Integrating Large Language Models with Reinforcement Learning for Generalization in Strategic Card Games

Extended Abstract

Wannian Xia*
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
xiawannian2020@ia.ac.cn

Meng Fang
University of Liverpool
Liverpool, United Kingdom
Meng.Fang@liverpool.ac.uk

Zihao Guo
King's College London
London, United Kingdom
zihao.1.guo@kcl.ac.uk

Yali Du
King's College London
London, United Kingdom
yali.du@kcl.ac.uk

Bo Xu
Institute of Automation, Chinese Academy of Sciences
Beijing, China
xubo@ia.ac.cn

ABSTRACT

Strategic card games, such as Hearthstone, offer a rich environment for exploring decision-making in reinforcement learning (RL). Yet, achieving generalization across diverse and evolving game scenarios remains a significant challenge. In this paper, we introduce a novel framework that integrates large language models (LLMs) with RL agents to improve generalization in strategic card games. Our approach leverages a fine-tuned T5 model to encode and interpret card strategies expressed in natural language, facilitating efficient policy learning across a large and continually expanding set of cards. Employing a self-play RL framework augmented with an auxiliary transition loss in the latent space, our agent captures and generalizes the complex, dynamic nature of card interactions. Experimental results show that our method not only enhances learning efficiency but also significantly improves the agent's ability to generalize, maintaining robust performance when encountering new cards.

KEYWORDS

Reinforcement Learning; Large Language Models; Self-Play Agents; World Model

ACM Reference Format:
Wannian Xia, Meng Fang, Zihao Guo, Yali Du, and Bo Xu. 2025. Integrating Large Language Models with Reinforcement Learning for Generalization in Strategic Card Games: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Language plays a fundamental role in human cognition and intelligence [4, 7–10] and is crucial in environments that merge natural language processing (NLP) with decision-making tasks, such as text-based and strategic card games [1, 3, 6, 13, 15]. These games provide a fertile ground for studying reinforcement learning (RL), as agents must interpret and act on textual information [2, 11] while continuously adapting to evolving game dynamics.

However, achieving robust generalization in strategic card games remains a significant challenge. RL agents must quickly assimilate new information, interpret complex language descriptions, and adjust their strategies to maintain optimal performance. This challenge is particularly acute in strategic card games like Hearthstone, where expanding card pools and continuously introduced new effects create a highly variable and evolving environment.

In this work, we introduce a novel framework that integrates a fine-tuned LLM with RL to improve generalization in Hearthstone. Our goal is to enable the agent to learn from diverse scenarios and generalize to evolving game dynamics. We design our agent using the encoder of a T5-base model [12], which has been fine-tuned on Hearthstone card and deck datasets. The fine-tuned T5 model processes game cards as tokens to generate corresponding natural language descriptions and encodes deck strategies to produce summaries of deck contents. While training the agent in a self-play manner using Monte Carlo sampling [5, 16], we incorporate a transition loss in the latent space as an auxiliary component. This auxiliary loss compels the agent to learn the inherent dynamics of the environment, as described in natural language and embedded by the fine-tuned T5 model. Our approach is the first to achieve high-level performance across the scale of available cards and diverse decks, maintaining a robust policy even when encountering new cards and demonstrating critical zero-shot performance. We release our dataset and code for further research [1].

2 APPROACH

2.1 Card Representation Learning using T5

We leverage a massive Hearthstone card dataset to capture card characteristics and deck strategies. To achieve this, we fine-tune a T5-base model and create two datasets: a card dataset and a deck

*Work done during his visit at King's College London

[1] https://github.com/WannianXia/GeneralizableHS

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

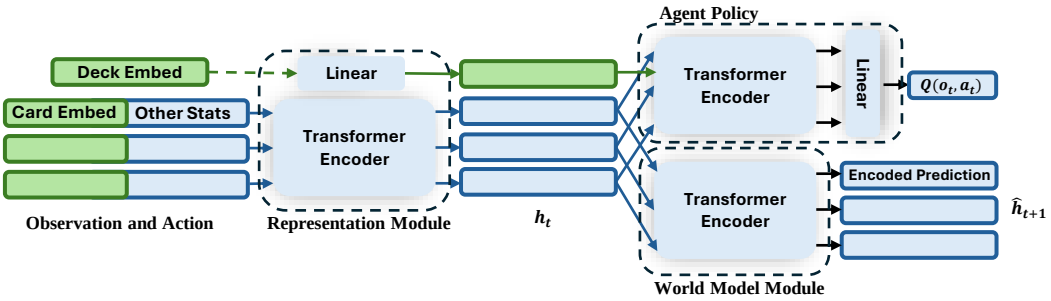

Figure 1: Architecture of Our Model.

strategy dataset. For the card dataset, each card is provided as input with its description as the output. We treat each card as a unique token. For the deck strategy dataset, we convert the contents of over 5,000 decks—including card names, descriptions, mana costs, attack values, health points, etc.—into a formal language description. We then prompt GPT-4 to generate concise, human-like strategy summaries for each deck. The generated deck strategy serves as the input, and its description as the output. We fine-tune T5 with these datasets to capture the semantics and effects of cards and learn the connections between strategy and card selection. Based on the fine-tuned T5 model, we use the T5 encoder to obtain representations for each card and deck strategy, applying additional mean pooling and an L2 normalization layer.

2.2 Self-Play Policy Training with Auxiliary Loss

We train our agent through self-play using Monte Carlo sampling and augment the training with an auxiliary loss from a world model, which incorporates future state information into the hidden representations.

Let O denote the set of observations and $\mathcal{A}$ the set of actions. At each time step t, the agent receives an observation $o_t \in O$, takes an action $a_t \in \mathcal{A}$, and transitions to o_{t+1}. We use representations based on the fine-tuned T5 encoder. As shown in Figure 1, our agent is composed of three key components:

(1) **Representation module:** A network f_θ encodes the cards and statistics into a hidden state: $h_t = f_\theta(o_t, a_t)$.
(2) **Policy module:** A network q_ϕ uses deck information, cards, and statistics to predict the action-value: $Q(o_t, a_t) = q_\phi(h_t)$.
(3) **World Model module:** A network w_ψ that forecasts the next hidden state: $\hat{h}_{t+1} = w_\psi(h_t)$.

The agent collects trajectories $\tau = \{(o_t, a_t, o_{t+1})\}$, and after each game, the discounted returns G_t are computed and stored in a buffer $\mathcal{D}$. The policy model is optimized by minimizing the mean squared error between the predicted Q-values and the returns:

$$\mathcal{L}_{\text{policy}}(\theta, \phi) = \mathbb{E}_{(o_t, a_t, G_t)\sim\mathcal{D}}\left[\left(q_\phi(f_\theta(o_t, a_t)) - G_t\right)^2\right].$$

For the world model module, we create a target hidden state $\bar{h}_{t+1}$ by encoding the next observation with the action masked:

$$\bar{h}_{t+1} = f_\theta(o_{t+1}, \text{mask}(a_{t+1})).$$

Table 1: Win rates compared to basic tree search AI of different levels on testing decks.

Tree Search	Our Agent	GPT-4o	Cardsformer
Level 0	**96%**	60%	82%
Level 1	**76%**	24%	68%
Level 2	**68%**	12%	48%

The world model module is trained by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{world}}(\theta, \psi) = \text{CrossEntropy}\Big(w_\psi(f_\theta(o_t, a_t)), \bar{h}_{t+1}\Big).$$

The overall loss is a weighted combination of the two:

$$\mathcal{L}_{\text{total}}(\theta, \phi, \psi) = \mathcal{L}_{\text{policy}}(\theta, \phi) + \beta\, \mathcal{L}_{\text{world}}(\theta, \psi),$$

with β regulating the influence of the world model loss.

3 EXPERIMENTS

We use the Hearthstone game for our experiments. We evaluate the zero-shot generalization capability of the agent in different game scenarios. To assess the agent's generalization ability, we design five testing decks with varying proportions of unseen cards; in one testing deck, up to 50% of the cards are unseen. Table 1 summarizes our results. Our agent achieves the highest win rates across all settings, outperforming even the most extensive tree search AI and other baselines, including GPT-4o (a direct LLM agent prompted with game observations as text) and Cardsformer [14]. Although GPT-4o performs well in simple scenarios, its win rate drops to 12% when tree search parameters are increased, demonstrating that LLMs alone struggle with competitive Hearthstone strategies.

4 CONCLUSIONS

We have presented a novel framework that integrates a fine-tuned T5-base model with reinforcement learning to enhance policy learning and generalization in Hearthstone. Evaluated on challenging test decks featuring unseen cards, our approach consistently outperforms current baselines. In future work, we plan to explore even tighter integration between the agent and the LLM to further improve performance and adaptability.

REFERENCES

[1] Krishnendu Chatterjee and Rasmus Ibsen-Jensen. 2016. The complexity of deciding legality of a single step of Magic: The Gathering. In *ECAI 2016*. IOS Press, 1432–1439.

[2] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. TextWorld: A Learning Environment for Text-based Games. *CoRR* abs/1806.11532 (2018).

[3] Fernando de Mesentier Silva, Rodrigo Canaan, Scott Lee, Matthew C Fontaine, Julian Togelius, and Amy K Hoover. 2019. Evolving the hearthstone meta. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[4] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 5, 3 (2023), 220–235.

[5] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*. PMLR, 1407–1416.

[6] Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17985–17993.

[7] Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to Active Learn: A Deep Reinforcement Learning Approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 595–605.

[8] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13, 1 (2022), 3094.

[9] Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14, 1 (2024), 3440.

[10] Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences* 120, 13 (2023), e2215907120.

[11] Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Pushkar Shukla, Sadhana Kumaravel, Gerald Tesauro, Kartik Talamadupula, Mrinmaya Sachan, and Murray Campbell. 2021. Text-based rl agents with commonsense knowledge: New challenges, environments and baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9018–9027.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[13] Adam Summerville and Michael Mateas. 2016. Mystical tutor: A magic: The gathering design assistant via denoising sequence-to-sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 12. 86–92.

[14] Wannian Xia, Yiming Yang, Jingqing Ruan, Dengpeng Xing, and Bo Xu. 2023. Cardsformer: Grounding Language to Learn a Generalizable Policy in Hearthstone. In *ECAI 2023*. IOS Press, 2720–2727.

[15] Yunqiu Xu, Meng Fang, Ling Chen, Yali Du, and Chengqi Zhang. 2021. Generalization in Text-based Games via Hierarchical Reinforcement Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 1343–1353.

[16] Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, and Ji Liu. 2021. Douzero: Mastering doudizhu with self-play deep reinforcement learning. In *international conference on machine learning*. PMLR, 12333–12344.