# A Survey on Use-Cases of Bandit Problems in Natural Language Processing

**Arnav Gurudatt**
University of California, Berkeley
arnavgurudatt@berkeley.edu

## Abstract

Multi-armed bandits have found use across various domains in sequential decision-making tasks, from assigning treatment arms in clinical trials to optimizing click-through rate at industry scale. This survey aims to evaluate the use-cases and effectiveness of bandit algorithms in natural language processing (NLP) settings, providing a unified compendium of existing frameworks and areas for further study. I motivate the use of bandits via structured prediction, where models receive partial, delayed, or implicit supervision from bandit feedback rather than full ground-truth labels. I examine use cases across a range of subfields including machine translation, dialog generation, question answering, and large language models. Through a synthesis of over 30 papers, I trace the historical development of bandit methods in NLP, outline key algorithmic frameworks, and identify emerging trends in both research and industry.

## 1 Introduction

### The Multi-Armed Bandit Problem

The multi-armed bandit problem in its simplest form, often called the stochastic bandit problem, models the fundamental tradeoff between exploration (gathering information about uncertain options) and exploitation (leveraging known information to maximize reward). A classic analogy is a gambler faced with various slot machines (or "arms") in a casino, each returning a random reward from an unknown probability distribution. The gambler's goal is to identify the slot machine which gives the most lucrative reward while simultaneously being rewarded as they pull the levers of various slot machines.

Formally, the problem consists of an agent faced with $K$ unknown probability distributions $\{D_1, D_2, \ldots, D_K\}$, each with an expected reward $\mu_k = \mathbb{E}[r_k]$ and variance $\sigma_k^2$. At each round $t = 1, 2, \ldots, T$, the agent selects an arm $j(t) \in \{1, \ldots, K\}$ and receives a reward $r(t) \sim D_{j(t)}$. The agent aims to maximize the total reward over $T$ rounds, ideally by identifying the optimal arm $k^* = \arg\max_k \mu_k$.

A key performance metric is the cumulative (expected) regret $R_T$, which quantifies the difference between the reward the agent would have earned by always pulling the optimal arm, and the reward actually received:

$$R_T = T\mu^* - \sum_{t=1}^{T} \mu_{j(t)}$$

where $\mu^* = \max_{k=1,\ldots,K} \mu_k$ is the expected reward of the best arm.

An equivalent expression using the expected number of times each arm $k$ is selected is:

$$R_T = T\mu^* - \sum_{k=1}^{K} \mu_k \cdot \mathbb{E}[T_k(T)]$$

where $T_k(T)$ denotes the (random) number of times arm $k$ is pulled in the first $T$ rounds.

Bandit algorithms aim to minimize this regret $R_T$, ideally achieving sublinear growth $R_T = o(T)$, ensuring that the average regret per round vanishes as $T \to \infty$ (Kuleshov and Precup, 2000).

In addition to stochastic bandits, variants of the original bandit problem have emerged in different settings, extending the base assumptions to model different agentic environments. The most prominent of these variants is the *contextual bandit* problem, which extend stochastic bandits by allowing the agent to observe side information (context) $x_t \in \mathcal{X}$ before selecting an arm. The expected reward depends on both the arm and the context:

$$\mu_k(x_t) = \mathbb{E}[r_k(t) \mid x_t]$$

The agent's goal is to learn a policy $\pi : \mathcal{X} \to \{1, \ldots, K\}$ that maps contexts to actions to max-

imize cumulative reward. Algorithms like LINUCB and contextual Thompson Sampling leverage structure in the context space (e.g., linear or neural models) to guide arm selection (Lu et al., 2010).

**Common Algorithms for Regret Minimization**

Several standard algorithms have been developed to tackle the exploration-exploitation tradeoff and minimize regret in stochastic bandit settings. Some of the most common approaches include, but are not limited to:

- **$\epsilon$-greedy**: At each time step, the agent chooses a random arm with probability $\epsilon$ (exploration), and with probability $1 - \epsilon$, it chooses the arm with the highest empirical mean reward (exploitation).
- **Upper Confidence Bound (UCB):** Selects the arm with the highest upper confidence bound on the estimated reward:

$$j(t) = \arg\max_k \left( \hat{\mu}_k + \sqrt{\frac{2 \log t}{T_k(t)}} \right)$$

  where $\hat{\mu}_k$ is the empirical mean reward of arm $k$ and $T_k(t)$ is the number of times arm $k$ has been pulled up to time $t$ (Kuleshov and Precup, 2000).
- **Thompson Sampling:** A Bayesian algorithm that samples a reward estimate from the posterior distribution of each arm and selects the arm with the highest sampled value. For Bernoulli rewards with Beta priors:

$$\theta_k \sim \text{Beta}(\alpha_k, \beta_k), \quad j(t) = \arg\max_k \theta_k$$

  Thompson Sampling achieves logarithmic expected regret in the stochastic setting. For the two-armed bandit case, the regret after $T$ rounds is:

$$R_T = O\left( \frac{\log T}{\Delta} + \frac{1}{\Delta^3} \right)$$

  where $\Delta = \mu^* - \mu_{\text{next-best}}$ is the gap between the expected reward of the best arm and the second-best arm.

  More generally, for $K$ arms, the regret satisfies:

$$R_T = O\left( \left( \sum_{i=2}^{K} \frac{1}{\Delta_i^2} \right) \log T \right)$$

  where $\Delta_i = \mu^* - \mu_i$ denotes the suboptimality gap for arm $i$ (Agrawal and Goyal, 2012).

## 2 Historical Development

### 2.1 Interactive Learning with Bandit Feedback (2016)

Sokolov et al. (2016) produced a seminal paper on the applications of bandits in training models for natural language tasks like machine translation; rather than train a language model on a supervised dataset containing ground-truth labels (e.g. correct gold-standard translations for a machine translation task), the model only gets feedback in the form of a loss value for its own prediction, such as a BLEU score or human rating. They describe this setting as a "bandit structured prediction," and it simulates realistic low-supervision environments, such as user feedback in production systems.

They validate this structured prediction framework across three natural language tasks – machine translation, sequence labeling, and text classification – each of which traditionally relies on full supervision. In machine translation, their bandit-trained model achieved a BLEU score of 0.2763, nearly matching the supervised baseline (0.2841), while in chunking, the bandit model reached 0.923 F1 versus a 0.935 supervised score.

Sokolov et al.'s (2016) paper was important within NLP literature for three reasons. First, it marked a shift in the supervision paradigm by laying the foundation for learning from partial, noisy, or implicit feedback (e.g. user clicks, time-spent metrics), thereby departing from the assumption that full ground-truth supervision is always available. Second, it modeled realistic human-in-the-loop feedback scenarios, such as BLEU-based or rating-based user feedback, which closely mirror the deployment conditions of NLP systems in production. Third, and most importantly, it helped bridge the methodological gap between NLP and bandits literature by applying bandit algorithms to common NLP tasks. By providing benchmark results across machine translation, sequence labeling, and classification, the paper catalyzed follow-up research in areas like summarization and dialogue modeling, helping shape research at the intersection of reinforcement learning and NLP.

### 2.2 Bandit Approaches for Neural Machine Translation (2017-18)

With the increased use of neural models for NLP tasks, especially with the advent of the transformer architecture (Vaswani et al., 2017), bandits research in NLP began exploring where reinforcement learn-

ing could improve the performance of sequence-to-sequence models, particularly for machine translation.

Kreutzer et al. (2017) were among the first to extend bandit structured prediction from Sokolov et al. (2016) to neural architectures by applying it to attention-based encoder-decoder models. Their work introduces variance-reduced stochastic learning algorithms for neural machine translation (NMT) using simulated partial feedback, such as sentence-level BLEU scores. In domain adaptation experiments, their model achieved BLEU improvements of up to 5.89 points over the out-of-domain baseline, showing that bandit learning with partial supervision could not only effectively fine-tune neural NMT systems, but also could scale to deep models and complex output spaces.

Earlier literature on NMT utilized *simulated* user feedback in bandit settings to overcome the limitations of requiring full supervision. Nguyen et al. (2017) modeled NMT as a reinforcement learning problem using an advantage actor-critic (A2C) algorithm, showing that models trained on noisy, high-variance reward signals could still achieve stable learning. Building on this, Kreutzer et al. (2018b) compared different forms of simulated human feedback, finding that standardized cardinal ratings yielded higher annotator agreement and enabled BLEU improvements of over 1 point using just 800 labeled examples.

This trajectory culminated in Kreutzer et al. (2018a), which marked a transition from simulation to real-world deployment. Using data collected from users on the eBay platform, they analyzed both explicit user ratings and implicit task-based feedback (e.g., clicks in cross-lingual search). While the 5-star ratings were found to be too noisy for training, the implicit feedback was successfully used to improve downstream task relevance and translation quality. This work provided an early example of how neural bandit methods could be integrated into production systems, showing that human reinforcement signals in real-world settings could improve NLP tasks like NMT at scale.

As MT systems moved toward deployment at scale, bandit algorithms proved useful not only for adapting models, but also for making decisions about which models or data to use. Naradowsky et al. (2020) proposed treating model selection itself as a bandit problem, where each "arm" corresponds to a pre-trained MT system with fixed parameters, with an agent that learns a selection policy from user feedback and dynamically chooses the best system for a given translation request. Similarly, data selection for model training can be framed as a bandit problem; since MT systems are trained on data from different domains, languages, levels of quality, and other "facets," Kreutzer et al. (2021) propose using dynamic sampling from diverse data sources to optimize model improvement, achieving up to 1.7 BLEU gain across various domain adaptation settings (e.g., selecting between natural vs. translated text, training multilingual models).

### 2.3 Industry Adoption (2019-2023)

The shift from simulation-focused study of bandit feedback in NLP systems to viable, real-world use cases prompted broader industry adoption of bandit algorithms, not just for machine translation like in previous literature, but also for better recommender systems and natural language understanding (NLU) in Automated Speech Recognition (ASR).

In particular, bandit algorithms found large success in recommender systems, where they enable personalization through efficient exploration under limited user feedback. ByteDance introduced *conversational contextual bandits*, which enhance exploration by allowing the system to query user preferences on key-terms that influence many items, accelerating learning and reducing cumulative regret across large-scale, real-world recommendation datasets like Yelp (restaurants) and Toutiao (news) (Zhang et al., 2020). Meta extended this idea with *Epistemic Neural Recommendation (ENR)*, a neural contextual bandit framework that combines Thompson Sampling with epistemic uncertainty estimation to deliver personalized recommendations from sparse interactions (Zhu and Van Roy, 2023). ENR achieved up to 9% higher click-through rates and required 29% fewer interactions compared to prior neural bandit methods. Together, these approaches show how bandits can scale exploration and personalization in industrial recommender systems with minimal supervision.

Amazon scientists also explored use cases of multi-armed bandits for personalized NLU models on their voice-controlled devices like Amazon Echo, particularly in music playback requests (Moerchen et al., 2020). In this setting, upstream NLU models produce multiple candidate interpretations of a voice query, and a contextual bandit model is used to re-rank these candidates using implicit feedback (specifically, whether the selected inter-

pretation led to music being played for at least $K$ seconds). Using Thompson Sampling, the model learns to personalize and correct interpretation errors by incorporating features such as customer-artist affinity and entity popularity. During A/B testing, this approach increased playback rate by up to 0.41% in the UK and 0.37% in the US, making it one of the most impactful improvements to Amazon Music's voice experience that year.

### 2.4 Bandits and Large Language Models (2023-present)

With Large Language Models (LLMs) increasingly at the forefront of both NLP research and industry scale alike, recent research has sought to examine how bandit algorithms can improve LLM performance on specific tasks across various stages of the modeling pipeline, including prompt optimization, data selection, and retrieval.

In prompt-based learning, both Kiyohara et al. (2025) and Google DeepMind's BanditBench paper (Nie et al., 2025) frame LLM behavior as a contextual bandit problem, optimizing output quality using only logged user feedback. These approaches avoid costly online interaction by either learning from generated responses (as in Direct Sentence Off-policy Gradient) or simulating bandit environments to benchmark in-context learning. Meanwhile, other work has adapted bandits to enhance retrieval-augmented generation (RAG) over knowledge graphs (KGs), where each retrieval method (e.g., dense retriever, SPARQL query generator, or LLM-based KG agent) is treated as an arm, and the system adaptively selects the best one for each query using real-time user feedback (Tang et al., 2024). Bandits have further been used to improve pretraining efficiency through data selection strategies that balance data quality, diversity, and influence in training corpora, where each data cluster is an arm receiving feedback from each cluster's "influence score," or impact on the model's performance (Zhang et al., 2025). Together, these papers show that bandits offer a scalable framework for guiding LLM learning when supervision is partial, delayed, or expensive to collect.

While the discussion of bandits and LLMs has thus far focused on bandits improving LLM performance, another branch of literature flips this relationship, examining how LLMs can enhance bandit algorithms in various sequential decision-making settings. Alamdari et al. (2024) and Sun et al. (2025) show that LLMs can simulate user preferences or serve as reward predictors to accelerate learning in contextual and dueling bandit settings. These methods avoid cold-start issues and reduce regret by leveraging the generalization capabilities of LLMs to warm-start bandit policies or predict outcomes across high-dimensional action spaces.

## 3 Applications by NLP Subfield

Historical development of bandit frameworks for NLP have seen substantial success in tasks such as machine translation, recommender systems that utilize natural language input, and LLM output generation. In this section, I aim to explore specific subfields of NLP that have benefited from the use of bandit algorithms as a way of augmenting the performance of NLP systems.

### 3.1 Dialog Generation

An NLP task that fairly naturally lends itself to extensions of the bandit problem is dialog generation. In particular, contextual bandits have been previously used in dialog generation literature to guide response selection based on the current conversation state while learning from partial feedback. Since contextual bandits leverage side information (i.e., context) to make decisions with only limited feedback, they are well-suited for dialog generation tasks where models must adaptively select or generate responses based on conversational history, user preferences, and sparse or implicit user feedback. This setting matches the typical interaction loop in dialog systems: observe the current context, select a response (or action), and learn from user behavior (e.g., clicks, replies, or ratings) without observing feedback on unchosen responses.

Many systems leverage contextual bandits to dynamically select responses or dialog skills based on conversational context, user signals, and long-term utility, enabling models to operate effectively without access to full supervision (Upadhyay et al., 2019). In particular, they have been used to guide online response selection in retrieval-based systems, where the model encodes the conversational context and candidate replies using neural models (e.g., LSTMs or RNNs), and uses Thompson Sampling to select the most appropriate response (Liu et al., 2018). Beyond reactive dialog, bandit-based approaches have also been used to enable proactive or self-improving systems that learn from ongoing user interactions; by incorporating memory or post-deployment feedback signals, these systems can

continually refine their policies based on implicit or explicit rewards without retraining from scratch (Perez and Silander, 2018; Hancock et al., 2019).

In addition to contextual bandits, *counterfactual* bandits have proven effective for training dialog systems from limited feedback by attributing observed outcomes to specific model components. Both in modular spoken language understanding (SLU) systems and multi-action dialog policy learning, counterfactual bandits enable learning from system-level or logged user feedback by estimating which parts of the model contributed to the outcome. Falke and Lehnen (2021) use attribution methods from multi-agent reinforcement learning to assign credit to SLU submodules like domain classification and slot filling, while Zhang et al. (2023) introduce BanditMatch, which combines pseudo-labeling and off-policy estimation to train dialog policies from real-world feedback signals. Across SLU benchmarks like SNIPS and TOP and dialog datasets like MultiWOZ, both approaches demonstrate strong performance, matching or surpassing supervised baselines under partial feedback.

### 3.2 Question Answering

Recent developments in question answering (QA) have combined user interaction patterns seen in dialog generation with information retrieval from large corpora, creating opportunities for bandit algorithms to improve answer selection from sparse feedback. Gao et al. (2022) formulate extractive QA as a contextual bandit problem, where the system proposes an answer span and learns from binary user feedback indicating correctness. To study this setting without requiring expensive real-time annotation, they introduce a simulation framework using SQuAD and Natural Questions to generate realistic user responses. Their experiments show that even simple bandit algorithms can outperform supervised learning when feedback is sparse or noisy, and that offline bandit learning achieves up to 2.4% F1 improvement in domain transfer settings.

### 3.3 Document Summarization

Dong et al. (2018) extend the contextual bandit problem to extractive document summarization, using the document itself as the context for the bandit algorithm and the selected set of sentences as the action. Rather than treating summarization as a sequential labeling or reinforcement learning task, their approach, BanditSum, models sentence selection as a one-shot decision, optimizing sentence subsets based on expected ROUGE scores using policy gradient methods. This allows the model to learn directly from summary-level feedback without needing extractive sentence-level labels or assuming strong sequential dependencies. On the CNN/Daily Mail benchmark, BanditSum achieves competitive ROUGE scores while mitigating early-sentence bias and converging faster than traditional reinforcement learning approaches.

### 3.4 Model Selection, Training, and Evaluation

Beyond augmenting domain-specific NLP tasks, bandit algorithms have increasingly been used to improve the model development pipeline itself, spanning from training data selection and hyperparameter tuning to reward optimization and evaluation. Urteaga et al. (2023) frame language model pretraining as an online optimization problem, applying Thompson Sampling to dynamically select optimal hyperparameter configurations, such as dropout and masking strategies, during training, reducing computational waste caused by more classic grid-search approaches and improving perplexity across resource-constrained settings. For tasks like text generation, where optimizing against a single reward metric (e.g., BLEU or ROUGE) can lead to overfitting, Pasunuru et al. (2020) introduce DORB, a multi-armed bandit framework that dynamically selects which evaluation metric to optimize at each training step using EXP3, yielding stronger generalization across multiple tasks and metrics. Finally, bandits can assist with benchmarking itself: Haffari et al. (2017) propose a method for selecting the best NLP system using Thompson Sampling, reducing the number of API calls required to statistically identify the top-performing model on tasks like NER, thereby minimizing cost in evaluation scenarios. Collectively, these approaches demonstrate the value of bandits not just for improving model behavior, but for making the overall modeling process more data-efficient, interpretable, and cost-effective.

Broadly, the use of bandits in what Bouneffouf and Rish (2019) call "Better Machine Learning" reflects a broader trend toward optimizing meta-level decisions in the ML pipeline. Bandit algorithms have been successfully applied to algorithm selection, hyperparameter tuning, and feature selection, all of which are key tasks that benefit from adaptive decision-making under uncertainty. For instance, Bouneffouf and Rish (2019) highlight how bandits

can dynamically choose among competing classification algorithms based on feedback, or iteratively select the most informative features to reduce dimensionality while preserving performance. Outside NLP, they also describe applications in areas such as anomaly detection in cloud security, energy usage prediction in smart grids, and optimizing financial trading strategies. These examples, while not specific to NLP, demonstrate the extensibility of bandit-based decision frameworks to a wide range of machine learning tasks, many of which could intersect with or inform future NLP research, especially in domains where data collection is expensive or real-time feedback is noisy and sparse.

## 4 Limitations

One limitation across early bandit learning papers in NLP is that the empirical gains over strong baselines are often marginal. For instance, in the multi-reward question generation task, Pasunuru et al. (2020) report that their SM-Bandit model improves BLEU-4 from 18.36 to only 18.68, with similarly modest gains in METEOR (+0.33), ROUGE-L (+0.05), and QAP (+0.41). On the WebNLG dataset, their HM-Bandit achieves BLEU of 63.38 compared to 63.00 for the baseline, and improves ROUGE-L by just 0.10 points. A similar pattern appears in Kreutzer et al. (2017), where their bandit-based NMT model achieves up to 5.89 BLEU improvement in domain adaptation, but this is measured relative to an out-of-domain baseline rather than an in-domain supervised model, which remains the stronger benchmark in practice.

Moreover, many of these evaluations rely on *simulated* bandit feedback for computed metrics like sentence-level BLEU or gGLEU, which do not fully capture the challenges of noisy or delayed real-world user responses. For example, even seminal papers like Sokolov et al. (2016), Kreutzer et al. (2017), and Kreutzer et al. (2018b) evaluate models using synthetic BLEU-based rewards, and even studies that incorporate human feedback (e.g., Kreutzer et al. (2018a)) often highlight its noisiness and inconsistency. While simulation is necessary for reproducibility, it raises questions about the ecological validity of the results, particularly when feedback assumptions (e.g., stability, reward variance) are unlikely to hold in production.

A particularly strong limitation of prior work is its dependence on online feedback, where rewards must be generated during training. Petrushkov et al. (2018) identify this constraint and propose a chunk-based feedback approach that allows learning from delayed or logged feedback. Although not a bandit algorithm per se, their method circumvents a major barrier to bandit learning in real-world deployments by allowing feedback to be collected asynchronously, thereby increasing flexibility in user-facing NLP systems.

Lastly, bandits in the NLP setting face the same challenges as multi-armed bandits do more generally in production environments. These include difficulties in defining good reward proxies, high implementation complexity, safety constraints around exploration, cold-start issues, and the challenge of ensuring consistent long-term evaluation (Abensur et al., 2019; van den Akker et al., 2023). As van den Akker et al. (2023) note, even simple decisions like offline policy evaluation versus online deployment can significantly affect performance and feasibility. As a result, while bandit methods offer a compelling theoretical framework, practical limitations remain a major barrier to their widespread adoption in high-stakes, production-grade NLP systems.

## References


David Abensur, Ivan Balashov, Shaked Bar, Ronny Lempel, Nurit Moscovici, Ilan Orlov, Danny Rosenstein, and Ido Tamir. 2019. Productization challenges of contextual multi-armed bandits. *Preprint*, arXiv:1907.04884.

Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. *Proceedings of the 25th Annual Conference on Learning Theory, PMLR*.

Parand A. Alamdari, Yanshuai Cao, and Kevin H. Wilson. 2024. Jump starting bandits with LLM-generated prior knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19821–19833, Miami, Florida, USA. Association for Computational Linguistics.

Djallel Bouneffouf and Irina Rish. 2019. A survey on practical applications of multi-armed and contextual bandits. *Preprint*, arXiv:1904.10040.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Tobias Falke and Patrick Lehnen. 2021. Feedback attribution for counterfactual bandit learning in multi-

domain spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1198, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ge Gao, Eunsol Choi, and Yoav Artzi. 2022. Simulating bandit learning from user feedback for extractive question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5167–5179, Dublin, Ireland. Association for Computational Linguistics.

Gholamreza Haffari, Tuan Dung Tran, and Mark Carman. 2017. Efficient benchmarking of NLP APIs using multi-armed bandits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 408–416, Valencia, Spain. Association for Computational Linguistics.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Haruka Kiyohara, Daniel Yiming Cao, Yuta Saito, and Thorsten Joachims. 2025. Prompt optimization with logged bandit data.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018a. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Julia Kreutzer, Artem Sokolov, and Stefan Riezler. 2017. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1503–1513, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018b. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics.

Julia Kreutzer, David Vilar, and Artem Sokolov. 2021. Bandits don't follow rules: Balancing multi-facet machine translation with multi-armed bandits. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3190–3204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Volodymyr Kuleshov and Doina Precup. 2000. Algorithms for the multi-armed bandit problem. *Journal of Machine Learning Research 1*.

Bing Liu, Tong Yu, Ian Lane, and Ole J. Mengshoel. 2018. Customized nonlinear bandits for online response selection in neural conversation models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Tyler Lu, Dávid Pál, and Martin Pál. 2010. Contextual multi-armed bandits. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR, Google Research*.

Fabian Moerchen, Patrick Ernst, and Giovanni Zappella. 2020. Personalizing natural language understanding using multi-armed bandits and implicit feedback. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2661–2668, New York, NY, USA. Association for Computing Machinery.

Jason Naradowsky, Xuan Zhang, and Kevin Duh. 2020. Machine translation system selection from bandit feedback. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 50–63, Virtual. Association for Machine Translation in the Americas.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. 2017. Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark. Association for Computational Linguistics.

Allen Nie, Yi Su, Bo Chang, Jonathan Lee, Ed H. Chi, Quoc V Le, and Minmin Chen. 2025. Evolve: Evaluating and optimizing LLMs for exploration.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2020. DORB: Dynamically optimizing multiple rewards with bandits. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7766–7780, Online. Association for Computational Linguistics.

Julien Perez and Tomi Silander. 2018. Contextual memory bandit for pro-active dialog engagement.

Pavel Petrushkov, Shahram Khadivi, and Evgeny Matusov. 2018. Learning from chunk-based feedback in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 326–331, Melbourne, Australia. Association for Computational Linguistics.

Artem Sokolov, Julia Kreutzer, Christopher Lo, and Stefan Riezler. 2016. Learning structured predictors from bandit feedback for interactive NLP. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1620, Berlin, Germany. Association for Computational Linguistics.

Jiahang Sun, Zhiyong Wang, Runhan Yang, Chenjun Xiao, John C. S. Lui, and Zhongxiang Dai. 2025. Large language model-enhanced multi-armed bandits. *Preprint*, arXiv:2502.01118.

Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. 2024. Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs.

Sohini Upadhyay, Mayank Agarwal, Djallel Bouneffouf, and Yasaman Khazaeni. 2019. A bandit approach to posterior dialog orchestration under a budget. *Preprint*, arXiv:1906.09384.

Inigo Urteaga, Moulay Zaidane Draidia, Tomer Lancewicki, and Shahram Khadivi. 2023. Multi-armed bandits for resource efficient, online optimization of language model pre-training: the use case of dynamic masking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10609–10627, Toronto, Canada. Association for Computational Linguistics.

Bram van den Akker, Olivier Jeunen, Ying Li, Ben London, Zahra Nazari, and Devesh Parekh. 2023. Practical bandits: An industry perspective. *Preprint*, arXiv:2302.01223.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Qiu Jiantao, Lei Cao, Ju Fan, Ye Yuan, Guoren Wang, and Conghui He. 2025. Harnessing diversity for important data selection in pretraining large language models. In *The Thirteenth International Conference on Learning Representations*.

Shuo Zhang, Junzhou Zhao, Pinghui Wang, Tianxiang Wang, Zi Liang, Jing Tao, Yi Huang, and Junlan Feng. 2023. Multi-action dialog policy learning from logged user feedback. *Preprint*, arXiv:2302.13505.

Xiaoying Zhang, Hong Xie, Hang Li, and John C.S. Lui. 2020. Conversational contextual bandit: Algorithm and application. In *Proceedings of The Web Conference 2020*, WWW '20, page 662–672, New York, NY, USA. Association for Computing Machinery.

Zheqing Zhu and Benjamin Van Roy. 2023. Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3636–3646, New York, NY, USA. Association for Computing Machinery.