Multi-Armed Bandits Meet Large Language Models

Djallel Bouneffouf, Raphael Feraud

IBM Research, Orange Lab
Djallel.bouneffouf@ibm.com, Raphael.feraud@orange.com

Abstract

Bandit algorithms and Large Language Models (LLMs) have emerged as powerful tools in artificial intelligence, each addressing distinct yet complementary challenges in decision-making and natural language processing. This survey explores the synergistic potential between these two fields, highlighting how bandit algorithms can enhance the performance of LLMs and how LLMs, in turn, can provide novel insights for improving bandit-based decision-making. We first examine the role of bandit algorithms in optimizing LLM fine-tuning, prompt engineering, and adaptive response generation, focusing on their ability to balance exploration and exploitation in large-scale learning tasks. Subsequently, we explore how LLMs can augment bandit algorithms through advanced contextual understanding, dynamic adaptation, and improved policy selection using natural language reasoning. By providing a comprehensive review of existing research and identifying key challenges and opportunities, this survey aims to bridge the gap between bandit algorithms and LLMs, paving the way for innovative applications and interdisciplinary research in AI.

Introduction

LLMS such as GPT (Brown et al. 2020), BERT (Devlin et al. 2019), and T5 (Raffel et al. 2020) have revolutionized natural language processing by excelling in tasks like text generation, summarization, and dialogue systems. However, optimizing LLMs for specific applications often involves challenges like balancing exploration (testing novel strategies) and exploitation (leveraging learned strategies). Bandit algorithms (Bouneffouf, Rish, and Aggarwal 2020; Bouneffouf 2023; Boufelja-Yacoubi, Bouneffouf, and Zhuk 2024; Tchrakian et al. 2025b), designed to address such trade-offs, provide a powerful framework for optimizing decision-making under uncertainty. By integrating bandit approaches into the training, tuning, and application of LLMs, researchers can achieve dynamic adaptability, efficiency, and enhanced performance.

Conversely, LLMs possess the ability to process and generate human-like text, enabling them to provide contextual insights and reasoning. This capability can enhance bandit algorithms by incorporating richer context, natural language understanding, and user feedback into the decision-making process. The interplay between these two fields opens the door to innovative AI applications that are both adaptive and contextually aware.

While Bandit algorithms and LLMs have independently demonstrated success across various domains, the exploration of their intersection remains underdeveloped. By investigating how these paradigms complement each other, researchers can unlock new possibilities, such as improving LLM performance in personalized recommendations or utilizing LLMs to address the contextual limitations of bandit-based decision systems. This exploration not only advances the theoretical understanding of both fields but also provides practical solutions for complex real-world problems, including adaptive dialogue systems, dynamic content generation, and human-AI collaboration.

This survey aims to provide a comprehensive overview of the emerging synergy between Bandit algorithms and LLMs (Bouneffouf and Féraud 2024). It explores:

- How Bandit algorithms can enhance the efficiency and adaptability of LLMs,
- How LLMs can contribute to the contextual and adaptive decision-making capabilities of Bandit algorithms,
- Applications where the integration of these paradigms offers tangible benefits,
- Discussion of open problems, research gaps, and potential future directions in combining Bandit algorithms and LLMs.

By synthesizing the current state of research and offering a roadmap for future exploration, this paper aims to serve as a foundational resource for researchers and practitioners interested in the intersection of Bandit algorithms and LLMs.

Leveraging Bandit to Enhance LLM

The integration of Bandit algorithms into machine learning and natural language processing has been an active area of research. While traditional applications of Bandit algorithms include many domains such as online recommendation systems (Idé et al. 2025; Chen et al. 2024), healthcare (Lin et al. 2020), automated machine learning (Riemer et al. 2019), question answering (Bouneffouf et al. 2021), telecommunication (Zafar et al. 2023; Delande et al. 2021), material science (Kishimoto et al. 2022), Reinforcement Learning (Noothigattu et al. 2019), recent works have explored their

role in optimizing LLMs. This section reviews key contributions in the following areas: fine-tuning and training, prompt optimization, adaptive response generation, and evaluation strategies.

Fine-Tuning and Training

Fine-tuning large-scale language models is a resource-intensive process that requires selecting the most informative data samples, optimizing hyperparameters, and ensuring generalization while keeping computational costs manageable. Traditional methods often rely on heuristics or grid search techniques, but recent advances have demonstrated that multi-armed bandits (MABs) and contextual bandits can enhance various aspects of the fine-tuning process.

Active Learning for Data Selection One of the critical challenges in fine-tuning LLMs is determining which data samples contribute most for improving the model. Instead of training on all available data, which can be computationally expensive and redundant, Bandit algorithms have been employed for active learning to prioritize the most informative data points.

Uncertainty-Based Sampling: Bandit models help identify samples where the model is most uncertain, selecting them for fine-tuning to improve performance in underrepresented areas (Settles 2009; Chen, Golrezaei, and Bouneffouf 2025; Bouneffouf 2025b; Bouneffouf, Feraud, and Lin 2025).

Diversity-Promoting Selection: Contextual bandits can be used to balance between selecting uncertain samples and ensuring diverse coverage of the data distribution, preventing biases in model adaptation(Xia et al. 2024b; Chang et al. 2023; Zeng, Chen, and Jin 2023; Bouneffouf et al. 2014; Wahed, Gruhl, and Lourentzou 2023; Bouneffouf 2025a; Tchrakian et al. 2025a).

Cost-Efficient Learning: By dynamically allocating resources to the most impactful data points, Bandit-based active learning reduces the amount of labeled data required while maintaining or improving model performance (Li 2025).

Hyperparameter Tuning Fine-tuning performance is highly dependent on selecting optimal hyperparameters such as learning rates, batch sizes, and dropout rates. Traditional methods like grid search or random search can be inefficient, whereas Bandit-based optimization methods, provide a more adaptive approach.

Bayesian Optimization for Hyperparameters: Framed as a bandit problem, Bayesian optimization iteratively refines hyperparameter choices by exploring promising configurations while minimizing unnecessary trials (Li et al. 2018; Snoek, Larochelle, and Adams 2012). Multi-Fidelity Optimization: Some Bandit approaches allocate resources dynamically to hyperparameter settings that show early promise, reducing wasteful full-scale evaluations (Mulakala et al. 2024).

Adaptive Gradient-Based Methods Beyond data selection and hyperparameters, Bandit algorithms have also been applied to optimize the learning process itself by improving gradient-based optimization techniques.

Thompson Sampling for Gradient Updates: Bandit-inspired approaches such as Thompson Sampling adaptively adjust gradient update rules, improving convergence rates and reducing overfitting (Liu, Wu, and Mozafari 2020). Exploration in Optimization Strategies: By dynamically selecting among different optimization techniques (e.g., Adam, RMSProp, SGD), Bandit models can guide learning toward more efficient convergence patterns (Song et al. 2024). Adaptive Batch Size Selection: Bandit algorithms have been proposed to adjust batch sizes in real time, optimizing trade-offs between convergence speed and computational efficiency (Lisicki, Nica, and Taylor 2023).

Impact on Fine-Tuning Efficiency The integration of Bandit-based decision-making into fine-tuning strategies has demonstrated significant improvements in both data efficiency and training performance for LLMs. By leveraging exploration-exploitation principles, these approaches lead to faster convergence with reduced training time, improved model generalization through selective data exposure, and more efficient allocation of computational resources. Future research can explore for instance, the selection of LLM experts for reducing the query cost, or the use of Bandit-driven curriculum learning to further enhance the adaptability and robustness of LLM training(Lisicki, Nica, and Taylor 2023).

Prompt Optimization

Prompt engineering plays a crucial role in determining the quality, relevance, and coherence of responses generated by large language models (LLMs). Manually crafting optimal prompts requires domain expertise and iterative experimentation, making it a time-consuming process. Bandit algorithms provide a systematic way to dynamically optimize prompts by continuously exploring different formulations and selecting those that yield the best performance based on predefined reward signals.

Multi-Armed Bandit for Dynamic Prompt Selection Recent research has formulated prompt selection as a multi-armed bandit (MAB) problem, where different prompt variants represent different "arms," and the LLM's response quality serves as the reward signal.

Exploration vs. Exploitation: Instead of relying solely on predefined prompts, Bandit-based approaches allow for continuous exploration of new formulations while exploiting the most successful ones. This adaptive strategy helps identify prompts that produce higher precision, coherence, and information (Shi et al. 2024; Bouneffouf 2025e,c).

Automated Prompt Discovery: MAB algorithms can iteratively refine prompts by adjusting keywords, phrasing, or sentence structure, optimizing for factors such as fluency, factual consistency, and response diversity (Gao et al. 2025; Kishimoto et al. 2025).

Evaluation Metrics as Rewards: Bandit models can use various reward functions based on response quality metrics, including BLEU scores (for linguistic similarity), factual consistency (measured via external verification models), or human feedback ratings(Grams, Betz, and Bartelt 2025; Bouneffouf 2025d).

Chain-of-Thoughts To enhance the ability of large language models (LLMs) to tackle complex reasoning prob-

lems, chain-of-thought (CoT) methods have been introduced to guide LLMs through step-by-step reasoning, enabling problem-solving from simple to complex tasks (Wei et al. 2022). State-of-the-art approaches to generate these reasoning chains are based on interactive collaboration, where the learner produces intermediate candidate thoughts that are evaluated by the LLM, influencing the generation of subsequent steps .

However, a significant, yet underexplored challenge is that LLM-generated evaluations are often noisy and unreliable, which can mislead the selection of promising intermediate thoughts. To address this issue, the authors in (Zhang et al. 2024) propose a bandit-based pairwise comparison framework instead of conventional point-wise scoring. In each iteration, the intermediate thoughts are randomly paired, and the LLM is directly prompted to select the most promising option from each pair. This iterative comparison process, framed as a dueling bandit problem (Urvoy et al. 2013; Idé et al. 2025), allows for adaptive exploration and exploitation of promising reasoning paths while reducing the impact of noisy feedback.

Contextual Bandits for Personalized Prompting Beyond general prompt selection, optimizing prompts for specific users or contexts is essential for improving interaction quality. Contextual bandits extend traditional MAB models by incorporating contextual features such as user intent, domain-specific requirements, and interaction history.

Adaptive Prompt Tuning: Contextual bandits dynamically adjust prompt attributes like length, specificity, and style based on the user's query type and preferences (Lau et al. 2024; Chen, Chen, and Buet-Golfouse 2024).

Personalized User Interaction: By analyzing past interactions, contextual bandit models can tailor prompts to individual users, ensuring more relevant and engaging responses (Dai et al. 2025).

Domain-Specific Adaptation: In specialized fields such as legal or medical AI applications, contextual bandits can fine-tune prompts to align with domain-specific jargon and information retrieval needs (Upadhyay et al. 2019).

Impact on Prompt Engineering Efficiency By leveraging Bandit algorithms, LLMs can be automatically guided toward more effective and adaptive prompt formulations, reducing reliance on manual prompt engineering. The key benefits of Bandit-based prompt optimization include:

- improved Response Quality, continuous refinement leads to prompts that generate more accurate and contextually relevant responses, - reduced Trial-and-Error Costs, automated exploration of prompt variations decreases the need for extensive manual experimentation, - scalability across domains, Bandits can adapt prompts for various applications, from chatbots and virtual assistants to domain-specific AI systems. Future research directions could explore integrating hierarchical Bandit models for multi-step prompt optimization, and developing hybrid approaches that combine rule-based heuristics with adaptive learning strategies.

Adaptive Response Generation

Generating high-quality responses requires balancing creativity and relevance, especially in conversational AI and dialogue systems. Bandit-based approaches have been used to dynamically adjust generation strategies based on user interactions and feedback.

Exploration-Exploitation Trade-Off in Response Selection: Thompson Sampling and Upper Confidence Bound (UCB) algorithms have been employed to explore diverse response strategies while gradually shifting towards more rewarding (i.e., coherent, engaging) responses (Xia et al. 2024b). Adaptive Sampling for Diversity: Multi-armed bandits have been used to balance novelty vs. coherence, ensuring that generated responses are neither too predictable nor too random (Hoveyda et al. 2024). Conversational Personalization: Bandit-driven methods allow for adaptive dialogue generation by continuously learning from user interactions and refining response styles accordingly (Cai et al. 2021).

These studies indicate that Bandit algorithms can improve LLM-generated responses by continuously adapting to changing user preferences and real-time feedback.

Evaluation Strategies

Evaluating the quality of outputs generated by large language models (LLMs) presents significant challenges, as assessment criteria such as fluency, coherence, factual accuracy, and user preference can be highly subjective. Traditional evaluation methods often rely on human annotations, rule-based metrics, or pre-trained scoring models, all of which have limitations in scalability, consistency, and adaptability. Bandit-based approaches provide a promising solution by dynamically optimizing evaluation strategies, reducing human annotation effort while improving the quality and efficiency of feedback collection (Xia et al. 2024a).

Reinforcement Learning with Human Feedback (RLHF): the human preferences guide model optimization (Kaufmann et al. 2023). While RLHF typically relies on policy optimization techniques, incorporating Bandit-based exploration strategies can enhance its efficiency: traditional RLHF may overfit to specific types of human feedback, but Bandit-based approaches ensure a better balance between learning from past feedbacks and exploring new feedbacks.

Optimizing Reward Model Updates: Bandit-inspired strategies help in dynamically adjusting the weight given to different types of feedback (e.g., explicit user ratings vs. implicit engagement signals).

Dynamic Reward Adjustment: As models improve over time, bandit algorithms dynamically update evaluation criteria to focus on emerging weaknesses, ensuring that reinforcement learning continues to drive meaningful improvements (Yang et al. 2024).

Adapting Metrics Based on Task-Specific Performance: Different tasks (e.g., summarization, translation, open-ended generation) require different evaluation criteria. Bandit algorithms adaptively select the most relevant metrics for each task (Xia et al. 2024b).

Continuous Optimization of Evaluation Pipelines: Instead of relying on static metric weighting, bandit algorithms iter-

atively adjust how much weight is assigned to each metric based on real-world feedback (Wu et al. 2024).

Summary

The Key Takeaways from Table 1:

The integration of Bandit algorithms into Large Language Models (LLMs) has demonstrated significant advancements across multiple aspects of model optimization, including fine-tuning, prompt engineering, adaptive response generation, and evaluation strategies.

Fine-Tuning and Training: Traditional fine-tuning methods rely on exhaustive data selection and hyperparameter tuning, often leading to inefficiencies. Bandit algorithms, particularly Multi-Armed Bandits (MABs) and contextual bandits, offer an adaptive alternative by prioritizing data samples based on uncertainty and diversity, optimizing hyperparameters dynamically, and refining gradient-based methods. These approaches improve model generalization, reduce computational costs, and accelerate convergence.

Prompt Optimization: Prompt engineering plays a critical role in LLM performance, yet manual tuning is labor-intensive. Bandit-based strategies, such as dynamic prompt selection, enable continuous exploration of different prompt formulations while exploiting the most effective ones. Contextual bandits further personalize prompt adaptation based on user preferences and domain-specific requirements. Additionally, a dueling bandit framework mitigates noisy LLM-generated evaluations in chain-of-thought reasoning.

The increasing focus on personalization: the ability to tailor responses and strategies based on individual user preferences. Bandit algorithms, particularly contextual bandits, are effective tools for enabling this personalization by dynamically adjusting to different user needs and contexts. This customization enhances user experiences and optimizes interactions, offering more relevant and targeted outputs.

Reinforcement Learning with Human Feedback (RLHF): Bandit algorithms enhance RLHF by optimizing reward model updates, prioritizing human annotations, and balancing exploration-exploitation trade-offs. This leads to more efficient learning, reducing overfitting to specific feedback patterns.

Leveraging LLMs to Enhance Bandit Algorithms

While Bandit algorithms have proven useful for optimizing LLM performance, the reverse interaction—using LLMs to improve Bandit-based decision-making—remains an emerging research frontier. This section explores key areas where LLMs can enhance Bandit algorithms.

Contextual Understanding

Bandit algorithms, particularly contextual bandits, rely on feature representations to make informed decisions. Traditionally, these features are manually engineered, often constrained by domain knowledge and predefined structures. However, LLMs offer a powerful alternative by automatically extracting high-dimensional semantic-rich representations from unstructured textual data.

Feature Extraction from Textual Inputs: LLMs can process raw text (e.g., user queries, product descriptions, or conversational history) and generate embeddings that encode deep contextual relationships (Baheri and Alm 2023). These embeddings can serve as input features for contextual bandits, improving their ability to distinguish between different contexts.

Disambiguation and intention recognition (Kelley et al. 2012): In many applications, reward signals depend on understanding nuanced user intent. LLMs can classify user intentions, sentiment, and preferences from interactions, providing a more informed contextual representation for bandit decision-making.

Example of application, Personalized Recommendation Systems: In online platforms, contextual bandits are used to recommend articles, advertisements, or products. Instead of relying on static user profiles, LLMs can dynamically extract real-time user preferences from chat logs, search history, or reviews, enhancing bandit-based recommendations. By enriching contextual bandits with deeper, more adaptive feature representations, LLMs enable more accurate and flexible decision-making in dynamic environments.

Policy Adaptation

Traditional Bandit algorithms adjust their policies based on numerical reward feedback, which may not always provide high-level strategic insights about changing environments. LLMs, with their ability to analyze trends, summarize past interactions, and predict future conditions, can enhance Bandit policy adaptation in several ways:

Generating Adaptive Exploration Strategies (Khoramnejad and Hossain 2025; Zhang et al. 2021) : Instead of using fixed exploration heuristics (e.g., epsilon-greedy or UCB), LLMs can analyze historical data to dynamically suggest exploration rates based on environmental changes.

Policy Updates in Non-Stationary Environments: Many real-world applications involve evolving reward distributions (e.g., user interests shift over time) (de Curtò i Díaz et al. 2023). LLMs can predict reward drift by analyzing sequential user behavior and adjust exploration-exploitation trade-offs accordingly.

Example Application: Dynamic Content Moderation: In social media platforms, moderation policies must continuously evolve based on emerging trends and user reports. LLMs can monitor discourse changes and recommend real-time policy updates to bandit-based content moderation systems. By integrating LLM-driven reasoning, bandit models can proactively adapt policies instead of merely reacting to numerical rewards, leading to more robust and resilient decision-making strategies.

Exploration-Exploitation Insights

Balancing exploration (trying new actions) and exploitation (favoring known high-reward actions) is a core challenge in Bandit algorithms. Traditional approaches rely on statistical methods to manage this trade-off, but they often lack long-term foresight. LLMs can enhance this balance by incorporating historical insights, domain knowledge, and predictive modeling:

Aspect	Challenges	Bandit Solutions	Impact
Training	High cost, inefficiency	Adaptive sampling, tuning	Faster, cheaper training
Prompt Optimization	Manual tuning is slow	Contextual, dueling bandits	Better responses
Personalization	User preference variation	Contextual bandits, fine-tuning	Tailored, user-centric responses
RLHF	Overfitting, annotation cost	Optimized feedback selection	Fair, efficient learning
Future	Scalability, robustness	Hierarchical, hybrid models	Smarter, adaptable LLMs

Table 1: Bandit Algorithms for LLM Improvement

Forecasting Long-Term Reward Trajectories: LLMs can analyze past interactions to predict potential reward distributions for different actions, allowing Bandit models to make more informed exploration choices (Chen, Golrezaei, and Bouneffouf 2023; Yacobi and Bouneffouf 2023).

Semantic Similarity for Knowledge Transfer (Xu et al. 2024): In cases where limited feedback is available, LLMs can assess the semantic similarity between different arms, enabling bandit algorithms to transfer knowledge across related decisions .

Example Application – In medical treatment recommendation systems (Jin et al. 2024), exploration must be carefully balanced with patient safety. LLMs can analyze past case studies, clinical trial results, and medical literature to predict treatment effectiveness, guiding bandit-based treatment selection. By incorporating LLM-driven predictive modeling, Bandit algorithms can improve exploration efficiency and reduce suboptimal selections, leading to faster and more reliable convergence to optimal actions.

Natural Language Feedback

One of the major limitations of traditional Bandit learning is its reliance on explicit numerical reward signals, which can be sparse or difficult to obtain. In many real-world applications, user feedback is provided in natural language, requiring interpretation before it can be used to update Bandit policies. LLMs can bridge this gap by converting qualitative feedback into structured rewards:

Sentiment Analysis for Implicit Reward Extraction: Instead of relying on explicit ratings (e.g., 1–5 stars), LLMs can analyze customer reviews, chat logs, or social media comments to extract implicit satisfaction signals for Bandit learning (Parthasarathy et al. 2025).

Summarizing Feedback for Reward Calibration: LLMs can condense large volumes of user responses into structured insights, enabling Bandit algorithms to adjust their policies without requiring exhaustive manual labeling (Hoveyda et al. 2024).

Summary

This summary outlines some key takeaways from the previous section that explores the application of Large Language Models (LLMs) in various aspects of bandit problems. Here's a breakdown of each point:

Diversity of Applications: LLMs are being applied in a wide range of bandit-related tasks, enhancing both basic decision-making processes (like regret minimization) and more complex objectives, such as fairness in multilingual settings. This diversity shows the potential of LLMs to improve different facets of the bandit problem.

LLM Integration: Different studies utilize LLMs in varying ways. Some focus on reward modeling, where LLMs help predict the rewards of certain actions. Others employ LLMs for adaptive learning, adjusting strategies over time based on new data. There are also cases where LLMs refine bandit strategies, optimizing decision-making in real-time.

Impact on Bandit Algorithms: LLMs appear to improve decision-making by making it more adaptable and reducing biases in dynamic environments. This highlights the ability of LLMs to enhance the performance of bandit algorithms, especially in uncertain or changing contexts.

Limited Exploration on Efficiency: While LLMs contribute to better decision-making, the studies seem to overlook their computational cost, scalability, and robustness under adversarial conditions. These are important factors when deploying LLM-enhanced bandit systems in real-world applications.

Applications and Use Cases

The combination of Bandit algorithms and LLMs has the potential to revolutionize various domains by enabling more adaptive, intelligent, and efficient decision-making systems. Below, we explore key applications where these technologies can be integrated.

Personalization and Recommendation Systems

Modern recommendation engines aim to provide users with personalized content, whether in e-commerce, media streaming, or online advertising. Traditional recommendation systems rely on collaborative filtering, reinforcement learning, or rule-based heuristics. The integration of Bandit algorithms and LLMs enhances these systems in several ways:

Dynamic Adaptation: Bandit algorithms optimize content selection by continuously learning from user feedback, ensuring that recommendations remain relevant even as user preferences evolve. Enhanced User Understanding: LLMs improve user profiling by analyzing explicit feedback (e.g., product reviews, search queries) and implicit signals (e.g., browsing history, click patterns). Cold-Start Problem Mitigation: In cases where new users or items enter the system, LLMs can infer preferences from textual metadata, while Bandit algorithms optimize exploration strategies to accelerate personalization (Ye, Yoganarasimhan, and Zheng 2024).

Aspect	Challenges	Bandit Solutions	Impact
Applications	Complex decision-making	LLMs enhance fairness, optimization	Improved task performance
Integration	Varying LLM usage	Reward modeling, adaptive learning	Better strategy refinement
Adaptability	Bias, uncertainty	LLMs improve response to changes	More robust decision-making
Efficiency	High computation cost	Few studies on scalability	Limited deployment feasibility
Benchmarking	Lack of standardization	Need for common evaluation metrics	Hard to assess generalizability

Table 2: LLM Applications and Challenges in Bandit Problems

Dialogue Systems

Conversational AI plays a crucial role in virtual assistants, customer support, and interactive chatbots (Hoveyda et al. 2024). While LLMs enable human-like text generation, they often struggle with long-term optimization and engagement strategies. By incorporating Bandit algorithms, dialogue systems can optimize Response Selection, where Bandits help balance response diversity and informativeness, ensuring that the AI remains engaging without becoming repetitive. Personalize Interactions: LLMs extract user sentiment and intent, while Bandits optimize response styles and topic transitions for a more adaptive conversation. Improve Customer Satisfaction: In customer support applications, Bandit algorithms prioritize high-value responses, continuously refining strategies based on real-time feedback.

Healthcare and Education

Adaptive learning and personalized healthcare are two fields that benefit greatly from AI-driven optimization. The combination of Bandit algorithms and LLMs enables:

Personalized Treatment Recommendations: Bandits refine treatment plans based on patient responses, while LLMs analyze medical records and clinical notes to enhance decision-making. Adaptive Learning Pathways: In education, Bandits determine optimal learning resources, while LLMs provide contextual explanations and engagement-driven adaptations. Enhanced Diagnostics: LLMs extract insights from medical literature, and Bandit algorithms optimize test selection to minimize unnecessary procedures.

Challenges and Open Problems

Despite the promising potential of combining Bandit algorithms with LLMs, several challenges remain to be addressed. This section outlines key open problems in this research domain. Adaptive Reasoning Pathways: Employing bandit strategies to select among various reasoning pathways or prompts during inference could enhance LLMs' performance on complex tasks. By dynamically choosing the most promising reasoning strategies, models might achieve better accuracy and coherence in their outputs (DeepSeek-AI et al. 2025).

Trust and Interpretability: Ensuring transparency in decision-making processes is crucial for real-world deployment. Bandit-driven decisions influenced by LLMs could benefit from some theoretical guarantees such as regret upper bound, and should be interpretable and explainable, particularly in critical applications like healthcare and autonomous systems.

Multi-Agent Scenarios: Integrating Bandit algorithms and LLMs in multi-agent environments introduces complexity in coordination, communication, and decision-making. New approaches are required to ensure optimal cooperation and competition among multiple agents using these technologies.

Addressing these challenges is essential for advancing the integration of Bandit algorithms and LLMs into practical, scalable, and robust AI systems.

Future Directions

The intersection of Bandit algorithms and LLMs presents numerous opportunities for future research. This section outlines key directions for advancing this interdisciplinary field.

Opportunities for Interdisciplinary Research: The fusion of Bandit algorithms and LLMs benefits from collaborations across multiple disciplines, including reinforcement learning, natural language processing, cognitive science, and human-computer interaction. By integrating insights from these fields, researchers can develop more robust and adaptive decision-making frameworks.

Emerging Trends: Recent advancements in multi-modal LLMs and hybrid algorithms open new avenues for research. Multi-modal models, which process diverse data types (e.g., text, images, audio), can enrich Bandit learning by incorporating varied contextual signals. Similarly, hybrid algorithms that combine deep learning and Bandit strategies can enhance exploration-exploitation balance in complex environments.

Conclusion

The integration of Bandit algorithms and Large Language Models (LLMs) represents a promising frontier in AI-driven decision-making. This survey has explored how Bandit algorithms can enhance LLM optimization and how LLMs, in turn, can improve Bandit-based strategies through contextual understanding, policy adaptation, and natural language feedback. The synergy between these approaches enables more intelligent, adaptive, and human-aligned AI systems across various applications, including recommendation systems, dialogue agents, healthcare, and autonomous systems.

References

Baheri, A.; and Alm, C. O. 2023. LLMs-augmented Contextual Bandit. *arXiv preprint arXiv:2311.02268*.

Boufelja-Yacoubi, S.; Bouneffouf, D.; and Zhuk, S. 2024. Machine Learning Using Robust Stochastic Multi-Armed Bandits with Historical Data. US Patent App. 18/128,454.

Bouneffouf, D. 2023. Multi-Armed Bandit Problem and Application.

Bouneffouf, D. 2025a. Contextual bandit with trending reward function. US Patent App. 18/343,104.

Bouneffouf, D. 2025b. Contextual thompson sampling with corrupted and missing context. US Patent App. 18/373,033.

Bouneffouf, D. 2025c. Evolutionary contextual bandits. US Patent App. 18/454,108.

Bouneffouf, D. 2025d. Multi-armed bandit with optimum exploration-exploitation distribution parameter. US Patent App. 18/432,577.

Bouneffouf, D. 2025e. Online system and method for solving context-attentive combinatorial bandit with observations. US Patent App. 18/454,106.

Bouneffouf, D.; and Féraud, R. 2024. A Tutorial on Multi-Armed Bandit Applications for Large Language Models. In *KDD*.

Bouneffouf, D.; Feraud, R.; and Lin, B. 2025. Multi-Armed Bandit with Sparse and Noisy Feedback. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5. IEEE.

Bouneffouf, D.; Féraud, R.; Upadhyay, S.; Khazaeni, Y.; and Rish, I. 2021. Double-Linear Thompson Sampling for Context-Attentive Bandits. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 3450–3454. IEEE.

Bouneffouf, D.; Laroche, R.; Urvoy, T.; Féraud, R.; and Allesiardo, R. 2014. Contextual bandit for active learning: Active thompson sampling. In *ICONIP 2014*.

Bouneffouf, D.; Rish, I.; and Aggarwal, C. 2020. Survey on applications of multi-armed and contextual bandits. In *CEC*. IEEE.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.

Cai, W.; Grossman, J.; Lin, Z. J.; Sheng, H.; Wei, J. T.-Z.; Williams, J. J.; and Goel, S. 2021. Bandit algorithms to personalize educational chatbots. *Machine Learning*, 110(9): 2389–2418.

Chang, J. D.; Brantley, K.; Ramamurthy, R.; Misra, D.; and Sun, W. 2023. Learning to generate better than your llm. *arXiv preprint arXiv:2306.11816*.

Chen, Q.; Golrezaei, N.; and Bouneffouf, D. 2023. Non-stationary bandits with auto-regressive temporal dependency. *Advances in Neural Information Processing Systems*, 36: 7895–7929.

Chen, Q.; Golrezaei, N.; and Bouneffouf, D. 2025. Online system with bandit feature and auto-regressive temporal structure. US Patent App. 18/627,702.

Chen, Q.; Liang, J. C. N.; Golrezaei, N.; and Bouneffouf, D. 2024. Interpolating Item and User Fairness in Multi-Sided Recommendations. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Chen, Z.; Chen, P.-Y.; and Buet-Golfouse, F. 2024. Online personalizing white-box llms generation with neural bandits. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 711–718.

Dai, X.; Xie, Y.; Liu, M.; Wang, X.; Li, Z.; Wang, H.; and Lui, J. 2025. Multi-Agent Conversational Online Learning for Adaptive LLM Response Identification. *arXiv preprint arXiv:2501.01849*.

de Curtò i Díaz, J.; de Zarzà i Cubero, I.; Roig, G.; Cano, J. C.; Manzoni, P.; and Calafate, C. T. 2023. LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Delande, D.; Stolf, P.; Feraud, R.; Pierson, J.-M.; and Bottaro, A. 2021. Horizontal Scaling in Cloud Using Contextual Bandits. In *Euro-Par 2021: Parallel Processing*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Gao, S.; Wang, C.; Gao, C.; Jiao, X.; Chong, C. Y.; Gao, S.; and Lyu, M. 2025. The Prompt Alchemist: Automated LLM-Tailored Prompt Optimization for Test Case Generation. *arXiv preprint arXiv:2501.01329*.

Grams, T.; Betz, P.; and Bartelt, C. 2025. Disentangling Exploration of Large Language Models by Optimal Exploitation. *arXiv preprint arXiv:2501.08925*.

Hoveyda, M.; de Vries, A. P.; de Rijke, M.; Oosterhuis, H.; and Hasibi, F. 2024. AQA: Adaptive Question Answering in a Society of LLMs via Contextual Multi-Armed Bandit. arXiv:2409.13447.

Idé, T.; Murugesan, K.; Bouneffouf, D.; and Abe, N. 2025. Sequential uncertainty quantification with contextual tensors for social targeting. *Knowl. Inf. Syst.*, 67(3): 2881–2910.

Jin, Q.; Wang, Z.; Floudas, C. S.; Chen, F.; Gong, C.; Bracken-Clarke, D.; Xue, E.; Yang, Y.; Sun, J.; and Lu, Z. 2024. Matching patients to clinical trials with large language models. *Nature communications*, 15(1): 9074.

Kaufmann, T.; Weng, P.; Bengs, V.; and Hüllermeier, E. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Kelley, R.; Tavakkoli, A.; King, C.; Ambardekar, A.; Nicolescu, M.; and Nicolescu, M. 2012. Context-based bayesian intent recognition. *IEEE Transactions on Autonomous Mental Development*, 4(3): 215–225.

Khoramnejad, F.; and Hossain, E. 2025. Generative AI for the optimization of next-generation wireless networks: Basics, state-of-the-art, and open challenges. *IEEE Communications Surveys & Tutorials*.

Kishimoto, A.; Bouneffouf, D.; Marinescu, R.; Ram, P.; Rawat, A.; Wistuba, M.; Palmes, P. P.; and Botea, A. 2022. Bandit Limited Discrepancy Search and Application to Machine Learning Pipeline Optimization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 10228–10237. AAAI Press.

Kishimoto, A.; Hama, T.; Hsu, H. H.; and Bouneffouf, D. 2025. Optimization of multiple molecules. US Patent 12,334,195.

Lau, A.; Choi, Y.; Balazadeh, V.; Chidambaram, K.; Syrgkanis, V.; and Krishnan, R. G. 2024. Personalized Adaptation via In-Context Preference Learning. *arXiv preprint arXiv:2410.14001*.

Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; and Talwalkar, A. 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185): 1–52.

Li, Y. 2025. LLM Bandit: Cost-Efficient LLM Generation via Preference-Conditioned Dynamic Routing. *arXiv preprint arXiv:2502.02743*.

Lin, B.; Cecchi, G. A.; Bouneffouf, D.; Reinen, J. M.; and Rish, I. 2020. A Story of Two Streams: Reinforcement Learning Models from Human Behavior and Neuropsychiatry. In Seghrouchni, A. E. F.; Sukthankar, G.; An, B.; and Yorke-Smith, N., eds., *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, 744–752. International Foundation for Autonomous Agents and Multiagent Systems.

Lisicki, M.; Nica, M.; and Taylor, G. W. 2023. Bandit-Driven Batch Selection for Robust Learning under Label Noise. *arXiv preprint arXiv:2311.00096*.

Liu, R.; Wu, T.; and Mozafari, B. 2020. Adam with bandit sampling for deep learning. *Advances in Neural Information Processing Systems*, 33: 5393–5404.

Mulakala, B.; Saini, M. L.; Singh, A.; Bhukya, V.; and Mukhopadhyay, A. 2024. Adaptive Multi-Fidelity Hyperparameter Optimization in Large Language Models. In *CSITSS*.

Noothigattu, R.; Bouneffouf, D.; Mattei, N.; Chandra, R.; Madan, P.; Varshney, K. R.; Campbell, M.; Singh, M.; and Rossi, F. 2019. Teaching AI agents ethical values using reinforcement learning and policy orchestration. *IBM J. Res. Dev.*, 63(4/5): 2:1–2:9.

Parthasarathy, A.; Subramanian, C.; Senrayan, G.; Adappanavar, S.; Taneja, A.; Ravindran, B.; and Tambe, M. 2025. Multilinguality in LLM-Designed Reward Functions for Restless Bandits: Effects on Task Performance and Fairness. *arXiv preprint arXiv:2501.13120*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*.

Riemer, M.; Klinger, T.; Bouneffouf, D.; and Franceschini, M. 2019. Scalable Recollections for Continual Lifelong Learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 1352–1359. AAAI Press.

Settles, B. 2009. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison.

Shi, C.; Yang, K.; Chen, Z.; Li, J.; Yang, J.; and Shen, C. 2024. Efficient prompt optimization through the lens of best arm identification. In *NEURIPS*.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *NIPS*.

Song, Y.; Yin, D.; Yue, X.; Huang, J.; Li, S.; and Lin, B. Y. 2024. Trial and error: Exploration-based trajectory optimization for llm agents. *arXiv preprint arXiv:2403.02502*.

Tchrakian, T. T.; Zayats, M.; Zhuk, S.; and Bouneffouf, D. 2025a. Adaptive spinal cord stimulation policy generation. US Patent App. 18/233,699.

Tchrakian, T. T.; Zhuk, S.; Bouneffouf, D.; Zayats, M.; and Rogers, J. L. 2025b. Patient treatment recommendations. US Patent App. 18/215,960.

Upadhyay, S.; Agarwal, M.; Bounneffouf, D.; and Khazaeni, Y. 2019. A Bandit Approach to Posterior Dialog Orchestration Under a Budget.

Urvoy, T.; Clerot, F.; Féraud, R.; and Naamane, S. 2013. Generic exploration and K-armed voting bandits. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*.

Wahed, M.; Gruhl, D.; and Lourentzou, I. 2023. MArBLE: Hierarchical Multi-Armed Bandits for Human-in-the-Loop Set Expansion. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4857–4863.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NEURIPS*.

Wu, C.-K.; Tam, Z. R.; Lin, C.-Y.; Chen, Y.-N.; and Lee, H.-y. 2024. StreamBench: Towards Benchmarking Continuous Improvement of Language Agents. *arXiv preprint arXiv:2406.08747*.

Xia, Y.; Kong, F.; Yu, T.; Guo, L.; Rossi, R. A.; Kim, S.; and Li, S. 2024a. Convergence-Aware Online Model Selection with Time-Increasing Bandits. In *The Web Conference*.

Xia, Y.; Kong, F.; Yu, T.; Guo, L.; Rossi, R. A.; Kim, S.; and Li, S. 2024b. Which LLM to Play? Convergence-Aware Online Model Selection with Time-Increasing Bandits. In *Proceedings of the ACM on Web Conference 2024*, 4059–4070.

Xu, S.; Wu, Z.; Zhao, H.; Shu, P.; Liu, Z.; Liao, W.; Li, S.; Sikora, A.; Liu, T.; and Li, X. 2024. Reasoning before comparison: LLM-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv preprint arXiv:2402.11398*.

Yacobi, S. B.; and Bouneffouf, D. 2023. Robust Stochastic Multi-Armed Bandits with Historical Data. In *International World Wide Web Conference*.

Yang, R.; Pan, X.; Luo, F.; Qiu, S.; Zhong, H.; Yu, D.; and Chen, J. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Ye, Z.; Yoganarasimhan, H.; and Zheng, Y. 2024. LOLA: LLM-Assisted Online Learning Algorithm for Content Experiments. arXiv:2406.02611.

Zafar, S.; Féraud, R.; Blavette, A.; Camilleri, G.; and Ahmed, H. B. 2023. Multi-Armed Bandits Learning for Optimal Decentralized Control of Electric Vehicle Charging. In *2023 IEEE Belgrade PowerTech*.

Zeng, Y.; Chen, X.; and Jin, R. 2023. Ensemble Active Learning by Contextual Bandits for AI Incubation in Manufacturing.

Zhang, S. X.; Chan, W. S.; Tang, K. S.; and Zheng, S. Y. 2021. Adaptive strategy in differential evolution via explicit exploitation and exploration controls. *Applied Soft Computing*, 107: 107494.

Zhang, Z.-Y.; Han, S.; Yao, H.; Niu, G.; and Sugiyama, M. 2024. Generating Chain-of-Thoughts with a Pairwise-Comparison Approach to Searching for the Most Promising Intermediate Thought.