OCEAN: OFFLINE CHAIN-OF-THOUGHT EVALUATION AND ALIGNMENT IN LARGE LANGUAGE MODELS

Junda Wu[1*], **Xintong Li**[1*], **Ruoyu Wang**[2], **Yu Xia**[1], **Yuxin Xiong**[1], **Jianing Wang**[3], **Tong Yu**[4], **Xiang Chen**[4], **Branislav Kveton**[4], **Lina Yao**[2,5], **Jingbo Shang**[1], **Julian McAuley**[1]
[1]UC San Diego [2]The University of New South Wales
[3]East China Normal University [4]Adobe Research [5]CSIRO's Data61
{juw069,xil240,yux078,y7xiong,jshang,jmcauley}@ucsd.edu
{ruoyu.wang5}@unsw.edu.au lygwjn@gmail.com
{tyu,xiangche,kveton}@adobe.com lina.yao@data61.csiro.au

ABSTRACT

Offline evaluation of LLMs is crucial in understanding their capacities, though current methods remain underexplored in existing research. In this work, we focus on the offline evaluation of the chain-of-thought capabilities and show how to optimize LLMs based on the proposed evaluation method. To enable offline feedback with rich knowledge and reasoning paths, we use knowledge graphs (KGs) (*e.g.*, Wikidata5M) to provide feedback on the generated chain of thoughts. Due to the heterogeneity between LLM reasoning and KG structures, direct interaction and feedback from knowledge graphs on LLM behavior are challenging, as they require accurate entity linking and grounding of LLM-generated chains of thought in the KG. To address the above challenge, we propose an offline chain-of-thought evaluation framework, OCEAN, which models chain-of-thought reasoning in LLMs as a Markov Decision Process (MDP), and evaluate the policy's alignment with KG preference modeling. To overcome the reasoning heterogeneity and grounding problems, we leverage on-policy KG exploration and reinforcement learning to model a KG policy that generates token-level likelihood distributions for LLM-generated chain-of-thought reasoning paths, simulating KG reasoning preference. Then we incorporate the knowledge-graph feedback on the validity and alignment of the generated reasoning paths into *inverse propensity scores* and propose KG-IPS estimator. Theoretically, we prove the unbiasedness of the proposed KG-IPS estimator and provide a lower bound on its variance. With the off-policy evaluated value function, we can directly enable off-policy optimization to further enhance chain-of-thought alignment. Our empirical study shows that OCEAN can be efficiently optimized for generating chain-of-thought reasoning paths with higher estimated values without affecting LLMs' general abilities in downstream tasks or their internal knowledge.

1 INTRODUCTION

Offline policy evaluation aims to estimate a target policy model's performance with only collected data, without requiring direct interactions between the target policy and realistic environments. Previous offline evaluation methods focus on decision-making policies in recommender systems (Li et al., 2011), healthcare (Bang & Robins, 2005), and other scenarios where online experimentation is costly (Thomas et al., 2015; Bhargava et al., 2024), risky, and impractical (Yu et al., 2021). Recent studies in LLMs leverage human feedback to align models' behaviors with human preferences in single-turn generation (Ouyang et al., 2022; Rafailov et al., 2024) and multi-step reasoning tasks (Joshi et al., 2024). In addition, complicated LLM agentic frameworks, involving multi-agent collaboration, orchestration, and cooperation, rely heavily on efficient (Roucher et al., 2025; Wu et al., 2023a), robust (Masterman et al., 2024; Nguyen et al., 2024a), and proactive (Yao et al., 2023; Xia et al., 2025; Ma et al., 2023) chain-of-thought reasoning abilities, which needs to be finetuned offline

[*]These authors contributed equally to this work.

(Putta et al., 2024) before deploying them online. Due to the high cost of deploying LLMs online and interacting with human feedback, Bhargava et al. (2024) further enables offline evaluation of LLMs from logged human feedback to align LLMs' response generation.

However, considering annotators may not be comprehensive in various types of knowledge backgrounds, human feedback on chain-of-thought reasoning (Joshi et al., 2024) can be more challenging to collect. In addition, since chain-of-thought reasoning involves a sequential decision-making process, the volume of collected human feedback can be exponentially increased. Due to such challenges, conventional reinforcement learning from human feedback (RLHF) methods (Ouyang et al., 2022; Bai et al., 2022a) can suffer from training inefficiencies and scalability issues.

Motivated by recent works in using knowledge graphs (KGs) as side information to enable prompt engineering (Wang et al., 2024c; Xia et al., 2024b), self-correction (Zhao et al., 2023; Wang et al., 2023; Li et al., 2024b; Wu et al., 2024b), evaluating chain-of-thought (Nguyen et al., 2024b), and model fine-tuning (Wang et al., 2024b; Tang et al., 2024), we propose leveraging KGs as weak yet controllable knowledge reasoners to effectively measure the alignment between LLMs' multi-step chain-of-thought reasoning and multi-hop KG trajectories by *inverse propensity scores* (IPS) (Joachims et al., 2017). In contrast to the existing chain-of-thought evaluation (Nguyen et al., 2024b) method, which relies on accurate chain-of-thought grounding on specific KG, we propose verbalizing the KG trajectories and developing a KG policy that serves as a verbal reasoning mechanism over the graphs. Therefore, we can bridge the heterogeneity between KG and LLM reasoning forms, and the verbalized KG policy can be generalized to be compatible with various LLMs.

To enable controllable chain-of-thought alignment in LLMs, we principally track LLMs' decision-making process in generating chain-of-thought reasoning steps, by formulating the process as a Markov Decision Process (MDP) whose goal is to reach the correct final answer with minimal knowledge exploration and exploitation Lissandrini et al. (2020b;a); Wu et al. (2024a). Then, we propose offline chain-of-thought evaluation and alignment, `OCEAN`, which evaluates the generated chain of thoughts from off-policy LLMs through collected offline data samples with feedback from a KG. The improved Knowledge Graph - Inverse Propensity Scores (KG-IPS) approach considers the effects of feedback from the KG policy that aligns the model's chain-of-thought generation and the behavior policy, which prevents model degeneration. We prove that the KG-IPS estimator provides an unbiased estimate of the target policy, with a lower bound for the variance, and establish confidence intervals using sub-Gaussian concentration inequalities. To enable direct optimization of LLM policies, we leverage the proposed KG-IPS policy evaluation approach for LLM fine-tuning by directly maximize estimated policy values through gradient descent. Then we empirically evaluate the optimized LLM policy on three types of chain-of-thought reasoning tasks, and demonstrate the effectiveness of the proposed policy optimization method. We also observe relative performance improvements across evaluation tasks, without affecting LLMs' generalizability or generation quality. We summarize our contributions as follows:

- We propose an offline evaluation framework, `OCEAN`, which bridges the heterogeneity between LLM and KG reasoning, for effective evaluations of chain-of-thought.
- With the evaluation framework, we further develop a direct policy optimization method which enables efficient alignment with automatic feedback from the KG.
- To facilitate the evaluation and optimization, we model the KG preference and derive feedback by developing a policy which verbalizes KG trajectories.
- We provide a theoretical analysis of the unbiasedness and establish a lower bound for the variance of our KG-IPS estimator.
- Through comprehensive experiments, we demonstrate `OCEAN`'s effectiveness in aligning LLMs' chain-of-thought reasoning through direct optimization of the estimated policy value. `OCEAN` also achieves better performance on various downstream tasks without affecting LLMs' generalizability.

2 Related Work

Offline Policy Evaluation Offline policy evaluation (OPE) is essential when online deploying learned policies is risky and impractical (Levine et al., 2020). OPE has been applied to various practical applications, including evaluating the recommender system's behavior with offline collected

user feedback (Gilotte et al., 2018; Jeunen, 2019). Recent work (Gao et al., 2024) also develops an OPE estimator for LLM evaluation based on human feedback. Different from previous works, we study and formulate chain-of-thought generation in LLM as an MDP and use knowledge graph reasoning as automatic feedback to develop a KG-IPS policy value estimator.

LLM Alignment Reinforcement Learning from Human Feedback (RLHF) has been the dominant approach, optimizing LLMs using human-annotated data to align model behavior with user preferences (Ouyang et al., 2022; Bai et al., 2022a). DPO (Rafailov et al., 2024) and RRHF (Yuan et al., 2023) are proposed to reduce the training instability of RLHF. Wu et al. (2023b) utilizes varying densities of human feedback to offer fine-grained rewards for RL finetuning, and Sun et al. (2024a) focuses on aligning LLMs with reward models driven by human-defined principles. To address RLHF's limitations such as heavy reliance on human input, alternative approaches like Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022b; Lee et al., 2023; Liu et al., 2023) and self-alignment methods (Sun et al., 2024b) have been proposed, using AI-generated feedback to scale and automate alignment. Despite advancements, a key challenge remains in aligning LLMs' internal knowledge with their reasoning, resulting in flawed reasoning even after factual errors are corrected. Our approach focuses on improving chain-of-thought alignment by modeling reasoning paths as an MDP and using KGs to ensure both factual accuracy and human-like reasoning.

Chain-of-thought Reasoning Chain-of-thought prompting has been widely applied to elicit the strong reasoning abilities of LLMs (Wei et al., 2022; Chu et al., 2023; Xia et al., 2024a). By decomposing a complex problem into a sequence of intermediate sub-tasks, LLMs are able to focus on important details and solve the step-by-step (Huang & Chang, 2023; Yu et al., 2023). Despite the remarkable performance improvements, recent studies have found that LLMs often generate unfaithful chain-of-thought reasoning paths that contain factually incorrect rationales (Turpin et al., 2023; Lanham et al., 2023). To address this, a number of works leverage LLMs' self-evaluation abilities to verify and refine each reasoning step (Ling et al., 2023; Madaan et al., 2023). As the factual errors in the generated chain-of-thought may also be caused by the limited or outdated parametric knowledge of LLMs, recent methods incorporate external knowledge sources to further edit unfaithful content in the reasoning path (Zhao et al., 2023; Wang et al., 2023; Li et al., 2024b; Wang et al., 2024d;a). While these methods focus on knowledge augmentation and editing through prompts, our method, in comparison, directly aligns LLM internal knowledge with faithful and factual chain-of-thought, which avoids potential knowledge conflicts between parametric and non-parametric knowledge when generating reasoning paths.

3 PRELIMINARY

We first provide the formulation of chain-of-thought reasoning in LLMs as an MDP. Then we discuss conventional knowledge graph reasoning, as an alternative to free-form generation by verbalizing structured knowledge graph reasoning paths into natural language, which is more statistically controllable and generates faithful reasoning paths to the knowledge graph.

3.1 PROBLEM FORMULATION: CHAIN-OF-THOUGHT AS AN MDP

Given the prompt instruction q, chain-of-thought reasoning process in a causal language model π_θ includes the generation of a trajectory of reasoning steps $\boldsymbol{c} = (c_1, c_2, \ldots, c_T)$, before the final answer prediction y,

$$c_t \sim \pi_\theta(\cdot|q) = \prod_{i=1}^{t-1} \pi_\theta(c_i|q, c_{<i}), \quad y \sim \pi_\theta(\cdot|q) = \pi_\theta(y|q, \boldsymbol{c}) \prod_{i=1}^{T} \pi_\theta(c_i|q, c_{<i}),$$

where each reasoning step c_t comprises a sequence of tokens and the number of reasoning step T is determined by the model's generation. Controllable chain-of-thought generation can be challenging due to its nature in autoregressive sequential sampling (Lin et al., 2020), which produces a high-dimensional action space in sampling a reasoning step $\pi_\theta(c_t|q)$ containing multiple tokens.

Chain-of-thought reasoning can be viewed as a Markov Decision Process (MDP) (Sutton, 2018): starting with the instruction prompt q, the LLM sequentially decides and generates the next-step reasoning path c_t that navigates until it arrives at a target final answer y. Given the LLM policy π_θ, at time step t, each **state** $s_t \in \mathcal{S}$ comprises of the instruction prompt q and previously generated

reasoning paths $(c_i)_{i=0}^{t-1}$. The **action** space $\{1, \ldots, |\mathcal{V}|\}^{N_t}$ in LLMs is a sequence of N_t tokens sampled from an identical and finite vocabulary set $\mathcal{V}$. The LLM policy π_θ samples next-step thought based on current state as $a_t \sim \pi_\theta(a_t|s_t)$, which is a sub-sequence in the reasoning path $a_t \in c_t$ and the surrounding context $c_t \backslash a_t$ is deterministically generated by LLMs. The **transition** in chain-of-thought is concatenating each reasoning path to the current state as $s_{t+1} = [s_t, c_t]$. Then the **reward** function is to evaluate each thought given the state as $r_t = r(s_t, c_t)$. Although such formulation of chain-of-thought enables direct LLM on-policy optimization via reinforcement learning, direct interaction with knowledge graphs to collect per-step reward in LLMs can be practically challenging and require a large effort of engineering due to the discrepancy between the unstructured generation of LLMs and structured knowledge graphs (Pan et al., 2024). Therefore, we propose to offline evaluate and optimize the target policy aligning with knowledge graph preference.

3.2 Verbalized Knowledge Graph Reasoning

In contrast to chain-of-thought reasoning, conventional knowledge graph reasoning methods (Lin et al., 2018; Saxena et al., 2020) sample a entity-relation pair (r_t, e_t) at step t from a subset of the graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ consisting of the outgoing edges of current entity e_{t-1},

$$(r_t, e_t) \in \{(r', e')|(e_{t-1}, r', e') \in \mathcal{G}\}, \tag{1}$$

where the transition feasibility of the entity e_{t-1} to all the outgoing edges is entirely determined by $\mathcal{G}$. Knowledge graph reasoning starts with a decomposed triple (e_0, r_1, e_1) of the instruction q, and produces a chain of triplets $\boldsymbol{h} = (e_0, r_1, e_1, \ldots, r_T, e_T)$ by sampling from a policy μ,

$$(r_t, e_t) \sim \mu\left((r_t, e_t)|e_0, r_1, e_1, \ldots, r_{t-1}, e_{t-1}\right), \tag{2}$$

where the goal of such knowledge graph exploration is to arrive at the correct answer entity at the end of the search step T. By knowledge graph exploration, we can collect a set of trajectories $\mathbb{H} = \{\boldsymbol{h}_k\}_{k=1}^{K}$, which are used to estimate a parametric probabilistic policy μ_ϕ as a proxy to model the preference of the knowledge graph.

To align the action space between the knowledge graph preference policy μ_ϕ and the target policy π_θ, we leverage a small language model as the backbone of μ_ϕ and fine-tune the model on verbalized trajectories as natural language contexts. Inspired by existing efforts in verbalizing structured knowledge graphs into natural language query (Seyler et al., 2017) and context (Agarwal et al., 2020; Wang et al., 2022a), we leverage the GPT-4 (Achiam et al., 2023) model f to verbalize each chain of triplets $\boldsymbol{h}$ into a chain-of-thoughts $\boldsymbol{c} = f(\boldsymbol{h})$. The verbalized knowledge-graph trajectories are used to model knowledge graph preference in Section 4.2.

4 OCEAN: Offline Chain-of-thought Evaluation and AlignmeNt

We propose an offline evaluation of the chain-of-thought generation process aligned with knowledge graph preference. The off-policy estimator can be used for policy optimization that aligns LLMs with more faithful reasoning paths from knowledge graphs (Lin et al., 2023). We develop a small language model as a behavior policy that models the knowledge graph preference. In Figure 1, we illustrate the workflow of our proposed framework OCEAN.

4.1 Offline Evaluation and Optimization

One of the most broadly used offline evaluation approaches is *inverse propensity scores* (Ionides, 2008; Dudík et al., 2011), which has been used for LLM-based offline policy evaluation for various purposes (Bhargava et al., 2024; Dhawan et al., 2024; Wu et al., 2022). Given the offline logged chain-of-thought trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$, where $\tau_i = (s_t^{(i)}, c_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)})_{t=0}^{T_i}$, we propose a KG-IPS estimator considering two-folded weights of entity tokens in the knowledge graph preference policy μ_ϕ and of non-entity tokens in the base LLM policy π_0 ,

$$\hat{V}_{KG-IPS}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\frac{1}{|c_t^{(i)}|}\sum_{v \in c_t^{(i)}}\frac{\pi_\theta(v|s_t^{(i)})}{\lambda(v|s_t^{(i)})}\log \pi_0(v|s_t^{(i)}), \tag{3}$$

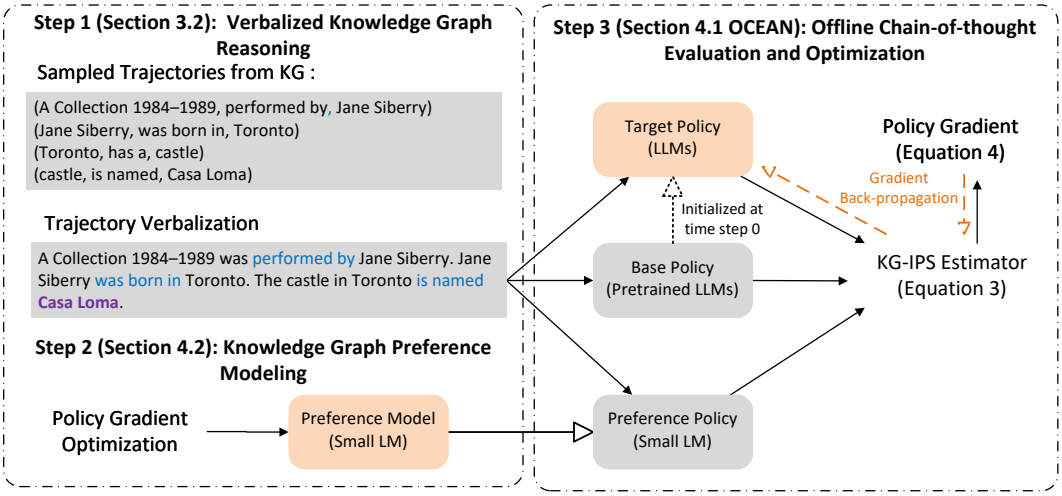


Figure 1: An illustration of our framework OCEAN. We illustrate the framework in three steps. **Step 1** samples trajectories from the Wikidata5M knowledge graph and verbalizes the reasoning trajectories. With the collected trajectories, **Step 2** trains a knowledge graph preference modeling, which is fixed and used during chain-of-thought evaluation and optimization in **Step 3** We highlight the learnable models in red and the frozen models in gray.

where $\lambda(v|s_t^{(i)}) = \mathbf{1}\{v \in a_t^{(i)}\} \cdot \mu_\phi(v|s_t^{(i)}) + \mathbf{1}\{v \in c_t^{(i)} \backslash a_t^{(i)}\} \cdot \pi_0(v|s^{(i)})$. We follow (Zhang et al., 2024) to use the log-likelihood score of each token in the base policy π_0 as the reward function. Establishing the unbiasedness of the KG-IPS estimator is essential for reliable policy evaluation (Jiang & Li, 2016; Bhargava et al., 2024). We formalize this in the following lemma:

Lemma 1. *The KG-IPS estimator provides an unbiased estimate of the target policy π_θ.*

Intuitively, by re-weighting the token-level likelihoods from the target policy π_θ with the behavior policy μ_ϕ (for entity tokens) and the base policy π_0 (for non-entity tokens), we ensure that our estimator compensates for the off-policy distribution, leading to an unbiased estimate of the true value function. The detailed proof is provided in Appendix A.

The standard IPS estimator is known to have high variances (Metelli et al., 2018) considering large behavior discrepancies ($\pi_\theta(v|s_t^{(i)})/\mu_\phi(v|s_t^{(i)})$) between the behavior policy μ_ϕ and the target policy π_θ. In addition, by separately weighting the entity and non-entity tokens with μ_ϕ and π_0 respectively, we avoid the increasing variance accumulated from the long chain-of-thought reasoning process and maintain the LLM's behaviors on non-entity tokens without model degeneration. To further formalize our approach and illustrate the variance inherent in the KG-IPS estimator, we present the following Lemma, which provides a lower bound on the variance,

Lemma 2. *The variance of the KG-IPS estimator is lower bounded by $\Omega(\frac{M^2}{n})$, where M denotes the maximum value of the weighted terms, and n is the number of samples. Given $V(\theta)$ is the true value function of the target policy π_θ, applying the concentration inequality for sub-Gaussian variables, the KG-IPS estimator satisfies the following confidence interval with probability at least $1-\delta$:*

$$\left|\hat{V}_{KG\text{-}IPS}(\theta) - V(\theta)\right| \leq O\left(M\sqrt{\log(1/\delta)/n}\right).$$

A detailed analysis of the variance and confidence interval can be found in Appendix B.

To further support our findings, we demonstrate that the optimal policy for the final reward is consistent with the optimal policy for the entity-based knowledge graph reward, which means the non-entity-based LLM reward can be considered as a regularization term that does not affect the optimal policy. See Appendix C for a complete analysis. In the end, we could directly optimize the target policy by maximizing the estimated value function through policy gradient,

$$\theta \leftarrow \theta + \nabla_\theta \hat{V}_{KG-IPS}(\theta). \tag{4}$$

4.2 KNOWLEDGE GRAPH PREFERENCE MODELING

To facilitate the evaluation and optimization, we model knowledge graph preference and derive feedback by developing the behavior policy μ_ϕ which verbalizes knowledge-graph trajectories. Randomly sampled trajectories $\mathbb{H}$ from $\mathcal{G}$ in Section 3.2 contain samples that may not be transformed into a chain of thoughts leading to a reasonable question-answering. Following conventional self-consistent measurement (Wang et al., 2022b; Manakul et al., 2023), given a sampled trajectory $\boldsymbol{h}$ and its verbalized chain of thoughts $\boldsymbol{c}$, we prompt the GPT-4 model to propose a question q related to the first entity $e_0 \in \boldsymbol{h}$ whose answer should be exactly the last entity $e_T \in \boldsymbol{h}$, and query the GPT-4 model with the proposed question,

$$\hat{q} \sim f(q|e_0, e_T, \boldsymbol{c}), \quad \hat{y} \sim f(y|\hat{q}, \boldsymbol{c}), \quad R(\boldsymbol{h}|\boldsymbol{c}) = \mathrm{E}\left[\mathbf{1}\left\{e_T = \hat{y}\right\}\right],$$

where the reward of the trajectory is determined by the answer accuracy. We estimate the reward function $R(\boldsymbol{h}|\boldsymbol{c})$ as the normalized question-answering accuracy (detailed in Appendix D). Then we fine-tune the preference policy μ_ϕ directly via policy gradient optimization,

$$\nabla_\phi J(\phi) = \nabla_\phi \sum_{k=1}^{K} \sum_{t=0}^{|\boldsymbol{c}_k|-1} R(\boldsymbol{h}_k|\boldsymbol{c}_k) \log \mu_\phi(y_{k,t}|q_k, y_{k,<t}).$$

Based on the distribution of relations (Figure 4b) and entities (Figure 4c) in the sampled knowledge graph trajectories, we observe that the relation distribution is relatively more skewed toward the most frequent relations. This suggests that the verbalized knowledge graph reasoning policy is likely to focus on more frequent reasoning transitions, potentially enhancing its ability to learn meaningful patterns. In contrast, the entity distribution shows a relatively short tail, which may help mitigate the risk of overfitting to specific entities or knowledge biases.

5 EXPERIMENTS

In this section, we evaluate our proposed method, OCEAN, by conducting chain-of-thought alignment on four LLM backbone models and evaluating several downstream tasks. We show our method's effectiveness in chain-of-thought alignment and its generalizability in various tasks to understand (i) whether the proposed optimization approach sufficiently aligns LLMs' chain-of-thought behaviors with higher estimated values on multi-hop question-answering tasks, (ii) how the proposed method performs on knowledge-intensive question-answering tasks and (iii) whether the post-alignment LLM generalizes on commonsense reasoning tasks.

5.1 IMPLEMENTATION DETAILS

Datasets. Following prior work (Zhang et al.), we evaluate our approach across three key aspects of question answering. For *knowledge-intensive reasoning*, we use datasets requiring deeper domain understanding. The **AI2 Reasoning Challenge (ARC)** (Clark et al., 2018) tests models' advanced reasoning abilities with grade-school science questions. **PubMedQA** (Jin et al., 2019) assesses biomedical reasoning from research abstracts. Finally, **SciQA** (Auer et al., 2023) challenges models to reason over scientific knowledge using the Open Research Knowledge Graph (ORKG).

The second focus is *multi-hop reasoning*, where models must combine information from multiple sources. **HotpotQA** (Yang et al., 2018) requires reasoning across multiple Wikipedia documents. **MuSiQue** (Trivedi et al., 2022) includes complex questions needing 2-4 inference hops. **StrategyQA** (Geva et al., 2021) tests implicit reasoning, where steps must be inferred. These datasets assess the models' ability to perform connected reasoning tasks for complex questions.

Lastly, we focus on *commonsense reasoning*, crucial for understanding everyday knowledge. We evaluate three different commonsenseQA benchmarks, **CSQA** (Talmor et al., 2021), **CSQA2** (Saha et al., 2018), and CSQA-COT1000 (Li et al., 2024a). In addition, **OpenBookQA** (Mihaylov et al., 2018) tests models' ability to combine common knowledge with science facts for elementary-level questions. **WinoGrande** (Sakaguchi et al., 2021) challenges models to avoid biases in commonsense reasoning. These tasks evaluate a model's general abilities in commonsense question-answering.

Baselines. We experiment with four backbone LLMs: Gemma-2 (Team, 2024) with 2B model parameters, Llama-3 (AI@Meta, 2024) with 8B model parameters, Phi-3.5-mini (Abdin et al., 2024)

with 3.8B model parameters, and Mistral-0.2 (Jiang et al., 2023) with 7B model parameters. We use the instruction fine-tuned version of backbone LLMs for better instruction following abilities in question-answering. For chain-of-thought alignment in OCEAN, we use the CWQ question-answering dataset (Talmor & Berant, 2018) as the source data, in which the question-answering pairs are developed from knowledge graphs. OCEAN only uses CWQ questions for the LLM to generate chain-of-thought reasoning paths, which are further aligned using the knowledge graph preference model, without directly supervised learning on the ground-truth answers. To compare with direct supervised learning, we also enable instruction-tuning as a baseline (SFT), which is fine-tuned with the question as instruction and the answer as the response.

5.2 Multi-hop Question Answering

We evaluate the chain-of-thought reasoning performance of OCEAN compared with base LLMs and supervised fine-tuning (SFT), in three multi-hop question-answering tasks in Table 1. Comparing SFT and Base LLMs, we observe similar knowledge inconsistency as in knowledge-intensive tasks. Although SFT improves on MuSiQue with Gemma-2 and Mistral-0.2 backbones whose base models' performance is relatively inferior on this task, such knowledge-inconsistent problems result in worse performance on other downstream tasks.

Model	Method	HotpotQA			MuSiQue		StrategyQA		
		w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$
Llama-3	Base	32.78	<u>33.54</u>	-10.35	11.59	-9.90	<u>77.73</u>	59.53	-9.25
	SFT	8.22 (-24.56)	16.49 (-17.05)	-22.28	1.80 (-9.79)	-17.09	66.52 (-11.21)	51.82 (-7.71)	-15.17
	OCEAN	<u>33.38</u> (+0.6)	**33.75** (+0.21)	**-8.10**	11.67 (+0.08)	**-9.77**	75.40 (-2.33)	59.83 (+0.3)	**-5.53**
Gemma-2	Base	26.33	18.58	-31.88	5.84	-26.41	76.71	<u>60.99</u>	-14.06
	SFT	29.75 (+3.42)	15.91 (-2.67)	-46.92	**12.53** (+6.69)	-40.25	64.77 (-11.94)	51.97 (-9.02)	-23.27
	OCEAN	26.20 (-0.13)	19.70 (+1.12)	**-26.43**	6.87 (+1.03)	**-22.15**	74.24 (-2.47)	**66.23** (+5.24)	**-13.52**
Phi-3.5	Base	32.13	26.14	-19.49	<u>11.85</u>	-15.30	73.51	58.37	-13.87
	SFT	21.99 (-10.14)	7.87 (-18.27)	-44.57	6.01 (-5.84)	-42.10	63.03 (-10.48)	50.95 (-7.42)	-21.66
	OCEAN	**35.13** (+3.0)	26.23 (+0.09)	**-14.84**	10.82 (-1.03)	**-13.47**	72.20 (-1.31)	57.64 (-0.73)	**-12.25**
Mistral-0.2	Base	26.82	28.13	-19.08	5.67	-6.40	**79.33**	58.22	-11.36
	SFT	20.88 (-5.94)	14.49 (-13.64)	-18.53	7.73 (+2.06)	-12.24	52.40 (-26.93)	51.53 (-6.69)	-15.39
	OCEAN	27.24 (+0.42)	27.54 (-0.59)	**-3.12**	5.15 (-0.52)	**-5.94**	77.29 (-2.04)	56.62 (-1.6)	**-11.21**

Table 1: Comparison results of OCEAN, base LLMs (Base), and supervised fine-tuning (SFT), on three **Multi-hop Question-answering** tasks. We report with context (**w/ ctx**) and without context (**w/o ctx**) answer results with the Exact Match (EM) metric on **HotpotQA** and the Accuracy metric on **StrategyQA**. Performance on **MuSiQue** dataset is EM with context. We also use each test/validation split for each dataset and report policy evaluation $\hat{V}(\theta)$ results. We highlight the best-performed metric in **bold font** and the second-best <u>underline</u> for each task.

Since OCEAN is aligned to incorporate more knowledge-faithful chain-of-thought reasoning patterns learned from knowledge graph reasoning policy without directly editing its internal knowledge, OCEAN maintains its generalizability in adapting to downstream tasks. We observe that OCEAN consistently improves on the policy estimated value $\hat{V}(\theta)$ through direct policy optimization proposed in equation 4, which demonstrates the effectiveness of the developed optimization method. Regarding the question-answering accuracy, OCEAN improves base LLMs, which achieves the best performance on HotpotQA and StrategyQA without context.

5.3 Knowledge-intensive Question Answering

To understand the effectiveness of OCEAN in knowledge-intensive question-answering tasks, we show performance comparison with base LLMs (Base) and supervised fine-tuning (SFT) in Table 2. Comparing SFT and Base LLMs, we observe that directly aligning knowledge graphs with LLMs may suffer from domain and knowledge inconsistency when downstream tasks require specific domain knowledge, conflicting with the knowledge graph in the fine-tuning stage. We also observe that SFT achieves 4.85% and 0.55% average improvements on the PubMedQA dataset, with and without context respectively, whereas it suffers from 29.60%, 8.35%, 13.6% average performance decreases on the remaining tasks. Such significant discrepancies in SFT's effects across different downstream tasks further show the risk in direct knowledge editing in LLMs.

Model	Method	ARC		PubMedQA			SciQA		
		w/o ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$
Llama-3	Base	79.93	**-10.38**	63.60	58.60	-25.40	83.10	57.10	-22.91
	SFT	61.87 (-18.06)	-18.42	**75.80** (+12.2)	58.00 (-0.6)	-26.03	67.10 (-16.0)	35.80 (-21.3)	-23.84
	OCEAN	80.60 (+0.67)	-12.45	66.00 (+2.4)	**59.80** (+1.2)	**-9.37**	83.20 (+0.1)	57.70 (+0.6)	**-16.63**
Gemma-2	Base	65.89	**-15.36**	34.40	40.60	-24.61	76.50	47.10	**-26.60**
	SFT	18.06 (-47.83)	-25.22	35.60 (+1.2)	21.00 (-19.6)	-26.55	79.80 (+3.3)	51.50 (+4.4)	-36.61
	OCEAN	63.21 (-2.68)	-16.20	44.60 (+10.2)	41.60 (+1.0)	**-18.72**	72.20 (-4.3)	47.50 (+0.4)	-26.77
Phi-3.5	Base	87.29	**-7.86**	70.40	41.80	-28.48	83.50	58.90	-14.46
	SFT	65.22 (-22.07)	-9.02	62.40 (-8.0)	50.20 (+8.4)	-28.40	76.90 (-6.6)	43.80 (-15.1)	-14.62
	OCEAN	**87.63** (+0.34)	-7.94	68.40 (-2.0)	47.60 (+5.8)	**-11.45**	**84.70** (+1.2)	**63.50** (+4.6)	**-13.40**
Mistral-0.2	Base	73.91	**-9.99**	51.60	36.20	-13.01	78.50	58.00	**-11.77**
	SFT	43.48 (-30.43)	-13.99	65.60 (+14.0)	50.20 (+14.0)	-21.87	64.40 (-14.1)	35.50 (-22.5)	-21.86
	OCEAN	68.90 (-5.01)	-10.89	52.60 (+1.0)	33.20 (-3.0)	**-12.42**	79.10 (+0.6)	58.40 (+0.4)	-12.00

Table 2: Comparison results of OCEAN, base LLMs (Base), and supervised fine-tuning (SFT), on three **Knowledge-intensive Question-answering** tasks. We report answers with context (**w/ ctx**) and without context (**w/o ctx**) on Exact Match (EM) metric on **PubMedQA** and **SciQA**. The EM performance on **ARC** dataset is without context. We also use the test/validation split for each dataset to report estimated policy values $\hat{V}(\theta)$. We highlight the best metric in **bold font** for each task.

With the enhancement of OCEAN, question-answering accuracy of knowledge-intensive tasks generally improved, while OCEAN fine-tuned LLMs achieving the best performance on all three datasets, except for PubMedQA without context where SFT achieves better performance due to knowledge transfer from knowledge graph dataset. We also observe consistent policy value improvement on PubMedQA and SciQA, where the original policy values of base LLMs are relatively lower. For tasks like ARC, which does not require additional reference knowledge from context and reasoning in an easier chain of thought, OCEAN still maintains comparable policy value to the base LLM, which demonstrates the robustness and generalizability of the proposed method.

5.4 COMMONSENSE REASONING

Model	Method	CSQA	CSQA-2	CSQA-COT1000	OpenBookQA	Winogrande	Average
Llama-3	Base	**65.03**	**71.39**	69.50	58.80	**43.09**	**61.56**
	SFT	51.19 (-13.84)	57.06 (-14.33)	49.00 (-20.5)	63.20 (+4.4)	34.73 (-8.36)	51.04
	OCEAN	**65.03** (0.0)	68.60 (-2.79)	**72.00** (+2.5)	**60.40** (+1.6)	41.36 (-1.73)	61.48
Gemma-2	Base	57.99	62.57	63.50	51.80	49.64	57.10
	SFT	14.66 (-43.33)	65.80 (+3.23)	15.00 (-48.5)	7.40 (-44.4)	50.04 (+0.4)	30.58
	OCEAN	**67.73** (+9.74)	**63.56** (+0.99)	**72.50** (+9.0)	**57.20** (+5.4)	**50.12** (+0.48)	**62.22**
Phi-3.5	Base	68.55	**64.70**	72.50	**72.40**	**50.51**	**65.73**
	SFT	69.94 (+1.39)	61.47 (-3.23)	72.00 (-0.5)	69.40 (-3.0)	50.51 (0.0)	64.66
	OCEAN	**69.62** (+1.07)	62.77 (-1.93)	**73.50** (+1.0)	71.20 (-1.2)	50.12 (-0.39)	65.44
Mistral-0.2	Base	61.18	68.48	65.00	**64.00**	46.25	60.98
	SFT	35.87 (-25.31)	22.47 (-46.01)	33.00 (-32.0)	34.80 (-29.2)	29.12 (-17.13)	31.05
	OCEAN	**63.80** (+2.62)	**69.19** (+0.71)	**67.00** (+2.0)	62.60 (-1.4)	**46.49** (+0.24)	**61.82**

Table 3: Comparison results of OCEAN, base LLMs (Base), and supervised fine-tuning (SFT), on five **Commonsense Reasoning** tasks. We report the Exact Match (EM) metric on these tasks and the average performance. We highlight the best method in **bold font** for each task and LLM.

Finally, to demonstrate OCEAN's generalizability in preserving commonsense knowledge and preventing knowledge catastrophic forgetting (Luo et al., 2023; Wu et al., 2025), we evaluate OCEAN with base LLMs (Base) and supervised fine-tuning (SFT) on five commonsense reasoning tasks in Table 3. Since such tasks do not require external domain knowledge, we only evaluate the accuracy of the model's generated answers. We observe that directly applying supervised fine-tuning (SFT) using knowledge graphs significantly impacts large language models (LLMs), potentially leading to catastrophic forgetting of commonsense knowledge. especially for the backbone LLMs of Gemma-2 and Mistral-0.2. In contrast, we show that OCEAN achieves robust performance on commonsense reasoning by leveraging off-policy evaluation and optimization from knowledge graph's feedback. OCEAN manages to maintain comparable performance of base LLMs (*e.g.*, Llama-3 and Phi-3.5),

which have strong zero-shot commmonsense reasoning abilities. In addition, we observe that for base LLM with relatively lower performance (*e.g.*, Gemma-2 and Mistral-0.2), OCEAN enables consistent improvements. Therefore, OCEAN serves as a robust off-policy alignment paradigm to incorporating knowledge graph reasoning without affecting the generalizability of pretrained LLMs.

6 Analysis

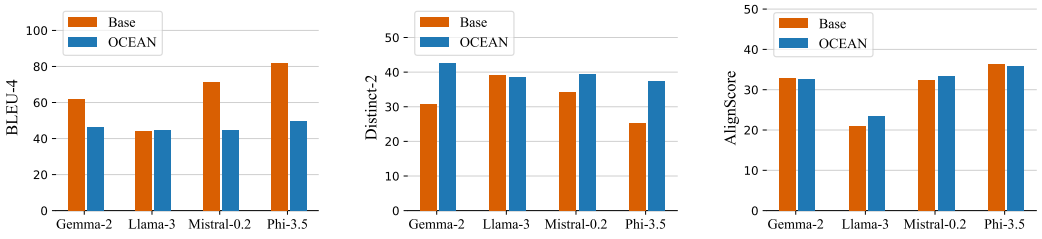
6.1 In-context Learning & Instruction tuning

We conduct further analysis to compare the performance of both the base model and our proposed model in the scenarios of In-Context Learning and instruction fine-tuning. Specifically, we conduct this analysis using the Gemma-2 and Phi-3.5 models across three benchmark datasets: SST2 (Socher et al., 2013) for sentiment classification, AgNews (Zhang et al., 2015) for topic classification, and BoolQ (Clark et al., 2019) for reading comprehension. In the In-context Learning setup, we provide the model with a single example for each task in the prompt. For the Instruction tuning experiments, we apply LoRA (Hu et al., 2021) to the pre-trained model and fine-tune it on each dataset for 10 epochs. Throughout these experiments, the rank parameter in LoRA is fixed at 16, and we set α in LoRA to 32 across all tasks. The results of the In-context Learning and Instruction Tuning are presented in Table 4. Overall, we observe that the performance of the base model and our proposed model is largely comparable across most scenarios, except in the AG News task with Gemma-2, where OCEAN demonstrates greater performance after instruction tuning.

Model	Method	In-Context Learning				Instruction-Tuning			
		SST2	BoolQ	AG News	Avg.	SST2	BoolQ	AG News	Avg.
Gemma-2	Base	87.16	56.12	16.47	53.25	96.21	69.03	47.03	70.76
	OCEAN	89.33	55.72	13.14	52.73	96.56	68.66	60.08	75.10
Phi-3.5	Base	41.28	60.06	31.89	44.41	96.44	68.13	86.43	83.67
	OCEAN	40.48	59.11	32.37	43.98	96.44	68.81	86.24	83.83

Table 4: Performance Comparison of In-Context Learning and Instruction Tuning. All datasets consist of classification tasks or true/false questions, so accuracy is used to evaluate the performance. The performance of the base model and our proposed model is largely comparable.

6.2 Evaluation of Generation Quality Post Alignment



(a) comparison on **Self-BLEU** (b) comparison on **Distinct-2** (c) comparison on **AlignScore**

Figure 2: Comparison results of base LLMs and OCEAN on three evaluation metrics, Self-BLEU, Distinct-2, and AlignScore. Lower Self-BLEU scores and higher Distinct-2 scores indicate better diversity of the generated text, while higher AlignScore indicates better faithfulness.

To further evaluate the generation quality of post-alignment LLMs, we use the *Self-BLEU* (Zhu et al., 2018) and *Distinct-2* (Li et al., 2015) scores to evaluate the diversity of the generation, concerning the similarity between generated texts and the uniqueness of generated 2-gram phrases respectively. In addition, *AlignScore* (Zha et al., 2023) is used to evaluate the faithfulness of the generated answer given the question context. The results are presented in Figure 2, which show that post-alignment LLMs achieve comparable or better performances in terms of generation diversity and faithfulness.

This demonstrates that while OCEAN aligns chain-of-thought reasoning with KGs, we maintain the text generation qualities of LLMs.

6.3 Case Study

In the previous Section 5.2 and 5.3 we observe efficient chain-of-thought alignment with improvement on the estimated policy value $\hat{V}(\theta)$. To further understand the effects of the alignment, we choose backbone LLMs, Gemma-2 and Llama-3, with significant improvements on $\hat{V}(\theta)$, and perform a sample analysis by comparing the outputs of the base model and OCEAN on the same set of questions. Our findings demonstrate that the application of our method enhances the precision and conciseness of the chain of thought in the generated responses. Some illustrative examples are provided in Figure 3. Specifically, in the first example, the base Llama-3 model incorrectly claims that singing is not a primary action associated with playing the guitar, which leads to an erroneous solution to the question. In contrast, our method enables the model to recognize that singing is a common activity when playing the guitar, while also understanding that making music serves as a broader term. In the second example, although both the base model and OCEAN on Gemma-2 provide reasonable answers to the question, our model demonstrates a more concise chain of thought, streamlining the reasoning process and arriving at the solution with greater simplicity.

Q: What do people typically do while playing guitar? Choose the correct answer from the following choices.
(A) cry (B) hear sounds (C) singing (D) arthritis (E) making music

Base Llama-3: Inaccurate Information

1. The question asks what people typically do while playing guitar. ..... 4. Playing the guitar typically involves plucking or strumming the strings to produce sound, which is a form of making music. 5. The other options don't make sense in the context of playing the guitar: ..... (C) Singing is an activity that can be done while playing guitar, but it's not the primary action people do while playing the instrument. ..... Therefore, the correct answer is (E) making music.

Wrong answer

OCEAN Llama-3: Accurate Information

1. The question is asking about what people typically do while playing guitar. ..... 4. Singing is a common activity that people do while playing guitar, as it is a common way to accompany oneself while playing the instrument. 5. Making music is a broad term that encompasses a wide range of activities, including playing the guitar.
Based on the above reasoning, the correct answer is: C) singing

Correct Answer

Q: What is the height above or below sea level called?
Sample Answer: elevation

Base Gemma-2: Redundant chain-of-thought

1. What is sea level? Sea level is the average level of the ocean. 2. What is the height of something above or below sea level? This is the vertical distance of an object from sea level. 3. What is the term for this vertical distance? This is the height above or below sea level. Final Answer: The height above or below sea level is called **altitude**.

OCEAN Gemma-2: A more concise thinking process

Answer: Elevation. Reasoning: 1. Elevation is the height of a place above or below sea level.

Figure 3: Sample comparison between the base model and OCEAN on Llama-3 and Gemma-2. Our method enables a more precise and concise chain of thought.

7 Conclusion

In this work, we propose OCEAN to address the challenge of offline chain-of-thought evaluation and optimization of LLMs. By modeling the knowledge-graph preference and deriving feedback by developing a policy that verbalizes knowledge-graph trajectories, we propose KG-IPS estimator to estimate policy values in the alignment of reasoning paths with knowledge graphs. Theoretically, we proved the unbiasedness of the KG-IPS estimator and provided a lower bound on its variance. Empirically, our framework effectively optimizes chain-of-thought reasoning while maintaining LLMs' general downstream task performance, offering a promising solution for enhancing reasoning capabilities in large language models. Our framework not only enhances chain-of-thought reasoning but can also offer a potential offline evaluation mechanism for agentic frameworks, enabling the safe assessment of autonomous decision-making processes. Future work could explore integrating this approach into reinforcement learning and multi-agent systems to further validate its utility in complex, dynamic environments.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv preprint arXiv:2010.12688*, 2020.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13 (1):7240, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a. URL https://arxiv.org/pdf/2204.05862.pdf.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b. URL https://arxiv.org/pdf/2212.08073.pdf.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Aniruddha Bhargava, Lalit Jain, Branislav Kveton, Ge Liu, and Subhojyoti Mukherjee. Off-policy evaluation from logged human feedback. *arXiv preprint arXiv:2406.10030*, 2024.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. *arXiv preprint arXiv:2407.07018*, 2024.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. Off-policy evaluation for human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 198–206, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Olivier Jeunen. Revisiting offline evaluation for implicit-feedback recommender systems. In *Proceedings of the 13th ACM conference on recommender systems*, pp. 596–600, 2019.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pp. 781–789, 2017.

Nitish Joshi, Koushik Kalyanaraman, Zhiting Hu, Kumar Chellapilla, He He, and Erran Li. Improving multi-hop reasoning in llms by learning from rich human feedback. 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023. URL https://arxiv.org/pdf/2309.00267.pdf.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. Focus on your question! interpreting and mitigating toxic cot problems in commonsense reasoning. *arXiv preprint arXiv:2402.18344*, 2024a.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=cPgh4gWZlz.

Chu-Cheng Lin, Aaron Jaech, Xin Li, Matthew R Gormley, and Jason Eisner. Limitations of autoregressive models and their alternatives. *arXiv preprint arXiv:2010.11939*, 2020.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. Multi-hop knowledge graph reasoning with reward shaping. *arXiv preprint arXiv:1808.10568*, 2018.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*, 2023.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36407–36433. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/72393bd47a35f5b3bee4c609e7bba733-Paper-Conference.pdf.

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. Graph-query suggestions for knowledge graph exploration. In *Proceedings of The Web Conference 2020*, pp. 2549–2555, 2020a.

Matteo Lissandrini, Torben Bach Pedersen, Katja Hose, and Davide Mottin. Knowledge graph exploration: Where are we and where are we going? *ACM SIGWEB Newsletter*, 2020(Summer): 1–8, 2020b.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*, 2023.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Kaixin Ma, Hongming Zhang, Hongwei Wang, Xiaoman Pan, Wenhao Yu, and Dong Yu. Laser: Llm agent with state-space exploration for web navigation. *arXiv preprint arXiv:2309.08172*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *ArXiv*, abs/2404.11584, 2024. URL https://api.semanticscholar.org/CorpusID:269187633.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, et al. Gui agents: A survey. *arXiv preprint arXiv:2412.13501*, 2024a.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. *arXiv preprint arXiv:2402.11199*, 2024b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 'smolagents': a smol library to build great agentic systems. https://github.com/huggingface/smolagents, 2025.

Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4498–4507, 2020.

Dominic Seyler, Mohamed Yahya, and Klaus Berberich. Knowledge questions from knowledge graphs. In *Proceedings of the ACM SIGIR international conference on theory of information retrieval*, pp. 11–18, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. SALMON: Self-alignment with principle-following reward models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=xJbsmB8UMx.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36, 2024b.

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*, 2018.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=qF7FlUT5dxa.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 491–500, 2024.

Gemma Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Jianing Wang, Wenkang Huang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Xiang Li, and Ming Gao. Knowledge prompting in pre-trained language model for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 3164–3177. Association for Computational Linguistics, 2022a. URL https://doi.org/10.18653/v1/2022.emnlp-main.207.

Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. Boosting language models reasoning with chain-of-knowledge prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 4958–4981. Association for Computational Linguistics, 2024a. URL https://doi.org/10.18653/v1/2024.acl-long.271.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. Instructgraph: Boosting large language models via graph-centric instruction tuning and preference alignment. *arXiv preprint arXiv:2402.08785*, 2024b.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19206–19214, 2024c.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*, 2024d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3436–3448, 2022.

Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. Coral: Collaborative retrieval-augmented large language models improve long-tail recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3391–3401, 2024a.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14073–14087, 2024b.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, et al. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023a.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b8c90b65739ae8417e61eadb521f63d5-Paper-Conference.pdf.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *arXiv preprint arXiv:2404.15676*, 2024a.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A Rossi, Haoliang Wang, and Julian McAuley. Knowledge-aware query expansion with large language models for textual and relational retrieval. *arXiv preprint arXiv:2410.13765*, 2024b.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, et al. From selection to generation: A survey of llm-based active learning. *arXiv preprint arXiv:2502.11767*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*, 2023.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EdIGMCHk4l.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

Chen Zhang, Dading Chong, Feng Jiang, Chengguang Tang, Anningzhe Gao, Guohua Tang, and Haizhou Li. Aligning language models using follow-up likelihood as reward signal. *arXiv preprint arXiv:2409.13948*, 2024.

Huiming Zhang and Song Xi Chen. Concentration inequalities for statistical inference. *arXiv preprint arXiv:2011.02258*, 2020.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Zongmeng Zhang, Yufeng Shi, Jinhua Zhu, Wengang Zhou, Xiang Qi, Houqiang Li, et al. Trustworthy alignment of retrieval-augmented large language models via reinforcement learning. In *Forty-first International Conference on Machine Learning*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5823–5840, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.320. URL https://aclanthology.org/2023.acl-long.320.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

A MEAN ANALYSIS

To prove that the KG-IPS estimator is unbiased, we need to demonstrate that the expected value of the IPS estimator equals the true expected reward under π_θ.

Proof. The value function of policy π_θ can be defined as:

$$\begin{aligned} V(\pi_\theta) &= \mathbb{E}_{a_t \sim \pi(\cdot|s_t)} \left[r(s_t, a_t) \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i |c_t^{(i)}|} \sum_{t=1}^{T_i} \mathbb{E}_{e \sim \pi_\theta(\cdot|s_t^{(i)})} \left[r(s_t^{(i)}, e) \right] \end{aligned}$$

where $r(s_t, a_t)$ represent the reward obtained by taking action a_t under state s_t.

Given that our value function consists of two cases: the first case considers the reward derived from the entity tokens under the knowledge graph preference policy μ_θ, and the second case focuses on the reward derived from the non-entity tokens under the base LLM policy π_0. We separately prove the unbiasedness by showing that the expected value of either the entity-based or non-entity-based estimators is equal to the true expected reward under their respective policies.

The expected value of the entity tokens in the knowledge graph is:

$$\begin{aligned}
\mathbb{E}\left[\hat{V}_{KG}(\theta)\right] &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\mu_\phi(\cdot|s_t^{(i)}),e\sim\mathcal{P}(e)}\left[\frac{\pi_\theta(e|s_t^{(i)})}{\mu_\phi(e|s_t^{(i)})}\log\pi_0(e|s_t^{(i)})\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\mu_\phi(\cdot|s_t^{(i)})}\left[\frac{\pi_\theta(e|s_t^{(i)})}{\mu_\phi(e|s_t^{(i)})}\mathbb{P}(e=\hat{y}|s_t^{(i)},e)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\pi_\theta(\cdot|s_t^{(i)})}\left[\mathbb{P}(e=\hat{y}|s_t^{(i)},e)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\pi_\theta(\cdot|s_t^{(i)})}\left[r(s_t^{(i)},e)\right] = V(\pi_\theta)
\end{aligned}$$

For non-entity tokens, the proof is similar:

$$\begin{aligned}
\mathbb{E}\left[\hat{V}_{base}(\theta)\right] &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\pi_0(\cdot|s_t^{(i)}),e\sim\mathcal{P}(e)}\left[\frac{\pi_\theta(e|s_t^{(i)})}{\pi_0(e|s_t^{(i)})}\log\pi_0(e|s_t^{(i)})\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\pi_0(\cdot|s_t^{(i)})}\left[\frac{\pi_\theta(e|s_t^{(i)})}{\pi_0(e|s_t^{(i)})}\mathbb{P}(e=\hat{y}|s_t^{(i)},e)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\mathbb{E}_{e\sim\pi_\theta(\cdot|s_t^{(i)})}\left[\mathbb{P}(e=\hat{y}|s_t^{(i)},e)\right] = V(\pi_\theta)
\end{aligned}$$

This completes the proof. □

B VARIANCE ANALYSIS

In this section, we derive a confidence bound from the confidence interval and calculate the lower bound of the variance for the KG-IPS estimator.

Proof. Let M be the maximum value of $\frac{\pi_\theta(e|s_t)}{\mu_\phi(e|s_t)}$ ranging over all entity tokens e. This quantity M represents the largest discrepancy between the target policy π_θ and the behavior policy μ_ϕ.

Since each reward r_t lies in $[0,1]$, it is $\frac{1}{4}$-sub-Gaussian. Consequently, multiplying r_t by a constant factor of at most M produces a random variable of the form

$$X_t = \frac{\pi_\theta(e|s_t)}{\mu_\phi(e|s_t)}r_t,$$

that is $(\frac{M^2}{4})$-sub-Gaussian. Let $\hat{V}_{\text{KG-IPS}}(\theta)$ be the average of n i.i.d. samples $\{X_t\}_{t=1}^n$. From the property of sub-Gaussian variables, if each X_t is $(\frac{M^2}{4})$-sub-Gaussian, then the average

$$\hat{V}_{\text{KG-IPS}}(\theta) = \frac{1}{n}\sum_{t=1}^{n}X_t$$

is $(\frac{M^2}{4n})$-sub-Gaussian. In particular, this implies that the variance cannot be smaller than $\Omega(\frac{M^2}{n})$, indicating an irreducible noise level of order $\frac{M}{\sqrt{n}}$.

For any sub-Gaussian random variable X with variance σ^2, the concentration inequality (Zhang & Chen, 2020) holds:

$$\left|\hat{X} - \mathbb{E}[X]\right| \le \sigma\sqrt{2\log\left(\frac{1}{\delta}\right)}.$$

Plugging the variables above into the inequality, we get the following bound for the KG-IPS estimator:

$$\left|\hat{V}_{\text{KG-IPS}}(\theta) - V(\theta)\right| \le M\sqrt{\frac{\log(1/\delta)}{2n}} = O(M\sqrt{\log(1/\delta)/n}),$$

with probability at least $1 - \delta$.

The variance of a sub-Gaussian random variable is close to its variance proxy, which means the lower bound on the variance of the KG-IPS estimator is $\Omega(\frac{M^2}{n})$. In addition, by standard concentration inequalities, we can get $O(M\sqrt{\log(1/\delta)/n})$ confidence intervals on our estimator for policy π_θ. $\square$

C Theoretical Analysis

The value function of policy π_θ is defined as:

$$V^{\pi_\theta}(s_t, a_t) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\left[r(s_t, a_t)\right].$$

Based on our settings, we optimize the target policy for entity tokens aligning with knowledge graph preference. The reward function is formulated as:

$$r^{\text{KG}}(s_t, a_t) = \sum_{e \in a_t} \frac{\pi_\theta(e|s_t)}{\mu_\phi(e|s_t)} \log \pi_0(e|s_t), \tag{5}$$

where μ_ϕ is the knowledge graph preference policy.

To reduce variance, the logged rewards for non-entity tokens under the base LLM policy π_0 are incorporated as a regularization term in the reward function, formulated as:

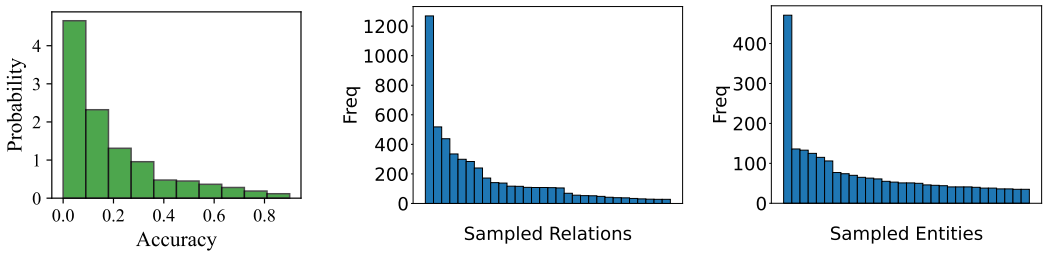
$$r^{\text{reg}}(s_t, a_t) = \sum_{v \in c_t \setminus a_t} \frac{\pi_\theta(v|s_t)}{\pi_0(v|s_t)} \log \pi_0(v|s_t), \tag{6}$$

where π_0 is the base LLM policy. This helps mitigate disturbances, ensuring the LLM's behavior on non-entity tokens remains stable and preventing model degeneration.

The final reward is:

$$r(s_t, a_t) = r^{\text{KG}}(s_t, a_t) + r^{\text{reg}}(s_t, a_t), \tag{7}$$

Since both $r^{\text{KG}}(s_t, a_t)$ and $r^{\text{reg}}(s_t, a_t)$ are reweightings of the log-based reward $\log \pi_0(v|s_t)$, they belong to the same equivalence class. By leveraging Lemma 2 from DPO (Rafailov et al., 2024), we show that the optimal policy for the task-specific reward r^ϕ aligns with the optimal policy for the final reward r. This implies that both rewards induce the same optimal policy.



(a) Distribution of QA Accuracy (b) Distribution of Relations (c) Distribution of Entities

Figure 4: Sampling distributions of (a) trajectories in the knowledge graph that are verbalized as multi-step QA tasks and successfully answered by the LLM itself, (b) relations, and (c) entities in the knowledge graphs and their frequencies of the appearance in the trajectories sampled from the *Wikidata5M* (Wang et al., 2021) knowledge graph.

D Details of Knowledge Graph Preference Modeling

The verbalized trajectories have in average 141.64 tokens with a standard deviation of 34.39. The average trajectory length is 4.79 steps with a standard deviation of 0.56. For each trajectory, there are in average 5.79 entities with a standard deviation of 0.56, and 4.11 relations with a standard

deviation of 0.90. In Figure 4a, we present the probability distribution of sampled trajectories, with respect to the number of correct answers generated per trajectory from ten differently sampled questions associated with each trajectory. Based on such self-consistency measurement, we estimate the reward function $R(\boldsymbol{h}|\boldsymbol{c})$ as the normalized question-answering accuracy.

Knowledge Graph Preference Model. The knowledge graph preference model is developed based on the pre-trained GPT2-Medium model (Radford et al., 2019). We collected 6K question-answering pairs from the Wikidata5M (Wang et al., 2021) knowledge graph based on the sampling strategy in Section 4.2. The sampled knowledge graph trajectories are composed into natural language prefixed by the corresponding questions by the GPT-4 model, which verbalizes the knowledge graph reasoning trajectories and aligns with generative language models' behaviors. The model is then fine-tuned with a base learning rate of $1e-4$ for 10 epochs with a linear learning scheduler.