A Dataset Description

This section provides a detailed description of the dataset, providing the Motivation, Composition, Collection Process, Preprocessing, Uses, Distribution, and Maintenance, which are the key stages of our proposed `ReviewMT` dataset lifecycle.

A.1 Motivation

For what purpose was the dataset created? The `ReviewMT` dataset was created to advance the research of LLMs in the academic peer review process. The primary purpose of the dataset is to capture the dynamic and iterative nature of real-world peer reviews, facilitating the development of LLMs capable of engaging in realistic, multi-turn dialogues. By providing a comprehensive resource that includes detailed interactions between authors, reviewers, and decision-makers.

A.2 Composition

What do the instance that comprise the dataset represent (e.g., documents, photos, people, countries?) Does the dataset contain all possible instances or is it a sample of instances from a larger set? What data does each instance consist of? The `ReviewMT` dataset comprises detailed instances of peer review interactions for academic papers. Each instance represents a complete peer review cycle, capturing the multi-turn dialogue between authors, reviewers, and decision-makers. The dataset includes documents such as the full texts of academic papers from the ICLR spanning 2017 to 2024 and NeurIPS spanning 2021 to 2023. Additionally, it includes initial reviews, rebuttals, final reviews, and decision notes associated with each paper.

How many instances are there in total (of each type, if appropriate)? We provide detailed statistics of the dataset in Table 1. The `ReviewMT` dataset encompasses a significant number of instances across different years and sources, reflecting its comprehensiveness and depth. In total, the `ReviewMT` dataset combines 30,854 papers, 110,642 reviews, and 661,935,412 tokens. This rich and comprehensive dataset is an invaluable resource for training and evaluating LLMs in the context of academic peer review. It captures the iterative and detailed nature of the peer review process, providing a robust foundation for developing advanced LLMs capable of engaging in realistic, multi-turn peer review dialogues.

Is there a label or target associated with each instance? Is any information missing from individual instances? The dataset ensures that each instance is rich with information, supporting multi-turn dialogues that accurately reflect real-world peer review processes. There are no explicit labels or targets associated with each instance, as the primary goal is to capture the iterative and dynamic nature of peer reviews rather than to classify or label the data. While the dataset aims to be comprehensive, it is a curated sample designed to include diverse examples from reputable sources.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No critical information is missing from individual instances, although the depth and detail of reviews can vary. This variability is an inherent characteristic of the peer review process, reflecting differences in reviewer engagement and thoroughness. Relationships between the components of each instance, such as the sequence of reviews and responses for a particular paper, are explicitly maintained to preserve the integrity of the dialogue structure.

Are there any errors, sources of noise, or redundancies in the dataset? Is the dataset self-contained, or does it link to or otherwise rely on external resources? The dataset may contain some sources of noise or redundancies, such as variations in review quality and discrepancies in feedback depth. These elements are preserved to provide a realistic representation of the peer review process. The dataset is self-contained and does not rely on external resources.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? Confidential data within the dataset has been anonymized to protect privacy. The dataset does not include sensitive information protected by legal privilege or doctor-patient confidentiality. Any potentially offensive, insulting, or threatening content is inherently excluded by those publication sources, as the focus is on academic peer review interactions.

A.3 Collection Process

How was the data associated with each instance acquired? What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? The data associated with each instance in the `ReviewMT` dataset was acquired through a combination of automated and manual processes. For papers from ICLR and NeurIPS, we utilized the official API [25] to extract titles, abstracts, and other relevant metadata. The full texts of the papers were obtained as PDF files, which were then converted to text using a software tool called Marker [29]. Marker was chosen for its ability to render text with markdown grammar, preserving the structural and formatting fidelity of the original documents.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? We were primarily responsible for the initial data extraction, validation, and preprocessing. Compensation for us was provided through research grants and funding.

Over what timeframe was the data collected? The data was collected over a timeframe of two months, from the initial planning and setup phases through to the validation and formatting stages.

A.4 Preprocessing

Was any preprocessing/cleaning/labeling of the data done? First, the raw data extracted from the sources was cleaned to remove any extraneous information and ensure consistency. This involved standardizing text formats, correcting any conversion errors from the PDF to text conversion process, and removing any non-text elements that were not relevant to the peer review process. Special attention was given to maintaining the structural integrity of the papers, preserving sections such as titles, abstracts, and main text in a clear and readable format. The dataset was then structured to facilitate multi-turn dialogues. As shown in Figure 2, each instance was organized to include fields for each turn of the review process: initial reviews, author rebuttals, final reviews, and meta reviews along with the final decision. All the data was stored in a JSON format.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? The raw data was saved in addition to the preprocessed, cleaned, and labeled data to support unanticipated future uses. This ensures that the dataset can be revisited and potentially reprocessed as new techniques and requirements emerge, offering flexibility for ongoing and future research. The scripts used for collecting and preprocessing the data are available.

Is the software that was used to preprocess/clean/label the data available? The software used included a combination of open-source tools and custom scripts. Marker [29] was utilized for the PDF to text conversion, ensuring the preservation of text structure and formatting.

A.5 Uses

Has the dataset been used for any tasks already? The `ReviewMT` dataset has been designed to support a variety of research tasks related to the peer review process, although its primary focus is to facilitate the development and evaluation of LLMs in the context of academic peer reviews. As of now, the dataset has been used to assess the performance of LLMs in generating realistic and constructive peer reviews. These initial studies have shown promising results, demonstrating the potential of the dataset to enhance LLM capabilities in complex academic dialogues.

Is there a repository that links to any or all papers or systems that use the dataset? A GitHub repository has been established to host the dataset. This repository is accessible to the broader research community.

What (other) tasks could the dataset be used for? The dataset has a wide range of potential applications related to the academic peer review process. It can be used to evaluate the performance of LLMs in generating reviews, assess the quality and effectiveness of peer reviews, and analyze the dynamics of multi-turn dialogues. The dataset can also be used to develop and evaluate new algorithms for summarization, sentiment analysis, and dialogue generation.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? The dataset is heavily focused on specific conferences and journals, which may introduce biases related to those particular academic communities. Researchers should be aware of these potential limitations and consider them when designing experiments and interpreting results.

Are there tasks for which the dataset should not be used? There are also certain tasks for which the dataset might not be suitable. Given the anonymization of personally identifiable information, the dataset should not be used for tasks that require identifying or analyzing individual authors or reviewers. Additionally, the dataset should not be used for any applications that could compromise the confidentiality or integrity of the peer review process, such as attempting to deanonymize reviews or use the data in ways that could influence ongoing review processes.

A.6 Distribution

Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created? The `ReviewMT` dataset will be made available to third parties. The dataset will be released under a Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license. This licensing will allow researchers to use, modify, and share the dataset as long as they provide appropriate credit, do not use it for commercial purposes, and distribute any derivative works under the same license.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? We will provide detailed scripts and processed datasets via GitHub.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? No third parties have imposed intellectual property (IP)-based or other restrictions on the data associated with the instances in the `ReviewMT` dataset. The dataset has been created from publicly accessible documents and reviews, which have been properly anonymized and processed to comply with ethical and legal standards.

A.7 Maintenance

Is there an erratum? An erratum will be maintained to address any errors or updates necessary after the initial release. The erratum will be accessible through the GitHub repository hosting the dataset, where users can find detailed descriptions of any corrections or changes made to the dataset. The dataset will be periodically updated to correct labeling errors, and add new instances if necessary.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances? Since the dataset does not contain personally identifiable information directly associated with living individuals, there are no specific limits on the retention of the data.

Will older versions of the dataset continue to be supported/hosted/maintained? If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Older versions of the dataset will continue to be supported, hosted, and maintained to ensure that ongoing research projects that rely on specific versions can proceed without disruption. Users will be able to access and reference these older versions via the GitHub repository. We will provide detailed scripts for researchers who want to contribute to our `ReviewMT` dataset.

A.8 Dataset Statistics

The `ReviewMT` dataset compiles comprehensive statistics from premier AI conferences, including ICLR and NeurIPS, providing valuable insights into the scale and evolution of scholarly contributions in the field. Figure 3 illustrates the temporal and venue-specific distribution of papers and reviews, highlighting the dataset's extensive coverage—comprising 30,854 papers, 110,642 reviews, and over 661 million tokens (tokenized using LLaMA-3). This visualization underscores the dataset's significance as a resource for advancing AI research, particularly in understanding peer-review dynamics.

A.9 Performance Analysis of Large Language Models

Figure 4 presents a radar chart that quantitatively compares the performance of various large language models (LLMs) on the ReviewMT dataset across multiple text similarity metrics. This analysis reveals pronounced disparities between zero-shot and supervised fine-tuned models, with the latter consistently outperforming the former across nearly all evaluated dimensions. These results validate the efficacy of supervised fine-tuning and emphasize the ReviewMT dataset's utility as a challenging benchmark for LLM evaluation in review generation tasks.

A.10 GPTScore Evaluation of Supervised Fine-Tuned Models

Table 5 provides a detailed evaluation of supervised fine-tuned LLMs on the test split of the ReviewMT dataset, using the GPTScore metric. This table reports performance across key dimensions—consistency, coherence, fluency, correctness, and semantics—alongside an overall score, offering a granular perspective on model capabilities in generating high-quality reviews. The data demonstrate significant performance variations among models, reinforcing the dataset's role in benchmarking advanced LLMs for peer-review applications.

A.11 Author Statement

The authors of this scientific paper bear full responsibility for any violation of rights that may arise from the collection of the data included in this research.

B Baseline Models

In this section, we present the baseline models used in our experiments to evaluate the effectiveness of the ReviewMT dataset. Each model is described briefly, highlighting its key features, training methodology, and relevance to the peer review task.

B.1 LLaMA-3

LLaMA-3 [21] is an advanced large language model developed by Meta AI. Building on the success of its predecessors [38, 39], LLaMA-3 features a tokenizer with a vocabulary of 128K tokens that encodes language more efficiently, leading to substantially improved performance. The model is renowned for its scalability and efficiency in handling long-context scenarios, making it well-suited for the multi-turn dialogues inherent in peer review processes. LLaMA-3 incorporates extensive pre-training on diverse datasets, followed by fine-tuning on task-specific data, enabling it to generate high-quality, context-aware responses.

B.2 Qwen

Qwen [3] is a comprehensive language model series that encompasses distinct models with varying parameter counts. It includes Qwen, the base pretrained language models, and Qwen-Chat, the chat models fine-tuned with human alignment techniques. The base language models consistently demonstrate superior performance across a multitude of downstream tasks, while the chat models, particularly those trained using Reinforcement Learning from Human Feedback (RLHF), are highly competitive. The chat models possess advanced tool-use and planning capabilities for creating agent applications, showcasing impressive performance when compared to larger models on complex tasks.

B.3 Qwen2

Qwen2 [44] represents the next iteration in the Qwen model series, incorporating a more extensive and diverse dataset, with a particular emphasis on expanding the model's understanding of code and mathematical reasoning. This enriched training data includes a wider array of linguistic sources, which is hypothesized to enhance the model's reasoning capabilities, particularly in areas that require precise logical and computational thinking. The improvements in Qwen2 are aimed at bolstering both its general language understanding and specialized performance in tasks involving complex problem-solving, code generation, and mathematical reasoning, positioning it as a highly capable successor to the original Qwen models.

B.4 Baichuan2

Baichuan2 [43] is a series of large-scale multilingual language models with 7 billion and 13 billion parameters, trained from scratch on 2.6 trillion tokens. Baichuan2 matches or outperforms other open-source models of similar size on public benchmarks like MMLU, CMMLU, GSM8K, and HumanEval. Furthermore, Baichuan2 excels in vertical domains such as medicine and law, making it particularly relevant for peer review tasks that require specialized knowledge.

B.5 ChatGLM3

ChatGLM3 [46] is a bilingual (English and Chinese) pre-trained language model with 130 billion parameters based on the GLM architecture [10] developed by Zhipu AI. It is an attempt to open-source a 100B-scale model at least as good as GPT-3 (davinci) and demonstrate how models of such scale can be successfully pre-trained. ChatGLM3's ability to handle bilingual tasks enhances its applicability in peer review contexts where multilingual support might be necessary.

B.6 GLM-4

GLM-4 [12] is a powerful language model pre-trained on an extensive dataset comprising ten trillion tokens, primarily in Chinese and English, supplemented by a smaller corpus from 24 additional languages. This model is meticulously aligned for optimal performance in both Chinese and English, achieved through a multi-stage post-training process that incorporates supervised fine-tuning and human feedback. The robust alignment techniques not only improve its linguistic capabilities but also enhance its applicability across various multilingual tasks, making GLM-4 a highly versatile tool for applications requiring nuanced understanding and generation of text in multiple languages.

B.7 Gemma

Gemma [37] is a family of lightweight, state-of-the-art open models derived from the research and technology used to create Gemini

Table 5: GPTScore evaluation of supervised fine-tuned LLMs on the test set of the ReviewMT dataset.

Model	Consistency ↑	Coherence ↑	Fluency ↑	Correctness ↓	Semantics ↑	Overall Score ↑
GPT-4o	54.42±1.10	37.19±1.89	19.44±1.23	67.31±1.03	36.94±1.66	43.06±1.42
LLaMA-3	32.40±2.70	27.38±2.60	9.32±2.42	40.22±2.83	21.44±2.90	26.15±2.70
Qwen	44.56±1.06	30.41±1.16	16.74±1.22	53.54±1.04	30.12±1.34	35.07±1.17
Qwen2	48.61±0.55	31.13±0.55	18.92±0.76	58.73±0.56	33.87±0.78	38.25±0.65
Baichuan2	34.70±2.31	27.87±2.28	10.75±2.21	43.31±2.31	23.09±2.44	27.94±2.31
ChatGLM3	38.21±1.85	28.31±1.78	12.32±1.80	47.20±1.86	25.76±2.19	30.36±1.90
GLM-4	47.03±0.72	31.41±0.78	18.69±0.96	56.54±0.62	32.84±0.91	37.30±0.81
Gemma	41.05±1.56	29.20±1.43	14.60±1.54	50.28±1.53	27.67±1.74	32.56±1.56
Gemma2	46.14±0.98	30.67±0.94	17.66±1.05	55.15±0.99	31.43±1.16	36.21±1.03
DeepSeek	33.37±2.45	27.54±2.41	9.88±2.27	41.42±2.59	22.47±2.62	26.94±2.47
Yuan-2	42.68±1.37	29.77±1.33	15.68±1.27	52.27±1.28	29.44±1.55	33.97±1.36
Falcon	39.98±1.76	29.32±1.64	13.28±1.71	48.20±1.68	27.06±1.85	31.57±1.73
Yi-1.5	37.02±2.10	28.51±1.93	11.69±1.95	44.67±2.23	24.50±2.31	29.28±2.11

models. Gemma models demonstrate strong performance across academic benchmarks for language understanding, reasoning, and safety. There are two sizes of Gemma models (2 billion and 7 billion parameters), and they provide both pre-trained and fine-tuned checkpoints. Gemma outperforms similarly sized open models on 11 out of 18 text-based tasks, making it a versatile choice for academic peer review.

B.8 Gemma2

Gemma2 [32] marks an exciting expansion of the Gemma family, featuring models that range from 2 billion to an impressive 27 billion parameters. This updated version incorporates several well-established technical enhancements to the Transformer architecture, including interleaved local-global attention mechanisms and group-query attention strategies. Additionally, the 2 billion and 9 billion parameter models are trained using knowledge distillation techniques, rather than relying solely on next-token prediction. As a result, models achieve exceptional performance relative to their size and provide competitive alternatives to larger models, often outperforming those that are 2 to 3 times their scale.

B.9 DeepSeek

DeepSeek [5] is a series of open-source models trained from scratch on a vast dataset of 2 trillion tokens in both English and Chinese. The models calibrate scaling laws from previous work and propose a new optimal model/data scaling-up allocation strategy. Additionally, DeepSeek introduces a method to predict the near-optimal batch size and learning rate with a given compute budget. These models also highlight that scaling laws are related to data quality, guiding the best hyper-parameters for pre-training. DeepSeek's comprehensive evaluation makes it a robust candidate for generating thorough and balanced reviews.

B.10 Yuan-2

Yuan-2 [41] introduces the Localized Filtering-based Attention (LFA) mechanism to incorporate prior knowledge of local dependencies of natural language into attention mechanisms. The model employs a data filtering and generation method to build high-quality pre-training and fine-tuning datasets. Additionally, a distributed training method with non-uniform pipeline, data, and optimizer parallels is proposed, significantly reducing the bandwidth requirements of intra-node communication. Yuan models display impressive abilities in code generation, math problem-solving, and chat, making them well-suited for the detailed analytical tasks required in peer review.

B.11 Falcon

Falcon [1] comprises a series of three causal decoder-only models, pretrained on an extensive corpus of 3.5 trillion tokens. A key aspect of the Falcon model series is its focus on scaling the quality of data rather than simply increasing its quantity. To achieve this, the team applies rigorous filtering and deduplication processes, resulting in the creation of a high-quality English web dataset of 5 trillion tokens, ensuring no repetition of data during the training phase.

B.12 Yi-1.5

Yi-1.5 [45] is built upon 6B and 34B parameter pretrained models, which are further extended into specialized variants, including chat models, long-context models with up to 200K tokens, depth-upscaled models, and vision-language models. The base models in the Yi-1.5 series achieve impressive results on a variety of benchmarks, such as MMLU, showcasing their capability in handling a broad range of linguistic tasks. The fine-tuned chat models are particularly notable, consistently achieving high human preference ratings on widely recognized evaluation platforms such as AlpacaEval and Chatbot Arena. Yi-1.5's versatility across multiple domains—including dialogue systems, long-context understanding, and multimodal tasks—positions it as a powerful tool for complex applications, such as generating in-depth reviews and critiques in peer-review settings.

C Implementation Details

We implement the above baseline models using the LLaMA-Factory framework [50], which provides a robust and scalable platform for training and deploying large language models. All experiments were conducted on a cluster of NVIDIA A100 GPUs with 80GB of VRAM each. The models were implemented using the PyTorch deep learning framework. We used the Hugging Face Transformers library for model definitions and pre-trained weights, and the LLaMA-Factory for distributed training and fine-tuning. The detailed configure files and training scripts are available in GitHub repository.

We provide the pseudocode for the inference process in Algorithm 1. The initial prompt instructs the LLMs to summarize and remember the paper content, which helps to reduce the context size. In this pseudocode, the process begins with an initial prompt to set the context for the LLMs, instructing them to summarize and retain the key points of the paper, thus reducing the context size. This is followed by either appending the title and abstract or the entire paper to the context, depending on the model's capabilities. For models that cannot handle long contexts (ChatGLM3 and Yuan), only the title and abstract are used. The initial conversation includes the summarized context, with responses generated by the LLMs being recorded in the conversation history. As the dialogue progresses through multiple turns, each interaction is appended to the conversation history, with the LLMs generating replies based on this accumulated context. Finally, a decision prompt is issued, instructing the LLMs to take on the role of a decision maker, tasked with providing an accept or reject recommendation for the paper, along with their reasoning.

Algorithm 1 Pseudocode of Inference

```
initial_prompt = "This is a peer-review system. You will be
     assigned with roles such as author, reviewer or decision
     maker to perform different tasks. "
context = initial_prompt + "Please summarize and remember this
      paper: "
context = (context + title_abs) if not full_context else (
     context + paper)

chat_reply = LLMs([{"role": "user", "content": context}])
conversation_history = [
    {"role": "user", "content": initial_prompt},
    {"role": "assistant", "content": chat_reply}
]

for current_turn in dialogue_turns:
    conversation_history.append({"role": "user", "content":
          current_turn})
    chat_reply = LLM(conversation_history)
    coversation_history.append({"role": "assistant", "content":
           chat_reply})

decision_prompt = "You are the Decision Maker. Task: Suggest
     Accept or Reject for this paper, and provide reasons."
coversation_history.append({"role": "user", "content":
     decision_prompt})
decision = LLM(coversation_history)
```

D Metric Details

In the main text, we introduce several metrics to evaluate the validity and quality of the responses generated by the LLMs, including paper hit rate, review hit rate, decision hit rate, mean absolute error (MAE) for review scores, and F1-score for decision accuracy. In this section, we provide detailed definitions of these metrics.

(1) Text quality evaluation For all text replies, including the reviewers' initial reviews, the authors' responses, the reviewers' final reviews, and the decision makers' meta reviews, we employ text similarity metrics to assess the quality of the generated text. These metrics include:

- **BLEU-2** and **BLEU-4** [28]: Measures n-gram precision by comparing the generated text to a reference text, focusing on 2-gram and 4-gram overlaps respectively.
- **ROUGE-1**, **ROUGE-2**, and **ROUGE-L** [18]: Measures f1-score of unigram, bigram, and longest common subsequence overlaps between the generated and the reference text.
- **METEOR** [4]: A measure of alignment between the generated and reference texts, incorporating stemming and synonymy for sensitivity to variations.
- **BERT Score** [49]: The similarity between the generated and reference texts using contextual embeddings from BERT [8].

(2) Validity of response Given the long-context nature of peer-review documents, which average over 20,000 tokens per paper, LLMs may occasionally fail to provide valid responses. We use the hit rates to evaluate the validity of the responses:

- **Paper hit rate (P-hr)**: Measures whether the LLM-generated response addresses the paper content. If the LLM fails to respond to the paper, the hit rate is 0.
- **Review hit rate (R-hr)**: Evaluates whether the LLM-generated final review includes a score. If it fails to provide the score, the hit rate is 0.
- **Decision hit rate (D-hr)**: Assesses whether the LLM-generated decision includes a clear accept or reject outcome. If the LLM fails to respond with a decision, the hit rate is 0.

(3) Score and decision evaluation To evaluate the accuracy of the final review scores provided by the reviewers and the decisions (accept or reject) made by the decision makers, we use:

- **Mean Absolute Error (MAE)**: Measures the average absolute difference between the scores given by the LLM and the actual scores provided by human reviewers.
- **F1-score**: Combines precision and recall to measure the binary classification of decisions.

(4) GPTScore evaluation GPTScore evaluates generated texts by leveraging the emergent capabilities of generative pre-trained models. While the GPTScore [11] was designed using GPT-3, the evolution of OpenAI's models has led to configurational differences in newer versions, rendering the original approach outdated. In our evaluation, we employ the OPT-6.7B [48] as the base model for GPTScore assessment. GPTScore measures the quality of text generated by LLMs based on five key criteria:

- Consistency: The degree to which the generated text aligns with the factual information.
- Conherence: The logical flow and organization of the text, ensuring that ideas are well-connected and clearly presented.
- Fluency: The grammatical correctness and natural flow of the text, ensuring readability.
- Correctness: The accuracy and error-free nature of the text.
- Semantics: The relevance and appropriateness of the content within its specific context.

(5) Human evaluation We incorporate blind human evaluation as a key component of our experimental setup. We engage five experienced reviewers, all of whom have extensive expertise in peer reviewing for top-tier AI/ML conferences. They are tasked with evaluating the outputs of the LLMs, including the initial peer reviews (H-R), the author rebuttals (H-A), and the final decisions (H-D). The human evaluators are instructed to assess each generated output based on a 10-point scale, where a score of 0 indicates the lowest quality, and 10 represents the highest level of quality.

E Spearman Correlation Analysis

In Table 4 of the main text, we present the Spearman correlation matrix, computed using the mean values of each evaluation metric across all human reviewers. To provide a more detailed analysis, we further computed the Spearman correlation between the individual ratings assigned by the five human reviewers. This analysis helps quantify the level of agreement among reviewers for each metric. Table 6 presents the sample-level pairwise Spearman correlations across different categories (R = Reviewer, A = Author, D = Decision Maker). As shown, the correlation values between individual reviewers vary across categories, which is expected due to differences in evaluative perspectives.

Additionally, to gain deeper insights into the relationship between human ratings and GPTScore, we computed the global-level Spearman correlation using the mean ratings aggregated across all reviewers, as presented in Table 7. This analysis allows us to assess the overall consistency of human evaluations. Notably, when individual ratings are aggregated into global mean scores, the correlations between mean scores for each category increase significantly, indicating stronger alignment and stability across reviewers. For example, as shown in Table 7, Human1-R exhibits a strong correlation with both Human2-R (0.9063) and Human3-R (0.9118), suggesting that aggregating individual ratings enhances consistency and reduces variability in evaluations.

F Test Data List

Table 6: Sample-wise Spearman correlation between the individual ratings of the five human reviewers.

Metric	Human1-R	Human1-A	Human1-D	Human2-R	Human2-A	Human2-D	Human3-R	Human3-A	Human3-D	Human4-R	Human4-A	Human4-D	Human5-R	Human5-A	Human5-D
Human1-R	1.0000	0.4157	0.3815	0.4007	0.4085	0.2321	0.2525	0.2350	0.2209	0.2565	0.4396	0.3870	0.3574	0.3842	0.4185
Human1-A	0.4157	1.0000	0.4045	0.3705	0.3961	0.2937	0.2543	0.2591	0.2422	0.2634	0.4157	0.3836	0.3915	0.4032	0.4181
Human1-D	0.3815	0.4045	1.0000	0.3933	0.3771	0.2809	0.2251	0.2301	0.2532	0.3077	0.3802	0.4090	0.4198	0.3807	0.3957
Human2-R	0.4007	0.3705	0.3933	1.0000	0.3954	0.2439	0.2344	0.2308	0.2565	0.2313	0.3994	0.3944	0.3786	0.4003	0.4312
Human2-A	0.4085	0.3961	0.3771	0.3954	1.0000	0.2644	0.2734	0.2851	0.3075	0.2632	0.4089	0.3955	0.4107	0.4083	0.3992
Human2-D	0.2321	0.2937	0.2809	0.2439	0.2644	1.0000	0.5041	0.5105	0.5054	0.5105	0.2949	0.2542	0.2249	0.2619	0.2334
Human3-R	0.2525	0.2543	0.2251	0.2344	0.2734	0.5041	1.0000	0.4671	0.5039	0.5009	0.2665	0.2696	0.2303	0.2156	0.2346
Human3-A	0.2350	0.2591	0.2301	0.2308	0.2851	0.5105	0.4671	1.0000	0.5179	0.4965	0.2456	0.2416	0.2123	0.2261	0.2157
Human3-D	0.2209	0.2422	0.2532	0.2565	0.3075	0.5054	0.5039	0.5179	1.0000	0.5221	0.2645	0.2537	0.2533	0.2374	0.2309
Human4-R	0.2565	0.2634	0.3077	0.2313	0.2632	0.5105	0.5009	0.4965	0.5221	1.0000	0.2706	0.2510	0.2284	0.2493	0.2436
Human4-A	0.4396	0.4157	0.3802	0.3994	0.4089	0.2949	0.2665	0.2456	0.2645	0.2706	1.0000	0.4800	0.4928	0.5098	0.4829
Human4-D	0.3870	0.3836	0.4090	0.3944	0.3955	0.2542	0.2696	0.2416	0.2537	0.2510	0.4800	1.0000	0.4352	0.4406	0.4426
Human5-R	0.3574	0.3915	0.4198	0.3786	0.4107	0.2249	0.2303	0.2123	0.2533	0.2284	0.4928	0.4352	1.0000	0.4793	0.4545
Human5-A	0.3842	0.4032	0.3807	0.4003	0.4083	0.2619	0.2156	0.2261	0.2374	0.2493	0.5098	0.4406	0.4793	1.0000	0.4192
Human5-D	0.4185	0.4181	0.3957	0.4312	0.3992	0.2334	0.2346	0.2157	0.2309	0.2436	0.4829	0.4426	0.4545	0.4192	1.0000

Table 7: Sample-wise Spearman correlation between the individual ratings of the five human reviewers.

Metric	Human1-R	Human1-A	Human1-D	Human2-R	Human2-A	Human2-D	Human3-R	Human3-A	Human3-D	Human4-R	Human4-A	Human4-D	Human5-R	Human5-A	Human5-D
Human1-R	1.0000	0.3333	0.7466	0.9063	0.3388	0.7796	0.9118	0.3333	0.8110	0.9239	0.4270	0.7834	0.9118	0.3774	0.7851
Human1-A	0.3333	1.0000	0.4160	0.3994	0.9945	0.3499	0.3719	1.0000	0.2759	0.4260	0.9835	0.3393	0.6143	0.9945	0.3333
Human1-D	0.7466	0.4160	1.0000	0.8072	0.4215	0.9780	0.8347	0.4160	0.9600	0.8409	0.4766	0.9710	0.7796	0.4380	0.9835
Human2-R	0.9063	0.3994	0.8072	1.0000	0.4050	0.7851	0.9945	0.3994	0.7890	0.9295	0.4766	0.7779	0.9339	0.4435	0.8127
Human2-A	0.3388	0.9945	0.4215	0.4050	1.0000	0.3554	0.3774	0.9945	0.2759	0.4315	0.9890	0.3448	0.6198	0.9890	0.3278
Human2-D	0.7796	0.3499	0.9780	0.7851	0.3554	1.0000	0.8127	0.3499	0.9821	0.8299	0.4215	0.9986	0.7686	0.3829	0.9725
Human3-R	0.9118	0.3719	0.8347	0.9945	0.3774	0.8127	1.0000	0.3719	0.8221	0.9405	0.4545	0.8055	0.9229	0.4160	0.8457
Human3-A	0.3333	1.0000	0.4160	0.3994	0.9945	0.3499	0.3719	1.0000	0.2759	0.4260	0.9835	0.3393	0.6143	0.9945	0.3333
Human3-D	0.8110	0.2759	0.9600	0.7890	0.2759	0.9821	0.8221	0.2759	1.0000	0.8670	0.3476	0.9862	0.7503	0.3034	0.9821
Human4-R	0.9239	0.4260	0.8409	0.9295	0.4315	0.8299	0.9405	0.4260	0.8670	1.0000	0.5145	0.8338	0.9239	0.4592	0.8631
Human4-A	0.4270	0.9835	0.4766	0.4766	0.9890	0.4215	0.4545	0.9835	0.3476	0.5145	1.0000	0.4138	0.6970	0.9890	0.3884
Human4-D	0.7834	0.3393	0.9710	0.7779	0.3448	0.9986	0.8055	0.3393	0.9862	0.8338	0.4138	1.0000	0.7641	0.3724	0.9683
Human5-R	0.9118	0.6143	0.7796	0.9339	0.6198	0.7686	0.9229	0.6143	0.7503	0.9239	0.6970	0.7641	1.0000	0.6584	0.7631
Human5-A	0.3774	0.9945	0.4380	0.4435	0.9890	0.3829	0.4160	0.9945	0.3034	0.4592	0.9890	0.3724	0.6584	1.0000	0.3554
Human5-D	0.7851	0.3333	0.9835	0.8127	0.3278	0.9725	0.8457	0.3333	0.9821	0.8631	0.3884	0.9683	0.7631	0.3554	1.0000

Table 8: Test data list of ReviewMT.

ID	Title
1	Breaking Physical and Linguistic Borders: Multilingual Federated Prompt Tuning for Low-Resource Languages
2	On the generalization capacity of neural networks during generic multimodal reasoning
3	Learning Mean Field Games on Sparse Graphs: A Hybrid Graphex Approach
4	Out-of-Variable Generalisation for Discriminative Models
5	Neural Sinkhorn Gradient Flow
6	Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting
7	Pricing with Contextual Elasticity and Heteroscedastic Valuation
8	FABRIC: Personalizing Diffusion Models with Iterative Feedback
9	Learning energy-based models by self-normalising the likelihood
10	Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Image Diffusion Models
11	AUTOPARLLM: GNN-Guided Automatic Code Parallelization using Large Language Models
12	Nemesis: Normalizing the Soft-prompt Vectors of Vision-Language Models
13	Fast and unified path gradient estimators for normalizing flows
14	Certified Robustness on Visual Graph Matching via Searching Optimal Smoothing Range
15	Conformal Prediction for Deep Classifier via Label Ranking
16	Binder: Hierarchical Concept Representation through Order Embedding of Binary Vectors
17	Long-Term Typhoon Trajectory Prediction: A Physics-Conditioned Approach Without Reanalysis Data
18	Rethinking RGB Color Representation for Image Restoration Models
19	CEIR: Concept-based Explainable Image Representation Learning
20	AttributionLab: Faithfulness of Feature Attribution Under Controllable Environments
21	Rigid Protein-Protein Docking via Equivariant Elliptic-Paraboloid Interface Prediction
22	Sparling: Learning Latent Representations With Extremely Sparse Activations
23	Using Machine Learning Models to Predict Genitourinary Involvement Among Gastrointestinal Stromal Tumour Patients
24	On the Joint Interaction of Models, Data, and Features
25	A ROBUST DIFFERENTIAL NEURAL ODE OPTIMIZER
26	Shadow Cones: A Generalized Framework for Partial Order Embeddings
27	How do Language Models Bind Entities in Context?
28	On the Limitations of Temperature Scaling for Distributions with Overlaps
29	If there is no underfitting, there is no Cold Posterior Effect
30	A Unified View on Neural Message Passing with Opinion Dynamics for Social Networks
31	FROSTER: Frozen CLIP is A Strong Teacher for Open-Vocabulary Action Recognition

ID	Title
32	Detecting Pretraining Data from Large Language Models
33	Dozerformer: Sequence Adaptive Sparse Transformer for Multivariate Time Series Forecasting
34	Crystals with Transformers on Graphs, for predictions of crystal material properties
35	Task Adaptation from Skills: Information Geometry, Disentanglement, and New Objectives for Unsupervised Reinforcement Learning
36	Explore, Establish, Exploit: Red Teaming Language Models from Scratch
37	Neural Processing of Tri-Plane Hybrid Neural Fields
38	DISCRET: a self-interpretable framework for treatment effect estimation
39	Adversarial Instance Attacks for Interactions between Human and Object
40	Mildly Overparameterized ReLU Networks Have a Favorable Loss Landscape
41	Error Norm Truncation: Robust Training in the Presence of Data Noise for Text Generation Models
42	CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling
43	Symmetry Leads to Structured Constraint of Learning
44	Protein Discovery with Discrete Walk-Jump Sampling
45	H-Rockmate: Hierarchical Approach for Efficient Re-materialization of Large Neural Networks
46	TCD: TEXT IMAGE CHANGE DETECTION FOR MULTILINGUAL DOCUMENT COMPARISON
47	Exploring the Impact of Information Entropy Change in Learning Systems
48	Agent Instructs Large Language Models to be General Zero-Shot Reasoners
49	Rethinking Spectral Graph Neural Networks with Spatially Adaptive Filtering
50	Stability Analysis of Various Symbolic Rule Extraction Methods from Recurrent Neural Network
51	Learning with Language Inference and Tips for Continual Reinforcement Learning
52	Collaboration! Towards Robust Neural Methods for Vehicle Routing Problems
53	SetCSE: Set Operations using Contrastive Learning of Sentence Embeddings
54	Exploiting Code Symmetries for Learning Program Semantics
55	Token Alignment via Character Matching for Subword Completion
56	Dynamic Mode Decomposition-inspired Autoencoders for Reduced-order Modeling and Control of PDEs : Theory and Design
57	AgentBench: Evaluating LLMs as Agents
58	Towards Greener and Sustainable Airside Operations: A Deep Reinforcement Learning Approach to Pushback Rate Control for Mixed-Mode Runways
59	Best Response Shaping
60	Sharp results for NIEP and NMF
61	GEOFFair: a GEOmetric Framework for Fairness
62	MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning
63	GETMusic: Generating Music Tracks with a Unified Representation and Diffusion Framework

ID	Title
64	Making Large Language Models Better Reasoners with Alignment
65	Perceptual Scales Predicted by Fisher Information Metrics
66	An Efficient Tester-Learner for Halfspaces
67	Neural functional a posteriori error estimates
68	DSparsE: Dynamic Sparse Embedding for Knowledge Graph Completion
69	Pre-training with Random Orthogonal Projection Image Modeling
70	Generalist Equivariant Transformer Towards 3D Molecular Interaction Learning
71	Revisit and Outstrip Entity Alignment: A Perspective of Generative Models
72	Massively Scalable Inverse Reinforcement Learning in Google Maps
73	iHyperTime: Interpretable Time Series Generation with Implicit Neural Representations
74	Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching
75	Generating Pragmatic Examples to Train Neural Program Synthesizers
76	Fine-Tuning Is All You Need to Mitigate Backdoor Attacks
77	Synthetic Data as Validation
78	MuseCoco: Generating Symbolic Music from Text
79	Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation
80	Communication-efficient Random-Walk Optimizer for Decentralized Learning
81	Stoichiometry Representation Learning with Polymorphic Crystal Structures
82	Confidence-driven Sampling for Backdoor Attacks
83	A Quadratic Synchronization Rule for Distributed Deep Learning
84	Equivariant Matrix Function Neural Networks
85	Towards Zero Memory Footprint Spiking Neural Network Training
86	An old dog can learn (some) new tricks: A tale of a three-decade old architecture
87	In defense of parameter sharing for model-compression
88	Image Translation as Diffusion Visual Programmers
89	Achieving Fairness in Multi-Agent MDP Using Reinforcement Learning
90	Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in ultra low-data regimes
91	Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion
92	Interaction-centric Hypersphere Reasoning for Multi-person Video HOI Recognition
93	Image Background Serves as Good Proxy for Out-of-distribution Data
94	Pre-Training and Fine-Tuning Generative Flow Networks
95	CI-VAE: a Generative Deep Learning Model for Class-Specific Data Interpolation
96	Rethinking the Noise Schedule of Diffusion-Based Generative Models
97	FPTQ: FINE-GRAINED POST-TRAINING QUANTIZATION FOR LARGE LANGUAGE MODELS

ID	Title
98	Incentivized Truthful Communication for Federated Bandits
99	Reinforcement Learning of Diverse Skills using Mixture of Deep Experts
100	GLASU: A Communication-Efficient Algorithm for Federated Learning with Vertically Distributed Graph Data

G An Example of LLM inference on `ReviewMT`

Context

This is a peer-review system. You will be assigned with roles such as author, reviewer or decision maker to perform different tasks. Please summarize and remember this paper:
Title: Provable Memory Efficient Self-Play Algorithm for Model-free Reinforcement Learning.
Abstract: The thriving field of multi-agent reinforcement learning (MARL) studies how a group of interacting agents make decisions autonomously in a shared dynamic environment. Existing theoretical studies in this area suffer from at least two of the following obstacles: memory inefficiency, the heavy dependence of sample complexity on the long horizon and the large state space, the high computational complexity, non-Markov policy, non-Nash policy, and high burn-in cost. In this work, we take a step towards settling this problem by designing a model-free self-play algorithm *Memory-Efficient Nash Q-Learning (ME-Nash-QL)* for two-player zero-sum Markov games, which is a specific setting of MARL. We prove that ME-Nash-QL can output an an ε-approximate Nash policy with remarkable space complexity $O(SABH)$, sample complexity $\widetilde{O}(H^4SAB/\varepsilon^2)$, and computational complexity $O(T\text{poly}(AB))$, where S is the number of states, $\{A, B\}$ is the number of actions for the two players, H is the horizon length, and T is the number of samples. Notably, our approach outperforms in terms of space complexity compared to existing algorithms for tabular cases. It achieves the lowest computational complexity while preserving Markov policies, setting a new standard. Furthermore, our algorithm outputs a Nash policy and achieves the best sample complexity compared with the existing guarantee for long horizons, i.e. when $\min\{A, B\} \ll H^2$. Our algorithm also achieves the best burn-in cost $O(SAB, mathrmpoly(H))$, whereas previous algorithms need at least $O(S^3AB\,\text{poly}(H))$ to attain the same level of sample complexity with ours.
Main Text: ...

Input

You are Reviewer 1, and your review style is Neutral. Please provide a review based on the paper provided, including a summary, strengths, weaknesses, and any questions you have.

GPT-4o

Summary: This paper proposes ME-Nash-QL, a model-free self-play algorithm for two-player zero-sum Markov games, which achieves provable memory and sample efficiency while outputting a Markov and Nash policy. The authors demonstrate that ME-Nash-QL outperforms existing algorithms in terms of space complexity, sample complexity (especially for long horizons), and computational complexity. They also extend the algorithm to multi-player general-sum Markov games and provide theoretical guarantees.
Strengths: Novel algorithm: ME-Nash-QL introduces an innovative combination of techniques, including reference-advantage decomposition and early-settlement, to improve the sample and computational efficiency of self-play algorithms. Strong theoretical guarantees: The paper provides rigorous proofs for the algorithm's space, sample, and computational complexity. These guarantees are well-supported by detailed analysis and mathematical derivations. Improved performance: ME-Nash-QL achieves the best known space complexity and computational complexity for a Markov/Nash policy algorithm in the tabular case. It also surpasses existing algorithms in sample complexity for long horizons.
Weaknesses: Focus on tabular case: While the paper presents impressive results for the tabular setting, the generalization of ME-Nash-QL to function approximation remains an open question. Addressing this would be crucial for real-world applications. Limited empirical evaluation: The paper focuses primarily on theoretical analysis and does not include any empirical evaluation of the algorithm. While the theoretical guarantees are compelling, empirical results would provide valuable insights and validate the practical effectiveness.
Questions: How does the early-settlement method impact the learning process, and how does it affect the convergence behavior of the algorithm? Could the authors elaborate on the potential challenges of extending ME-Nash-QL to function approximation and discuss possible approaches for addressing these challenges? Would the authors consider incorporating empirical evaluation to complement the theoretical analysis and demonstrate the practical performance of ME-Nash-QL? Is there any possibility for further reducing the burn-in cost of the algorithm? If so, could the authors provide insights into how this could be achieved?
Overall: This paper presents a valuable contribution to the theoretical understanding of self-play algorithms in Markov games. The proposed ME-Nash-QL algorithm offers compelling advantages in terms of memory and sample efficiency. However, the paper would benefit from addressing the identified weaknesses, particularly the lack of empirical evaluation and a more comprehensive discussion of limitations and potential future directions.

Qwen2

Summary: This paper proposes a memory-efficient self-play algorithm for model-free reinforcement learning in multi-agent settings. The algorithm, called Memory-Efficient Self-Play (ME-SP), uses a combination of deep neural networks and a memory-efficient replay buffer to enable self-play without requiring large amounts of memory. The paper also provides theoretical guarantees for the algorithm's performance.
The paper is well-written and the algorithm is clearly explained. The proposed algorithm is simple and efficient, and it is interesting to see how it can be applied to multi-agent settings. The paper also provides theoretical guarantees for the algorithm's performance, which is a significant contribution. However, there are some weaknesses in the paper. First, the experiments are limited to a few small-scale multi-agent games, and it is not clear how the algorithm will perform in more complex settings. Second, the paper does not provide a detailed comparison with other self-play algorithms, which makes it difficult to assess the algorithm's relative strengths and weaknesses. Third, the paper does not discuss the limitations of the algorithm, such as its ability to handle continuous action spaces or its scalability to large-scale multi-agent systems.
Overall: This paper presents an interesting and promising approach to memory-efficient self-play in multi-agent settings. However, there are some limitations and weaknesses that need to be addressed before the algorithm can be considered for wider adoption.

Input

You are the authors. Please respond to Reviewer 1's comments by clarifying the mentioned weaknesses and answering the posed questions.

GPT-4o

We appreciate Reviewer 1's insightful comments and thank them for their time and effort. We agree that the paper could be improved by addressing the identified weaknesses and providing further clarifications.
Focus on Tabular Case: We acknowledge that the current paper primarily focuses on the tabular case. We agree that extending ME-Nash-QL to function approximation is crucial for real-world applications. We are actively researching this extension and plan to include preliminary results in a future version of the paper. We believe that the reference-advantage decomposition technique, which significantly contributes to the sample efficiency of ME-Nash-QL, can be generalized to function approximation settings by employing techniques such as linear function approximation or neural networks. We will also explore the challenges associated with the curse of dimensionality and the need for more complex exploration strategies in the function approximation setting.
Limited Empirical Evaluation: We understand the need for empirical evaluation to validate the practical effectiveness of ME-Nash-QL. While the current paper focuses on theoretical analysis, we are in the process of conducting extensive empirical studies on various benchmark environments, including Go, Atari games, and poker, to demonstrate the practical performance of our algorithm. We will incorporate the results of these studies into a future version of the paper.
Discussion of Limitations: We agree that a more balanced discussion of limitations would strengthen the paper. We will add a dedicated section discussing the potential limitations of ME-Nash-QL, including the dependence of sample complexity on the number of players in the multi-player setting. We will also discuss potential challenges and future research directions related to these limitations.
Early-settlement impact: The early-settlement method significantly impacts the learning process by reducing the variance of the value function estimates. It allows the algorithm to settle on a desirable reference value quickly, leading to a lower burn-in cost and improved sample complexity, especially for long horizons. The convergence behavior is not affected significantly, as the algorithm still guarantees convergence to an ε-approximate Nash equilibrium.
Extending to function approximation: We recognize that extending ME-Nash-QL to function approximation poses challenges. The main challenge lies in effectively incorporating the reference-advantage decomposition technique in the presence of function approximation. We plan to explore different approaches, such as using linear function approximation or neural networks, along with exploration strategies for function approximation settings.

Qwen2

The author doesn't have any rebuttal.

Input

You are Reviewer 1, and your review style is Neutral. Based on the author's response, please provide a final score from 1 to 10 and a confidence from 1 to 5.

GPT-4o

Score: 7. Confidence: 3. The authors have addressed my concerns in a reasonable manner. While the paper still has some weaknesses, particularly the lack of empirical evaluation, the authors have acknowledged these limitations and are working on addressing them.

Qwen2

Score: 7. Confidence: 4. I look forward to seeing the improvements in the revised version of the paper.

Input

You are Reviewer 2, and your review style is Neutral. Please provide a review based on the paper provided, including a summary, strengths, weaknesses, and any questions you have.

GPT-4o

Summary: This paper presents a novel model-free self-play algorithm called Memory-Efficient Nash Q-Learning (ME-Nash-QL) for two-player zero-sum Markov games (TZMG). ME-Nash-QL boasts several advantages over existing algorithms, including: Optimal space complexity: It achieves the optimal space complexity of O(SABH) for tabular cases, significantly outperforming other algorithms. Improved computational complexity: ME-Nash-QL has a computational complexity of O(Tpoly(AB)), lower than previous algorithms like Nash-VI. Superior sample complexity for long horizons: For long horizons, the algorithm achieves the best known sample complexity. Markov and Nash policy output: It outputs a single Markov and Nash policy, unlike other algorithms that may produce non-Markov policies or nested mixtures. The authors also extend ME-Nash-QL to multi-player general-sum Markov games (Multi-ME-Nash-QL), achieving an ε-optimal policy with M players and Ai actions per player.
Strengths: Novel and effective algorithm: ME-Nash-QL is a novel algorithm that achieves significant improvements in memory, computational, and sample complexity, especially for long horizons. Strong theoretical analysis: The paper provides a detailed and thorough theoretical analysis of ME-Nash-QL, proving its desirable properties. Clear and well-structured presentation: The paper is well-written and well-structured, making it easy to follow the algorithm design and theoretical analysis.
Weaknesses: Lack of empirical evaluation: The paper does not present any empirical results to validate the algorithm's performance in practice. While the authors mention ongoing empirical studies, their absence limits the paper's impact and makes it difficult to assess the practical implications of the theoretical findings. Limited discussion on function approximation: While the authors acknowledge the importance of function approximation, the paper provides limited discussion on how ME-Nash-QL can be extended to handle large state spaces. A more detailed discussion of the challenges would be beneficial.
Questions: Could the authors provide insights into the practical implications of the burn-in cost for ME-Nash-QL? How does this cost compare to other algorithms in real-world settings? What are the authors' plans for addressing the limitations regarding function approximation in future work? What approaches are they considering? Do the authors plan to include empirical results in a future version of the paper? If so, could they provide a timeline for this?
Overall: This paper presents a promising theoretical approach to multi-agent reinforcement learning. The proposed ME-Nash-QL algorithm shows significant improvements in key metrics compared to existing methods. However, the lack of empirical evaluation and a detailed discussion of function approximation limits the paper's impact. Including empirical results and a more comprehensive discussion of function approximation in future work would significantly enhance the paper's contribution to the field.

Qwen2

This paper presents a memory-efficient self-play algorithm for model-free reinforcement learning in multi-agent settings. The algorithm, called Memory-Efficient Self-Play (ME-SP), uses a combination of deep neural networks and a memory-efficient replay buffer to enable self-play without requiring large amounts of memory. The paper also provides theoretical guarantees for the algorithm's performance. This paper is well-written and the algorithm is clearly explained. The proposed algorithm is simple and efficient, and it is interesting to see how it can be applied to multi-agent settings. The paper also provides theoretical guarantees for the algorithm's performance, which is a significant contribution.

Input

You are the authors. Please respond to Reviewer 2's comments by clarifying the mentioned weaknesses and answering the posed questions.

GPT-4o

We appreciate Reviewer 2's thoughtful review and insightful comments. We agree that the lack of empirical evaluation and a more in-depth discussion of function approximation are limitations of the current manuscript. We are actively addressing these issues, and we plan to incorporate them into a future version of the paper.
Regarding the lack of empirical evaluation: We are currently conducting extensive empirical studies to validate the performance of ME-Nash-QL on various benchmark environments. These studies involve comparing our algorithm with state-of-the-art model-free and model-based algorithms, including those listed in Table 1. We plan to include these empirical results, along with a detailed analysis of the algorithm's performance.
Regarding the limited discussion on function approximation: We understand that extending ME-Nash-QL to handle large state spaces using function approximation is crucial for real-world applications. We are actively investigating various approaches to address this challenge, including: Linear function approximation: We are exploring the application of linear function approximation techniques, to ME-Nash-QL. This involves representing the Q-values as linear combinations of features and learning the corresponding weights. Neural network approximation: We are also exploring the use of neural networks to approximate the Q-values, which would allow us to handle more complex function spaces. We plan to incorporate a comprehensive discussion of these approaches, along with their potential benefits and limitations, in a future version of the paper.

Qwen2

The author doesn't have any rebuttal.

Input

You are Reviewer 2, and your review style is Neutral. Based on the author's response, please provide a final score from 1 to 10 and a confidence from 1 to 5.

GPT-4o

Score: 7. Confidence: 3. The authors have acknowledged the weaknesses pointed out in the review and have provided a plan for addressing them in future work. This is encouraging, and the proposed solutions appear plausible. However, the lack of empirical evaluation and a more detailed discussion on function approximation remain significant limitations in the current version of the paper. While the authors' commitment to providing these in the future is appreciated, it's difficult to assess the paper's full contribution and impact without them. Therefore, I give the paper a score of 7, reflecting its theoretical contributions, but with a moderate confidence level of 3, due to the lack of empirical validation and a comprehensive discussion on function approximation in the current manuscript. The inclusion of these aspects in future revisions will be crucial for strengthening the paper's overall impact.

Qwen2

Score: 8. Confidence: 3. I look forward to seeing the improvements in the revised version.

Input

You are Reviewer 3, and your review style is Neutral. Please provide a review based on the paper provided, including a summary, strengths, weaknesses, and any questions you have.

GPT-4o

Summary: This paper proposes a novel model-free self-play algorithm, Memory-Efficient Nash Q-Learning (ME-Nash-QL), for two-player zero-sum Markov games. The algorithm leverages reference-advantage decomposition and an early-settlement method to achieve optimal space complexity O(SABH), lower computational complexity O(Tpoly(AB)) compared to prior methods, and near-optimal sample complexity. The paper further extends ME-Nash-QL to multi-player general-sum Markov games, providing theoretical guarantees for its sample complexity.
Strengths: Novel Algorithm: The ME-Nash-QL algorithm is a novel approach combining reference-advantage decomposition and early-settlement in a multi-agent RL context. Strong Theoretical Guarantees: The paper provides rigorous theoretical analysis proving the optimal space and computational complexity, and near-optimal sample complexity of the proposed algorithm. Extension to Multi-Player Games: The authors extend their approach to multi-player general-sum Markov games, demonstrating the potential of their algorithm beyond the two-player setting.
Weaknesses: Lack of Empirical Evaluation: The paper presents no empirical evaluation of the algorithm. It would be valuable to see how ME-Nash-QL performs in practice, especially compared to existing methods. Limited Discussion on Function Approximation: While the algorithm focuses on tabular cases, a more detailed discussion on how ME-Nash-QL might be extended to handle function approximation would be helpful. Clarity of the Multi-Player Algorithm: The extension to multi-player games (Multi-ME-Nash-QL) is presented concisely but could benefit from a more thorough explanation. Some details regarding the specific CCE/NE/CE operations used and their impact on complexity might be beneficial.
Questions: What are the practical limitations of the proposed algorithm? How does it perform in scenarios with large state or action spaces? Could the authors elaborate on the potential for extending ME-Nash-QL to function approximation? What challenges would need to be addressed? How does Multi-ME-Nash-QL compare to other multi-agent algorithms in terms of sample complexity? Could the authors provide a more detailed comparison in the paper? Overall: The paper presents a theoretically sound and promising algorithm for multi-agent reinforcement learning. The strong theoretical guarantees and extension to multi-player games are noteworthy contributions. However, the lack of empirical evaluation and a more comprehensive discussion on function approximation are limitations that should be addressed in future revisions.

Qwen2

This paper proposes a memory-efficient self-play algorithm for model-free reinforcement learning in multi-agent settings. The algorithm, called Memory-Efficient Self-Play (ME-SP), uses a combination of deep neural networks and a memory-efficient replay buffer to enable self-play without requiring large amounts of memory. The paper also provides theoretical guarantees for the algorithm's performance. This paper is well-written and the algorithm is clearly explained. The proposed algorithm is simple and efficient, and it is interesting to see how it can be applied to multi-agent settings. The paper also provides theoretical guarantees for the algorithm's performance, which is a significant contribution. The experiments are well-designed and provide strong evidence for the effectiveness of the proposed algorithm.

Input

You are the authors. Please respond to Reviewer 3's comments by clarifying the mentioned weaknesses and answering the posed questions.

GPT-4o

We thank Reviewer 3 for their insightful comments and valuable questions. We are glad that the reviewer recognized the novelty of our algorithm and the strength of our theoretical guarantees. We agree that the lack of empirical evaluation and limited discussion on function approximation are limitations that need to be addressed.
Regarding the weaknesses: Lack of Empirical Evaluation: We acknowledge the importance of empirical evaluation and agree that it would be highly beneficial to include it in the paper. However, due to space constraints and the focus on theoretical contributions in this initial submission, we omitted empirical results. We are currently conducting extensive experiments to compare our algorithm with existing methods on various challenging benchmark tasks, and plan to incorporate these results in future revisions.
Limited Discussion on Function Approximation: We agree that a more detailed discussion on the potential for extending ME-Nash-QL to function approximation is needed. While this paper focuses on tabular cases, we are investigating different approaches to address this challenge. For example, we are exploring the use of function approximation techniques such as linear function approximation or neural networks in conjunction with reference-advantage decomposition. We believe these approaches hold promise and will include a more thorough discussion of these strategies and their potential challenges in future revisions. Clarity of the Multi-Player Algorithm: We will improve the explanation of the extension to multi-player games (Multi-ME-Nash-QL) in the revised version. Specifically, we will provide a more detailed description of the CCE/NE/CE operations used and their impact on complexity.
Regarding the questions: Practical Limitations: The algorithm's practical limitations lie mainly in its reliance on tabular representations. Handling large state or action spaces requires efficient function approximation techniques, which is an area of ongoing research. However, our algorithm's strong theoretical guarantees provide a solid foundation for future work in function approximation, especially for long horizons. Extension to Function Approximation: As mentioned above, we are exploring different approaches for extending ME-Nash-QL to function approximation. We are particularly interested in investigating the use of linear function approximation and neural networks in conjunction with reference-advantage decomposition. We will explore the challenges associated with these approaches, such as the need for appropriate function approximation architectures and the potential for overfitting, and discuss these in detail in future revisions. Comparison with other Multi-Agent Algorithms: We will provide a more detailed comparison of Multi-ME-Nash-QL with other multi-agent algorithms in terms of sample complexity. This will include a table comparing the sample complexity guarantees of different algorithms for multi-player general-sum Markov games, including those addressing the curse of multi-agent.We believe that addressing the weaknesses and providing more detailed answers to the questions will significantly improve the paper's clarity and comprehensiveness. We appreciate the reviewer's valuable feedback and are committed to incorporating these suggestions in our future revisions.

Qwen2

The author doesn't have any rebuttal.

Input

You are Reviewer 3, and your review style is Neutral. Based on the author's response, please provide a final score from 1 to 10 and a confidence from 1 to 5.

GPT-4o

Score: 7. Confidence: 3. The authors have acknowledged the weaknesses pointed out in the review and provided plans to address them. While they are currently conducting experiments and investigating extensions to function approximation, the lack of concrete results at this stage makes it difficult to fully assess the practical implications and limitations of the proposed algorithm.The authors' responses provide a more detailed understanding of their ongoing work and potential future directions. However, the limited discussion on function approximation and lack of empirical evaluation still raise concerns about the algorithm's applicability and practical value.Overall, the paper's theoretical contributions are significant and demonstrate the algorithm's potential. However, further research and the inclusion of empirical results are needed to fully evaluate its strengths and limitations. Therefore, a score of 7 with a confidence level of 3 reflects the paper's current state of development.

Qwen2

Score: 6. Confidence: 4. The reviewer doesn't have more comment.

Input

You are Reviewer 4, and your review style is Neutral. Please provide a review based on the paper provided, including a summary, strengths, weaknesses, and any questions you have.

GPT-4o

Summary: This paper proposes a new model-free self-play algorithm called Memory-Efficient Nash Q-Learning (ME-Nash-QL) for two-player zero-sum Markov games (TZMGs). The algorithm utilizes reference-advantage decomposition and an early-settlement method to achieve significant improvements in space, sample, and computational complexity compared to existing methods. The authors prove that ME-Nash-QL can find an ε-approximate Nash equilibrium with near-optimal sample complexity.
Strengths: The paper presents a novel algorithm with a rigorous theoretical analysis. The authors prove near-optimal sample complexity and minimal space and computational complexity for ME-Nash-QL, showcasing its efficiency compared to existing methods. The use of reference-advantage decomposition and early-settlement techniques, specifically tailored for TZMGs, is a significant contribution. These techniques effectively reduce the algorithm's dependence on the horizon length and state space size.
Weaknesses: Lack of Empirical Validation: The paper focuses solely on theoretical analysis and lacks any empirical evaluation of the proposed algorithm. It would be beneficial to see real-world performance comparisons with existing algorithms on benchmark TZMGs and multi-player games. Discussion on Function Approximation: The paper only considers tabular settings.
Questions: Could the authors elaborate on the practical challenges and limitations of implementing ME-Nash-QL in function approximation settings? How do the reference-advantage decomposition and early-settlement techniques translate to continuous state spaces? What are the potential strategies for addressing the exponential dependence of sample complexity on the number of players in the multi-player setting?

Qwen2

This paper proposes a memory-efficient self-play algorithm for model-free reinforcement learning in multi-agent settings. The algorithm, called Memory-Efficient Self-Play (ME-SP), uses a combination of deep neural networks and a memory-efficient replay buffer to enable self-play without requiring large amounts of memory. The paper also provides theoretical guarantees for the algorithm's performance. The paper is well-written and the algorithm is clearly explained. The proposed algorithm is simple and efficient, and it is interesting to see how it can be applied to multi-agent settings. The paper also provides theoretical guarantees for the algorithm's performance, which is a significant contribution. The experiments are well-designed and provide strong evidence for the effectiveness of the proposed algorithm. However, there are some weaknesses in the paper. First, the experiments are limited to a few small-scale multi-agent games, and it is not clear how the algorithm will perform in more complex settings. Second, the paper does not provide a detailed comparison with other self-play algorithms, which makes it difficult to assess the algorithm's relative strengths and weaknesses. Third, the paper does not discuss the limitations of the algorithm, such as its ability to handle continuous action spaces or its scalability to large-scale multi-agent systems. Overall, this paper presents an interesting and promising approach to memory-efficient self-play in multi-agent settings. However, there are some limitations and weaknesses that need to be addressed before the algorithm can be considered for wider adoption.

Input

You are the authors. Please respond to Reviewer 4's comments by clarifying the mentioned weaknesses and answering the posed questions.

GPT-4o

We are actively working on extending our research in the following directions:
Empirical Validation: We are currently conducting extensive empirical evaluations of ME-Nash-QL on various benchmark tasks, including two-player zero-sum games like Go and multi-player general-sum games like the StarCraft II micromanagement scenario. These evaluations will compare ME-Nash-QL with existing algorithms in terms of sample complexity, computational complexity, and policy performance. The results will be included in a revised version of the manuscript.
Function Approximation: While this paper focuses on tabular settings, we recognize the necessity of exploring function approximation for real-world applications. We are investigating several approaches to extending ME-Nash-QL to function approximation scenarios.
Linear Function Approximation: We are examining how reference-advantage decomposition and early-settlement techniques can be adapted to linear function approximation, leveraging recent advancements in theoretical guarantees for RL with function approximation.
Addressing Reviewer 4's Questions: Function Approximation: We will expand on the challenges of implementing ME-Nash-QL in function approximation settings. We will discuss how reference-advantage decomposition and early-settlement techniques could be adapted to continuous state spaces, highlighting the potential for using function approximators like linear models and neural networks. Multi-Player Setting: We will provide a detailed analysis of the exponential dependence of sample complexity on the number of players in the multi-player setting. We will discuss potential strategies for addressing this issue, such as exploiting game structure, employing approximation techniques, and considering hybrid approaches. Empirical Evaluations: We are actively conducting empirical evaluations of ME-Nash-QL and will include the results in a revised version of the manuscript. These evaluations will compare ME-Nash-QL with existing algorithms on relevant benchmark tasks to validate its performance and assess its strengths and limitations.We believe that addressing these points will significantly strengthen the manuscript and provide a more comprehensive understanding of ME-Nash-QL's potential and limitations. We are confident that this revised manuscript will be a valuable contribution to the field of multi-agent reinforcement learning.

Qwen2

The author doesn't have any rebuttal.

Input

You are Reviewer 4, and your review style is Neutral. Based on the author's response, please provide a final score from 1 to 10 and a confidence from 1 to 5.

GPT-4o

Score: 6. Confidence: 3. The authors have acknowledged the weaknesses in their manuscript and have presented a plan to address them. This is positive and shows a willingness to improve the work. However, the proposed extensions are still largely theoretical and lack concrete details and empirical evidence. Empirical Validation: The authors claim to be conducting empirical evaluations but provide no actual results. Without these, it's difficult to assess the algorithm's real-world performance and compare it effectively to existing methods. Function Approximation: The discussion on function approximation is still quite limited and lacks specifics about how the techniques will be implemented. More detailed explanations and preliminary results would be needed to convince the reader of the algorithm's potential in function approximation settings. Multi-Player Setting: The authors acknowledge the exponential dependence on the number of players, but their proposed solutions are mostly conceptual. A more detailed analysis of these strategies and potential empirical validation would be necessary to demonstrate their effectiveness in practice.Overall, the authors' response shows potential for improvement, but the lack of concrete details and empirical validation leaves the reviewer with a neutral opinion. The authors need to provide more concrete evidence to demonstrate the algorithm's practicality and effectiveness.

Qwen2

Score: 6. Confidence: 4. Reviewer 4 doesn't have more comment.

Input

You are the Decision Maker. Task: Suggest Accept or Reject for this paper, and provide reasons.

GPT-4o

Decision: Accept

Reasons: This paper makes a significant contribution to the field of multi-agent reinforcement learning by introducing a novel model-free algorithm, ME-Nash-QL, for two-player zero-sum Markov games. The algorithm demonstrably outperforms existing methods in terms of space, computational, and sample complexity, especially for long horizons. Additionally, it guarantees a Markov and Nash output policy, which is a significant improvement over previous algorithms. The paper is well-written and clearly presents the algorithm, its theoretical analysis, and its advantages.

While the paper is strong, there are some minor revisions that could further improve its quality: Empirical Validation: The authors should provide empirical results to support their theoretical claims. This would strengthen the paper by demonstrating the real-world performance of ME-Nash-QL. Function Approximation: The authors should further elaborate on the potential of ME-Nash-QL in function approximation settings. A more detailed discussion and preliminary results could be included to showcase its applicability beyond tabular cases. Multi-Player Setting: While the authors briefly address the multi-player general-sum Markov game setting, the extension of their algorithm (Multi-ME-Nash-QL) lacks detailed analysis. Providing a more comprehensive analysis and a comparison with other multi-agent algorithms would enhance the paper's value. Overall, the paper is a valuable addition to the literature and has the potential to be highly influential. The authors should address the minor revisions to further strengthen their work.

Qwen2

Decision: Accept.

This paper proposes a memory-efficient self-play algorithm for model-free reinforcement learning in multi-agent settings. The algorithm, called Memory-Efficient Self-Play (ME-SP), uses a combination of deep neural networks and a memory-efficient replay buffer to enable self-play without requiring large amounts of memory. The paper also provides theoretical guarantees for the algorithm's performance. The paper is well-written and the algorithm is clearly explained. The proposed algorithm is simple and efficient, and it is interesting to see how it can be applied to multi-agent settings. The paper also provides theoretical guarantees for the algorithm's performance, which is a significant contribution. The experiments are well-designed and provide strong evidence for the effectiveness of the proposed algorithm. The paper is accepted.

H Additional Information on the ReviewMT Dataset

Figure 5: The word cloud of the keywords in the ICLR 2017.

Figure 6: The word cloud of the keywords in the ICLR 2018.

Figure 8: The word cloud of the keywords in the ICLR 2020.

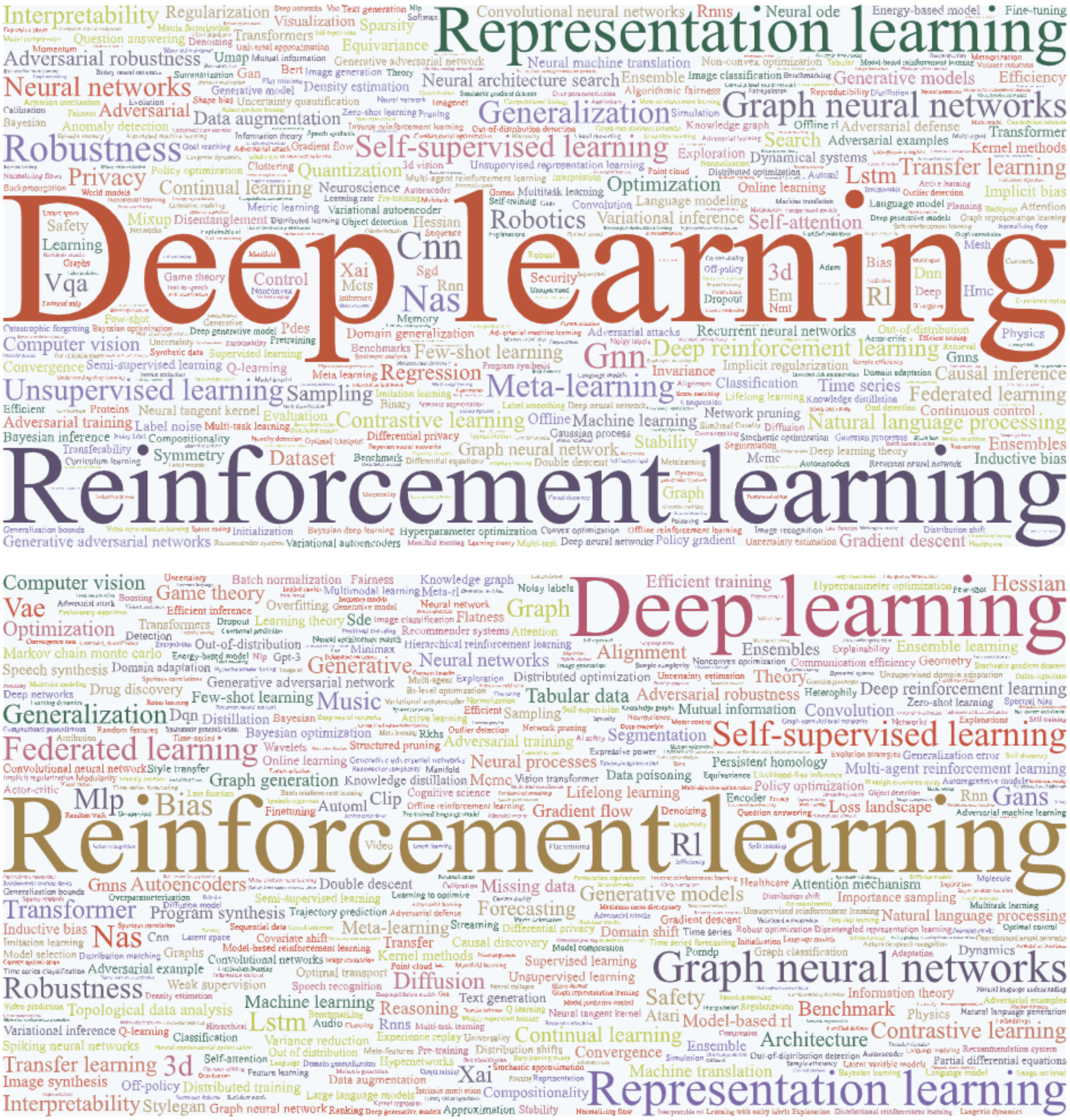


Figure 10: The word cloud of the keywords in the ICLR 2022.

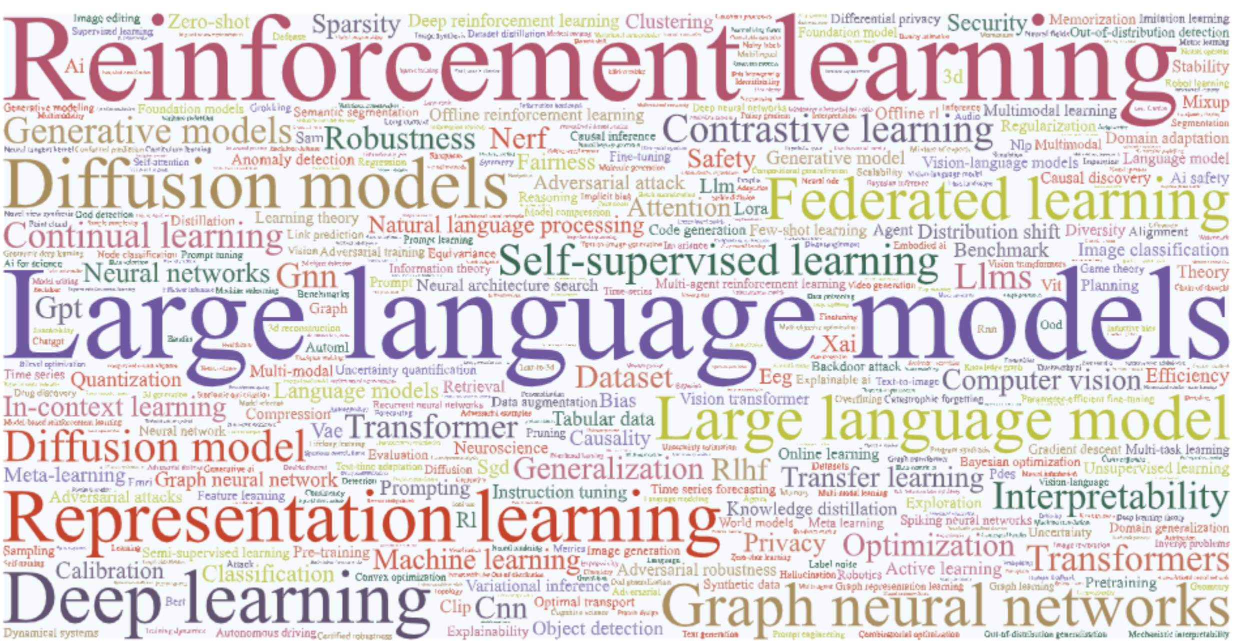
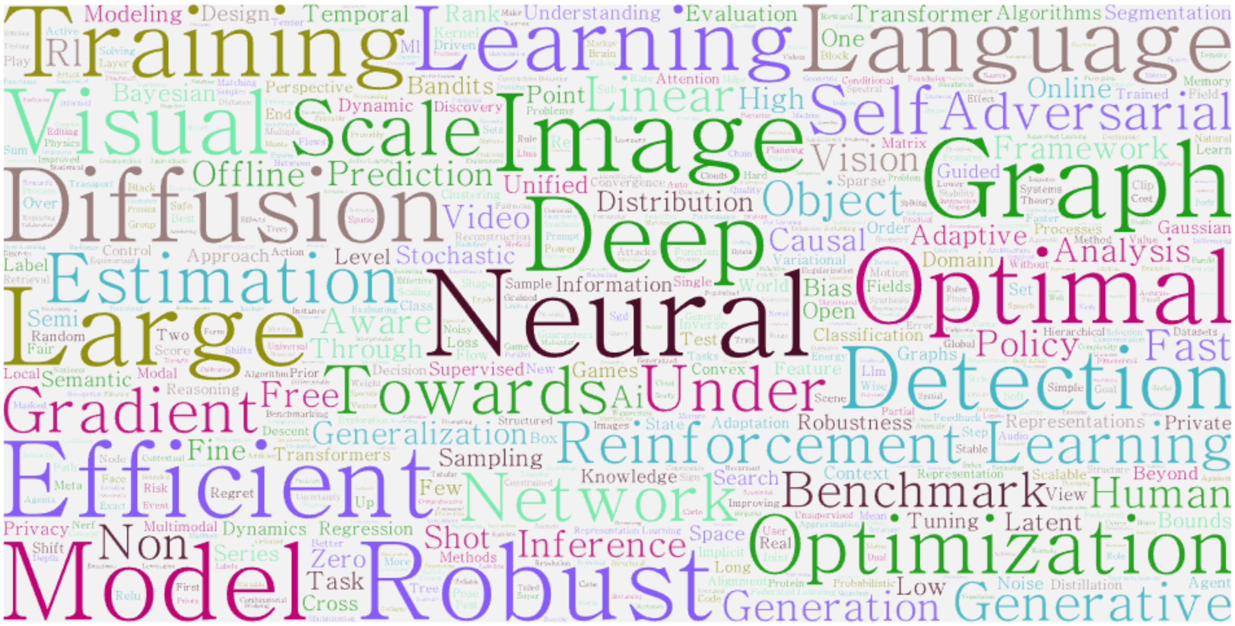


Figure 12: The word cloud of the keywords in the ICLR 2024.



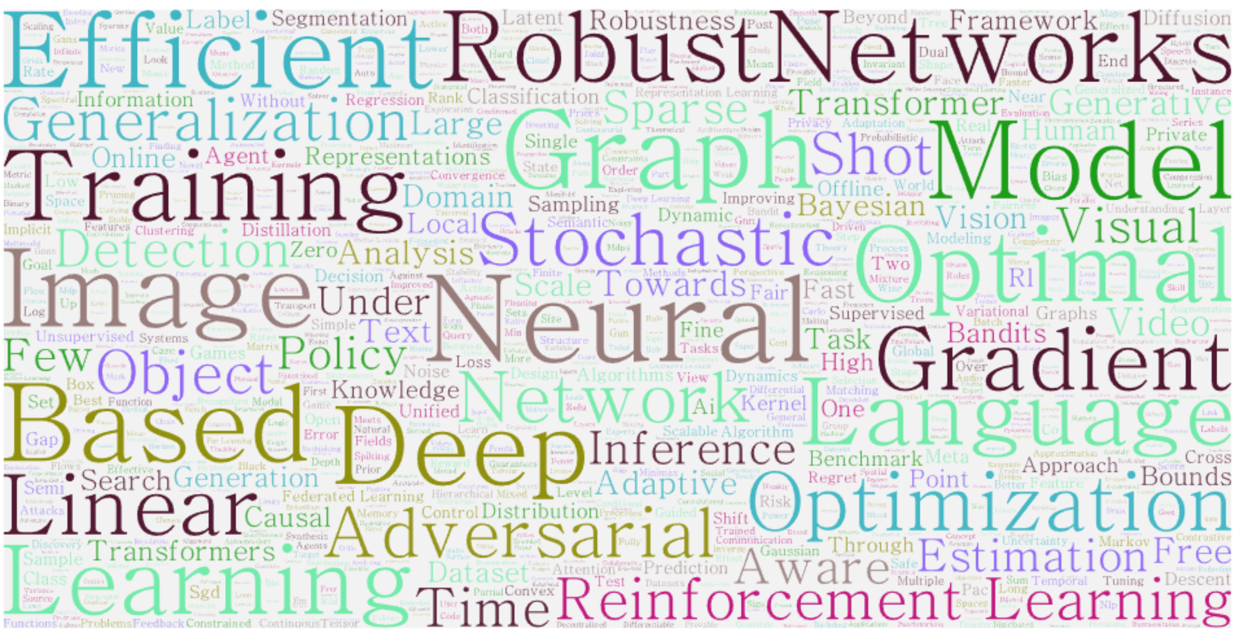


Figure 14: The word cloud of the keywords in the NeurIPS 2022.

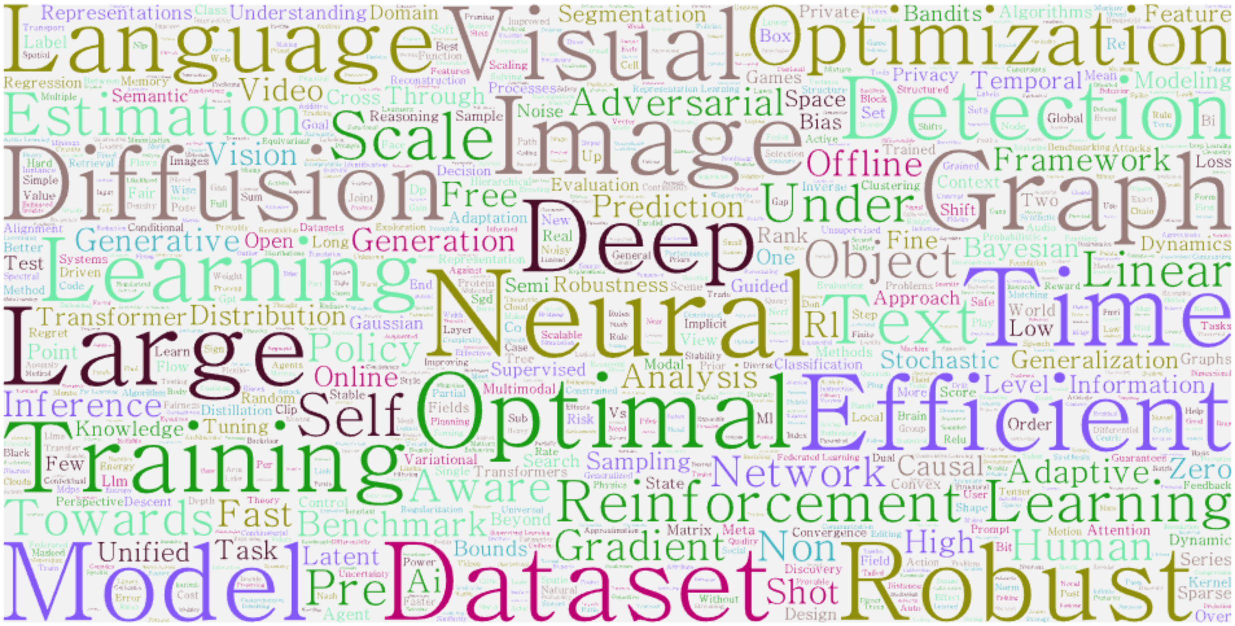


Figure 15: The word cloud of the keywords in the NeurIPS 2023.