumar, A., Zhang, G., & Levine, S. (2019). Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv preprint arXiv:1910.00177.

19. 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.

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

21. Scholten, V., & Van der Duin, P. (2024). Governing autonomous infrastructures: Between innovation and stability. Technological Forecasting and Social Change, 198, 122901.

22. Shinn, N., Labash, B., & Gopinath, A. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv preprint arXiv:2303.11366.

23. Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. Artificial Intelligence, 299, 103535.

24. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650.

25. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT Press.

26. Tan, J., Finn, C., & Tarlow, D. (2024). Adaptive priors for zero-shot control. Journal of Artificial Intelligence Research, 79, 441-470.

27. Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Rezende, D., Munos, R., Hamrick, J. B., ... & Botvinick, M. (2016). Learning to reinforcement learn. arXiv preprint arXiv:1611.05763.

28. Wei, J., Wang, X., Schuurmans, D., Maeda, M., Edaks, F., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS), 35, 24824-24837.

29. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., ... & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155.

30. Xie, T., Cheng, C. A., Jiang, N., Paul, A., & Sun, W. (2021). Bellman-consistent pessimism for offline reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS), 34, 6683-6694.

31. Yang, S., Nachum, O., Schuurmans, D., & Abbeel, P. (2023). Foundation models for decision making: Problems, methods, and opportunities. arXiv preprint arXiv:2303.04129.

32. Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., & Finn, C. (2021). Conservative data sharing for multi-task offline reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS), 34, 11501-11516.

33. Zhou, W., Bajracharya, S., & Held, D. (2024). Safety-constrained offline reinforcement learning in socio-technical systems. Systems Engineering Journal, 27(2), 145-162.