From Words to Actions: Unveiling the Theoretical Underpinnings of LLM-Driven Autonomous Systems

Jianliang He [*1] **Siyu Chen** [*2] **Fengzhuo Zhang** [3] **Zhuoran Yang** [2]

Abstract

In this work, from a theoretical lens, we aim to understand why large language model (LLM) empowered agents are able to solve decision-making problems in the physical world. To this end, consider a hierarchical reinforcement learning (RL) model where the LLM Planner and the Actor perform high-level task planning and low-level execution, respectively. Under this model, the LLM Planner navigates a partially observable Markov decision process (POMDP) by iteratively generating language-based subgoals via prompting. Under proper assumptions on the pretraining data, we prove that the pretrained LLM Planner effectively performs Bayesian aggregated imitation learning (BAIL) through in-context learning. Additionally, we highlight the necessity for exploration beyond the subgoals derived from BAIL by proving that naively executing the subgoals returned by LLM leads to a linear regret. As a remedy, we introduce an ϵ-greedy exploration strategy to BAIL, which is proven to incur sublinear regret when the pretraining error is small. Finally, we extend our theoretical framework to include scenarios where the LLM Planner serves as a world model for inferring the transition model of the environment and to multi-agent settings, enabling coordination among multiple Actors.

1. Introduction

The advent of large language models (LLMs) such as GPT-4 (OpenAI, 2023) and Llama 2 (Touvron et al., 2023) has marked a significant leap in artificial intelligence, thanks to their striking capabilities in understanding language and performing complex reasoning tasks. These capabilities of LLMs have led to the emergence of LLM-empowered agents (LLM Agents), where LLMs are used in conjunction with tools or actuators to solve decision-making problems in the physical world. LLM Agents have showcased promising empirical successes in a wide range of applications, including autonomous driving (Wang et al., 2023b; Fu et al., 2024), robotics (Brohan et al., 2023; Li et al., 2023a), and personal assistance (Liu et al., 2023; Nottingham et al., 2023). This progress signifies a crucial advancement in the creation of intelligent decision-making systems, distinguished by a high degree of autonomy and seamless human-AI collaboration.

LLMs only take natural languages as input. To bridge the language and physical domains, LLM-agents typically incorporate three critical components: an LLM Planner, a physical Actor, and a multimodal Reporter, functioning respectively as the brain, hands, and eyes of the LLM-agent, respectively. Specifically, upon receiving a task described by a human user, the LLM Planner breaks down the overall task into a series of subgoals. Subsequently, the Actor implements each subgoal in the physical world through a sequence of actions. Meanwhile, the Reporter monitors changes in the physical world and conveys this information back to the LLM Planner in natural language form. This dynamic interaction among Planner, Actor, and Reporter empowers LLM Agents to understand the environment, formulate informed decisions, and execute actions effectively, thus seamlessly integrating high-level linguistic subgoals with low-level physical task execution.

The revolutionary approach of LLM Agents represents a paradigm shift away from traditional learning-based decision-making systems. Unlike these conventional systems, LLM Agents are not tailored to any specific task. Instead, they rely on the synergy of their three distinct componentseach trained separately and often for different objectives. In particular, the LLM Planner is trained to predict the next token in a sequence on vast document data. Moreover, when deployed to solve a task, the way to interact with the LLM Planner is via prompting with the LLM fixed. The Actor, as language-conditioned policies, can be trained

[*]Equal contribution [1]Fudan University [2]Yale University [3]National University of Singapore. Correspondence to: Jianliang He <hejl20@fudan.edu.cn>, Zhuoran Yang <zhuoran.yang@yale.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

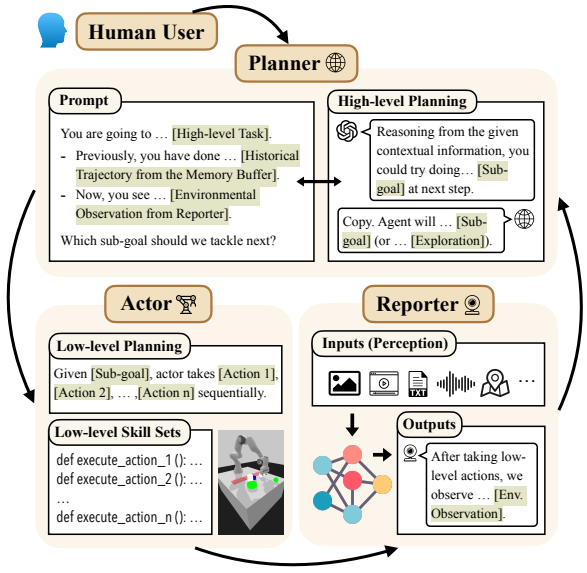

Figure 1. Overview of the Planner-Actor-Reporter (PAR) system as LLM Agents. Acting as a central controller, the Planner conducts the high-level planning by storing the history and reasoning through the iterative use of the ICL ability of LLMs, coupled with explorations. The Actor handles low-level planning and executes subgoals using pre-programmed skill sets, and the Reporter perceives and processes multimodal information from environment to reinforce the ongoing planning.

by RL or imitation learning. Moreover, the Reporter, as a multimodal model, is trained to translate the physical states (e.g., images) into natural language. This unique configuration prompts critical research questions regarding the theoretical underpinnings of LLM Agents, particularly concerning their decision-making effectiveness.

In this work, we make an initial step toward developing a theoretical framework for understanding the dynamics and effectiveness of LLM Agents. Specifically, we aim to answer the following questions: (a) What is a theoretical model for understanding the performance of LLM Agents? (b) How do pretrained LLMs solve decision-making problems in the physical world via prompting? (c) How does an LLM Agent address the exploration-exploitation tradeoff? (d) How do the statistical errors of the pretrained LLM and Reporter affect the overall performance of the LLM Agent?

To address Question (a), we propose analyzing LLM Agents within a hierarchical reinforcement learning framework (Barto & Mahadevan, 2003; Pateria et al., 2021), positioning the LLM Planner and the Actor as policies operating within high-level POMDPs and low-level MDPs, respectively (§3.1). Both levels share the same state space-namely, the physical statethough the LLM Planner does not directly observe this state but instead receives a language-based description from the Reporter, effectively navigating a POMDP. The action space of the high-level POMDP is the set of language subgoals. Meanwhile, the state transition kernel is determined by the pretrained Actor, and thus is associated with a variable z that summarizes its dependency on low-level Actor. Such a variable is unknown to the LLM Planner. After pretraining, without prior knowledge of the Actor's quality or the physical environment, the LLM Planner attempts to solve the high-level POMDP by iteratively generating a sequence of subgoals based on feedback from Reporter via prompting. Under this framework, the overall performance of the LLM Agent can be captured by the regret in terms of finding the optimal policy of the hierarchical RL problem in the online setting (§3.2).

Furthermore, to answer Question (b), we prove that when pretraining data includes a mixture of expert trajectories, during the prompting stage, the pretrained LLM Planner essentially performs Bayesian aggregated imitation learning (BAIL) through in-context learning (Theorem 4.2). This process involves constructing a posterior distribution over the hidden parameter z of the transition kernel, followed by generating subgoals that emulate a randomly selected expert policy, weighted according to this posterior distribution. Such a Bayesian learning mechanism is encoded by the LLM architecture and is achieved via prompting.

However, since the LLM has no prior knowledge of the physical environment, it needs to guide the Actor to explore the physical environment. We prove that merely adhering to BAIL-derived subgoals can lead to the inadequate exploration, resulting in a linear regret (Proposition 4.3). To mitigate this, i.e., Question (c), we introduce an ϵ-greedy exploration strategy, which occasionally deviates from BAIL subgoals in favor of exploration, significantly enhancing learning efficacy by ensuring a sublinear regret (Theorem 4.6). Specifically, to address Question (d) we establish that the regret is bounded by a sum of two terms (Theorem 5.7): a $\sqrt{T}$-regret related to the number of episodes the LLM Agent is deployed to the hierarchical RL problem, and an additional term representing the statistical error from pretraining the LLM Planner and Reporter via maximum likelihood estimation (MLE) and contrastive learning, respectively (Theorem 5.3, 5.5).

Finally, we extend our analysis to scenarios where the Planner utilizes the LLM as world model for inferring the upper-level POMDP's transition model via Bayesian model aggregation (Proposition B.1, Corollary B.3). Our theoretical framework also accommodates a multi-agent context, where the LLM Planner coordinates with a collaborative team of low-level actors (Corollary B.4).

2. Preliminaries and Related Works

Large Language Models and In-Context Learning. The Large Language Models (LLMs) such as ChatGPT (Brown

et al., 2020), GPT-4 (OpenAI, 2023), Llama (Touvron et al., 2023), Gemini (Team et al., 2023), are pretrained on vast text corpora to predict in an *autoregressive* manner. Starting from an initial token $\ell_1 \in \mathfrak{L} \subseteq \mathbb{R}^d$, where d denotes the token vector dimension and $\mathfrak{L}$ denotes the language space, the LLM, with parameters $\theta \in \Theta$, predicts the next token with $\ell_{t+1} \sim \texttt{LLM}_\theta(\cdot \,|\, S_t)$, where $S_t = (\ell_1, \ldots, \ell_t)$ and $t \in \mathbb{N}$. Each token $\ell_t \in \mathfrak{L}$ specifies a word or the word's position, and the token sequence S_t resides in the space of token sequences $\mathfrak{L}^*$. Such an autoregressive generating process terminates when stop sequence token is generated.

Unlike fine-tuned models customized for specific domains or tasks, LLMs showcase comparable capabilities by learning from the informative *prompts* (Li et al., 2022; Liu et al., 2022b), which is known as in-context learning (ICL, Brown et al., 2020). Assume that prompt $\texttt{pt}_t = (\ell_1, \ldots, \ell_t) \in \mathfrak{L}^*$, is generated based on a latent variable $z \in \mathcal{Z}$ autoregressively. The token follows a generating distribution such that $\ell_t \sim \mathbb{P}(\cdot \,|\, \texttt{pt}_{t-1}, z)$ and $\texttt{pt}_t = (\texttt{pt}_{t-1}, \ell_t)$, where $\mathcal{Z}$ represents the space of hidden information or latent concepts. This latent structure is commonly employed in language models, including topic models like LDA (Blei et al., 2003), BERT (Devlin et al., 2018), generative models like VAE (Kusner et al., 2017), T5 (Raffel et al., 2020). Such an assumption is also widely adopted in the theoretical analysis of ICL (Xie et al., 2021; Zhang et al., 2023). Following this, we build upon the framework attributing the ICL capability to Bayesian inference Xie et al. (2021); Jiang (2023); Zhang et al. (2023), which posits that the pretrained LLM predicts the next token with probability by aggregating the generating distribution concerning latent variable $z \in \mathcal{Z}$ over the posterior distribution. Moreover, a series of practical experiments, including Wang et al. (2023a); Ahuja et al. (2023), provide empirical support for Bayesian statement.

LLM Agents. LLMs, as highlighted in (OpenAI, 2023), are powerful tools for task planning (Wei et al., 2022a; Hu & Shu, 2023). The success of LLM Agents marks a shift from the task-specific policies to a pretrain-finetune-prompt paradigm (Mandi et al., 2023; Brohan et al., 2023; Lin et al., 2023a; Hao et al., 2023; Liu et al., 2023). By breaking down the complex tasks into subgoals, LLM Agent facilitates effective zero-shot resource allocation across environments. Envision a scenario where a robotic arm is tasked with "*move a teapot from the stove to a shelf*", a task for which the robotic arm may not be pretrained. However, leveraging LLMs allows the decomposition of the task into a sequence of executable subgoals: "*grasp the teapot*", "*lift the teapot*", "*move the teapot to the shelf*", and "*release the teapot*". We formalize this approach into a hierarchical LLM-empowered planning framework and provide a detailed theoretical analysis of its performance.

More related works are deferred to §A.1 due to space limit.

3. Theoretical Framework for LLM Agent

To formalize the architecture of LLM Agents, we propose a general theoretical framework—Planner-Actor-Reporter (PAR) system. Furthermore, the problem is modeled as a hierarchical RL problem (Pateria et al., 2021). Specifically, the Planner, empowered by LLMs, conducts high-level task planning within the language space; the Actor, pretrained before deployment, undertakes low-level motion planning within the physical world; and the Reporter, equipped with a sensor to sense the physical environment, processes the information and feeds it back to the Planner, bridging the gap between language space and the physical world (see §3.1). Additionally, we present the performance metric and pretraining methods of LLMs for the Planner and translators for the Reporter in §3.2.

3.1. Planner-Actor-Reporter System

In this section, we delve into details of the PAR system under Hierarchical Markov Decision Process (HMDP). At the high level, the Planner empowered by LLM handles task planning by decomposing tasks into subgoals to solve a language-conditioned Partially Observable Markov Decision Process (POMDP) with a finite horizon H. At the low level, the Actor translates these subgoals into the actionable steps in the physical world to handle a language-conditioned Markov Decision Process (MDP) with a finite horizon H_a[1]. Please refer to Figure 2 for an overview of the hierarchical interactive process.

Low-level MDP. Let $\mathcal{G} \subseteq \mathfrak{L}$ be the space of language subgoals, $\mathcal{S}$ and $\mathcal{A}$ respectively denote the space of physical states and actions. At high-level step h, the low-level MDP is specified by a transition kernel $\mathbb{T}_h = \{\mathbb{T}_{h,\bar{h}}\}_{\bar{h}\in[H_a]}$ and the rewards that depends on a subgoal $g \in \mathcal{G}$. Following this, the Actor is modelled as a language-conditioned policy $\mu = \{\mu_g\}_{g\in\mathcal{G}}$, where $\mu_g = \{\mu_{\bar{h}}(\cdot|\cdot, g)\}_{\bar{h}\in[H_a]}$ and $\mu_{\bar{h}} : \mathcal{S} \times \mathcal{G} \mapsto \Delta(\mathcal{A})$. Assume that the Actor stops at step H_a+1, regardless of the subgoal achievement. Subsequently, the Planner receives the observation of the current state $\bar{s}_{h,H_a+1}$ from the Reporter, and sends a new subgoal to the Actor based on the historical feedback.

High-level POMDP. Assume that each low-level episode corresponds to a single high-level action of Planner. Thus, the high-level POMDP reuses the physical state space $\mathcal{S}$ as the state space, but takes the subgoal space $\mathcal{G}$ as the action space instead. Following this, the high-level transition kernel is jointly determined by the low-level policy μ and the physical transition kernel $\mathbb{T}$ such that

$$\mathbb{P}_{z,h}(s' \,|\, s, g) = \mathbb{P}\big(\bar{s}_{h,H_a+1} = s' \,|\, \bar{s}_{h,1} = s, a_{h,1:\bar{h}} \sim \mu_g, \bar{s}_{h,2:\bar{h}+1} \sim \mathbb{T}_h\big),$$

[1]Throughout the paper, we use the notation $\bar{\cdot}$ to distinguish low-level elements from their high-level counterparts.

where we write $z = (\mathbb{T}, \mu)$. Since the LLM-empowered Planner cannot directly process the physical states, it relies on some (partial) observations generated by the Reporter. Specifically, let $o_h \in \mathcal{O}$ describe the physical state $s_h \in \mathcal{S}$ in language through a translation distribution $\mathbb{O} : \mathcal{O} \mapsto \Delta(\mathcal{S})$, where $\mathcal{O} \subseteq \mathfrak{L}$ denotes the space of observations. At each step $h \in [H]$, a reward $r_h(o_h, \omega) \in [0, 1]$ is obtained, which depends on both the observation and the task $\omega \in \Omega$ assigned by human users. Here, $\Omega \subseteq \mathfrak{L}$ denotes the space of potential tasks in language.

Interactive Protocol. The Planner aims to determine a sequence of subgoal $\{g_h\}_{h\in[H]}$ such that when the Actor is equipped with policy $\pi = \{\pi_h\}_{h\in[H]}$, these subgoals maximize the expected sum of rewards. During task planning, the Planner must infer both Actor's intention, i.e., policy μ, and the environment, i.e., physical transition kernel $\mathbb{T}$, from the historical information. Thus, z constitutes all the latent information to the high-level Planner, and denote $\mathcal{Z}$ as the space of all potential latent variables with $|\mathcal{Z}| < \infty$. To summarize, the interactive protocol is as below: at the beginning of each episode t, Planner receives a task ω_t. At step h, each module follows:

Module 1: Planner. After collecting o_h^t from Reporter, the Planner leverages LLMs for recommendations on task decomposition, and the policy is denoted by $\pi_{h,\texttt{LLM}}^t : \mathcal{T}^* \times (\mathcal{O} \times \mathcal{G})^{h-1} \times \mathcal{O} \times \Omega \mapsto \Delta(\mathcal{G})$, where $\mathcal{T}^*$ represents the space of the trajectory sequence with arbitrary length. LLM's recommendations are obtained by invoking the ICL ability with the history-dependent prompt:

$$\texttt{pt}_h^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t\}, \quad \mathcal{H}_t = \bigcup_{i=1}^{t-1} \{\omega^i, \tau_H^i\}, \quad (1)$$

where $\mathcal{H}_t \in \mathcal{T}^*$ denotes the historical context and $\tau_h^t = \{o_1^t, g_1^t, \ldots, o_h^t\}$ is the trajectory until h-th step. In the PAR system, Planner retains autonomy and is not obligated to follow LLM's recommendations. Let π_h^t be the Planner's policy, which partially leverages the LLM's recommendation $\pi_{h,\texttt{LLM}}^t(\cdot \,|\, \tau_h^t, \omega^t) = \texttt{LLM}_\theta(\cdot \,|\, \texttt{pt}_h^t)$. The Planner selects $g_h^t \sim \pi_h^t(\cdot \,|\, \tau_h^t, \omega^t)$, and sends it to the Actor.

Module 2: Actor. Upon receiving g_h^t from Planner, the Actor plans to implement g_h^t in physical world with given skill sets, denoted by a subgoal-conditioned policy $\mu = \{\mu_g\}_{g\in\mathcal{G}}$. Then, a sequence of actions $\{a_{h,\bar{h}}\}_{\bar{h}\in[H_a]}$ is executed, where the dynamics follows $a_{h,\bar{h}} \sim \mu_{\bar{h}}(\cdot \,|\, \bar{s}_{h,\bar{h}}, g_h^t)$, $\bar{s}_{h,\bar{h}+1} \sim \mathbb{T}_{h,\bar{h}}(\cdot \,|\, \bar{s}_{h,\bar{h}}, a_{h,\bar{h}})$ starting from $\bar{s}_{h,1} = s_h^t$. The low-level episode concludes at $s_{h+1}^t = \bar{s}_{h,H_a+1}$.

Module 3: Reporter. After the low-level episode concludes, the Reporter collects and reports the current state s_h^t via observation o_{h+1}^t generated from $\mathbb{O}_\gamma(\cdot \,|\, s_{h+1}^t)$, where $\mathbb{O}_\gamma : \mathcal{S} \mapsto \Delta(\mathcal{O})$ denotes the distribution of the pretrained translator. Subsequently, the observation o_{h+1}^t of the current state is sent back to the Planner, reinforcing to the ongoing task planning.

The strength of the PAR system lies in its resemblance to RL (Sutton & Barto, 2018), allowing the Planner to iteratively adjust its planning strategy based on feedback from the Reporter. Moreover, the Reporter empowers the system to process the real-time information and the integration of multiple modalities of raw data like RGB, images, LiDAR, audio, and text (Li et al., 2023b; Xu et al., 2023). The Actor's skill sets can be pretrained using goal-conditioned RL (Chane-Sane et al., 2021; Liu et al., 2022a), language-to-environment grounding (Brohan et al., 2023; Huang et al., 2022) or pre-programmed manually (Singh et al., 2023).

3.2. Performance Metric and Pretraining

Performance Metric. In this paper, we focus on the *performance of high-level Planner*, and regard the low-level Actor as an autonomous agent that can use the pretrained skill sets following a fixed policy. For any latent variable $z \in \mathcal{Z}$ and policy $\pi = \{\pi_h\}_{h\in[H]}$ with $\pi_h : (\mathcal{O} \times \mathcal{G})^{h-1} \times \mathcal{O} \times \Omega \mapsto \Delta(\mathcal{G})$, the value function is defined as

$$\mathcal{J}_z(\pi, \omega) := \mathbb{E}_\pi \left[\sum_{h=1}^{H} r_h\left(o_h, \omega\right) \right], \quad (2)$$

where the expectation is taken concerning the initial state $s_1 \sim \rho$, policy π, ground-truth translation distribution $\mathbb{O}$, and transition kernel $\mathbb{P}_z$. For all $(z, \omega) \in \mathcal{Z} \times \Omega$, there exists an optimal policy $\pi_z^*(\omega) = \operatorname{argmax}_{\pi\in\Pi} \mathcal{J}_z(\pi, \omega)$, where $\Pi = \{\pi = \{\pi_h\}_{h\in[H]}, \pi_h : (\mathcal{O}\times\mathcal{G})^{h-1} \times \mathcal{O} \times \Omega \mapsto \Delta(\mathcal{G})\}$.

To characterize the performance under practical setting, we denote $\widehat{\mathcal{J}}_z(\pi, \omega)$ as the value function concerning the pretrained translator $\mathbb{O}_{\widehat{\gamma}}$, and for all $\omega \in \Omega$, let $\widehat{\pi}_z^*(\omega) = \operatorname{argmax}_{\pi\in\Pi} \widehat{\mathcal{J}}_z(\pi, \omega)$ be the optimal policy in practice. Then, the regret under practical setting is defined as

$$\mathrm{Reg}_z(T) := \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t} \left[\widehat{\mathcal{J}}_z(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_z(\widehat{\pi}^t, \omega^t) \right], \quad (3)$$

where $\{\widehat{\pi}^t\}_{t\in[T]}$ represents the Planner's policy empowered by a pretrained $\texttt{LLM}_{\widehat{\theta}}$ and the expectation is taken with respect to the context $\mathcal{H}_t$ defined in (1) generated by taking $\{\widehat{\pi}^i\}_{i<t}$ sequentially. Here, we focus on the performance when the Planner collaborates with a pretrained PAR system in an environment characterized by z and pretrained Reporter. Our goal is to design a sample-efficient algorithm that achieves a sublinear regret, i.e., $\mathrm{Reg}_z(T) = o(T)$.

Pretraining Dataset Collection. The pretraining dataset contains N_{p} independent T-episode samples such that $\mathcal{D} = \{D_n\}_{n\in[N_{\mathrm{p}}]}$ with $D_n = \{z\} \cup \{\omega^t, \tau_H^t, g_{1:H}^{t,*}, s_{1:H}^t\}_{t\in[T_{\mathrm{p}}]}$.

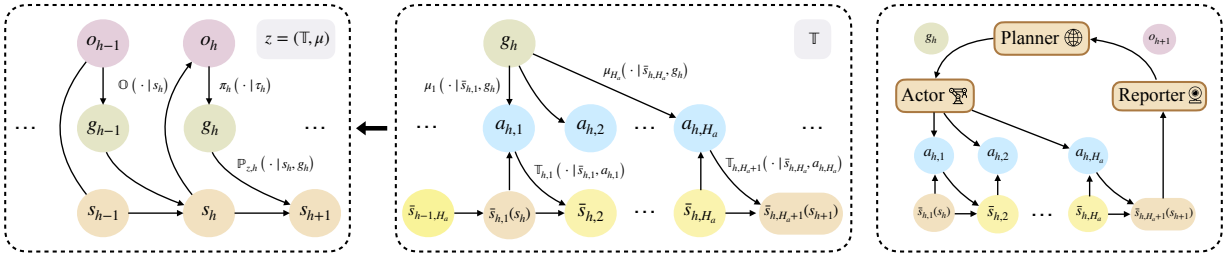

Figure 2. Illustration of structure of HMDP. The low-level MDP is featured by transition kernel $\mathbb{T}$, which characterizes the dynamics of the physical environment. The high-level transition is a result of a sequence of low-level actions in the physical environment, guided by policies $\mu = \{\mu_g\}_{g\in\mathcal{G}}$. Thus, high-level POMDP incorporates latent information $z = (\mathbb{T}, \mu)$ originated from the low-level.

For each sample, $z \sim \mathcal{P}_{\mathcal{Z}}$ specifies a low-level MDP with language-conditioned policies and $\omega^t \sim \mathcal{P}_\Omega$ specifies the sequence of high-level tasks. Here, $\mathcal{P}_{\mathcal{Z}}$ and $\mathcal{P}_\Omega$ denote the prior distributions. We assume that the joint distribution of each data point D in the dataset, denoted by $\mathbb{P}_{\mathcal{D}}$, follows

$$\mathbb{P}_{\mathcal{D}}(D) = \mathcal{P}_{\mathcal{Z}}(z) \cdot \prod_{t=1}^{T_{\mathrm{p}}} \mathcal{P}_\Omega(\omega^t) \cdot \prod_{h=1}^{H} \pi^*_{z,h}(g_h^{t,*} \,|\, \tau_h^t, \omega^t) \cdot \mathbb{O}(o_h^t \,|\, s_h^t) \cdot \pi_h^b(g_h^t \,|\, \tau_h^t, \omega^t) \cdot \mathbb{P}_{z,h}(s_{h+1}^t \,|\, s_h^t, g_h^t), \quad (4)$$

where $\pi^b = \{\pi_h^b\}_{h\in[H]}$ is the behavior policy that features how the contextual information is collected, and then additionally the label, i.e., optimal subgoal, is sampled from the optimal policy π_z^* by experts. Subsequently, the latent information z is hidden from the context.

LLM Pretraining. To pretrain LLMs, we adopt a supervised learning approach concerning the transformer structure, aligning with the celebrated LLMs such as BERT and GPT (Devlin et al., 2018; Brown et al., 2020). Specifically, the pretraining data is constructed based on $\mathcal{D}$. For clarity, we extract the language data without expert knowledge and write the collected data into a sequence of ordered tokens, i.e., sentences or paragraphs. For the n-th sample D_n, let

$$(\ell_1^n, \ldots, \ell_{\bar{T}_{\mathrm{p}}}^n) := \left(\omega^{n,t}, o_1^{n,t}, g_1^{n,t}, \ldots, o_{H-1}^{n,t}, g_{H-1}^{n,t}, o_H^{n,t}\right)_{t\in[T_{\mathrm{p}}]},$$

with a length of $\bar{T}_{\mathrm{p}} = 2HT_{\mathrm{p}}$, which contains T_{p} episodes with one task, H observations and $H-1$ subgoals. Following this, LLM's pretraining dataset is autoregressively constructed with the expert guidance, denoted by $\mathcal{D}_{\texttt{LLM}} = \{(\tilde{\ell}_t^n, S_t^n)\}_{(n,t)\in[N_{\mathrm{p}}]\times[\bar{T}_{\mathrm{p}}]}$, where $S_{t+1}^n = (S_t^n, \ell_t^n)$ and let

$$\begin{cases} \tilde{\ell}_{t'}^n = g_h^{n,t,*} & \text{if } t' = 2H(t-1) + 2h + 1, \\ \tilde{\ell}_{t'}^n = g_h^{n,t} & \text{otherwise.} \end{cases}$$

In other words, when pretraining to predict the next subgoal, we replace the one sampled from the behavior policy with the one from the optimal policy. In practice, sentences with expert knowledge can be collected from online knowledge platforms such as Wikipedia (Merity et al., 2016; Reid et al., 2022). Following the pretraining algorithm of BERT and GPT, the objective is to minimize the cross-entropy loss, which can be summarized as $\widehat{\theta} = \mathrm{argmin}_{\theta\in\Theta} \mathcal{L}_{\mathrm{CE}}(\theta; \mathcal{D}_{\texttt{LLM}})$ with

$$\mathcal{L}_{\mathrm{CE}}(\theta; \mathcal{D}_{\texttt{LLM}}) := \widehat{\mathbb{E}}_{\mathcal{D}_{\texttt{LLM}}} \left[-\log \texttt{LLM}_\theta(\ell \,|\, S)\right], \quad (5)$$

and $\texttt{LLM}_{\widehat{\theta}}$ is the pretrained LLM by algorithm in (5). More details are deferred to §5.1.

Translator Pretraining. To pretrain translators, we employ a self-supervised contrastive learning approach, which aligns with celebrated vision-language models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). Let $\mathcal{D}_{\texttt{Rep}}$ be the contrastive pretraining dataset for translators, which is also constructed upon the dataset $\mathcal{D}$. Following the framework adopted in (Qiu et al., 2022; Zhang et al., 2022), for each observation-state pair $(o, s) \in \mathcal{D}$, a positive or a negative data point, labelled as $y = 1$ and $y = 0$, is generated with equal probability, following that

- **Positive Data:** Collect (o, s) with label $y = 1$.
- **Negative Data:** Collect (o, s^-) with label $y = 0$, where s^- is sampled from negative sampling distribution $\mathcal{P}^- \in \Delta(\mathcal{O})$ with a full support on the domain of interest.

Denote $\mathbb{P}_{\mathcal{C}}$ as the joint distribution of data collected by the process above. The learning algorithm follows that $\widehat{\gamma} = \mathrm{argmin}_{\gamma\in\Gamma} \mathcal{L}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\texttt{Rep}})$, where the contrastive loss $\mathcal{L}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\texttt{Rep}})$ is defined as

$$\mathcal{L}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\texttt{Rep}}) := \widehat{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}[y \cdot \log\left(1 + f_\gamma(o,s)^{-1}\right) + (1-y) \cdot \log\left(1 + f_\gamma(o,s)\right)]. \quad (6)$$

Consider function class $\mathcal{F}_\gamma$ with finite elements with $\mathcal{F}_\gamma \subseteq (\mathcal{S} \times \mathcal{O} \mapsto \mathbb{R})$ serving as a set of candidate functions that approximates the ground-truth likelihood ratio $f^*(\cdot,\cdot) =$

$\mathbb{O}(\cdot\,|\,\cdot)/\mathcal{P}^-(\cdot)$ (see Lemma D.2 for justification). Following this, the pretrained translator for the Reporter by the algorithm in (6) is thus defined as $\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,\cdot) = f_{\widehat{\gamma}}(\cdot,\cdot)\cdot\mathcal{P}^-(\cdot)$. More details are deferred to §5.2.

Remark 3.1. In (4), we assume that all pretraining data is generated from a joint distribution $\mathbb{P}_{\mathcal{D}}$, and then split for pretraining of LLM and Reporter. In practice, the pretraining dataset for the Reporter can consist of paired observation-state data collected from any arbitrary distribution, as long as (i) the LLM and Reporter "speak" the same language, i.e., shared $\mathbb{O}$, and (ii) the coverage assumption can hold (see Assumption 5.6).

4. LLM Planning via Bayesian Aggregated Imitation Learning

In this section, we show that LLMs can conduct high-level planning through Bayesian aggregated imitation learning (BAIL) in §4.1, leveraging the ICL ability of LLMs with the history-dependent prompts. However, depending solely on LLM's recommendations proves insufficient for achieving sample efficiency under the worst case (see Proposition 4.3). Following this, we propose a planning algorithm for Planner in §4.2, leveraging LLMs for expert recommendations, in addition to an exploration strategy.

4.1. Bayesian Aggregated Imitation Learning

In this subsection, we show that the LLM conducts high-level task planning via BAIL, integrating both Bayesian model averaging (BMA, Hoeting et al., 1999) during the online planning and imitation learning (IL, Ross & Bagnell, 2010) during the offline pretraining. Intuitively, pretrained over $\mathcal{D}_{\texttt{LLM}}$, LLM approximates the conditional distribution $\texttt{LLM}(\ell = \cdot\,|\,S) = \mathbb{P}_{\mathcal{D}}(\ell = \cdot\,|\,S)$, where $\mathbb{P}_{\mathcal{D}}$ is the joint distribution in (4) and the randomness introduced by the latent variable is aggregated, i.e., $\mathbb{P}_{\mathcal{D}}(\ell = \cdot\,|\,S) = \mathbb{E}_{z\sim\mathbb{P}_{\mathcal{D}}(\cdot|S)}\left[\mathbb{P}_{\mathcal{D}}(\ell = \cdot\,|\,S, z)\right]$. Here, $\mathbb{P}_{\mathcal{D}}(\ell = \cdot\,|\,S, z)$ can be viewed as a generating distribution with a known z and is then aggregated over the posterior distribution $\mathbb{P}_{\mathcal{D}}(z = \cdot\,|\,S)$, aligning with the form of BMA (Zhang et al., 2023). We temporarily consider the perfect setting.

Definition 4.1 (Perfect Setting)**.** We say the PAR system is perfectly pretrained if (i) $\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s) = \mathbb{O}(\cdot\,|\,s)$ for all $s \in \mathcal{S}$, (ii) $\texttt{LLM}_{\widehat{\theta}}(\cdot\,|\,S_t) = \texttt{LLM}(\cdot\,|\,S_t)$ for all $S_t = (\ell_1, \ldots, \ell_t) \in \mathfrak{L}^*$ with length $t \le \bar{T}_{\mathrm{p}}$.

The assumption states that the Reporter and LLMs can report and predict with ground-truth distributions employed based on the joint distribution $\mathbb{P}_{\mathcal{D}}$. During ICL, we invoke LLMs by history-dependent $\texttt{pt}_h^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t\} \in \mathfrak{L}^*$ for all $(h,t) \in [H]\times[T]$. Conditioned on latent variable z and $\texttt{pt}_h^t$, the generating distribution is the optimal policy such that $\mathbb{P}_{\mathcal{D}}(\cdot\,|\,\texttt{pt}_h^t, z) = \pi^*_{z,h}(\cdot\,|\,\tau_h^t, \omega^t)$, which is independent of historical $\mathcal{H}_t$. In this sense, LLMs imitate expert policies during pretraining. The proposition below shows that LLMs conduct task planning via BAIL.

Algorithm 1 Planning with PAR System - Planner

Input: Policy $\pi_{\texttt{exp}}$ with $\eta \in (0,1)$, $c_{\mathcal{Z}} > 0$, $|\mathcal{Z}| \in \mathbb{N}$.
Initialize: $\mathcal{H}_0 \leftarrow \emptyset$, $\epsilon \leftarrow (H\log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.
for episode t from 1 to T **do**
  Receive the high-level task ω^t from the human user.
  Sample $\mathcal{I}_t \sim$ Bernuolli(ϵ).
  for step h from 1 to H **do**
    Collect the observation o_h^t from the Reporter.
    Set $\texttt{pt}_h^t \leftarrow \mathcal{H}_t \cup \{\omega^t, \tau_h^t\}$.
    Sample $g_{h,\texttt{LLM}}^t \sim \texttt{LLM}(\cdot\,|\,\texttt{pt}_h^t)$ via prompting.
    If $\mathcal{I}_t = 1$ **then** set $g_h^t \leftarrow g_{h,\texttt{LLM}}^t$,
      else sample $g_h^t \sim \pi_{h,\texttt{exp}}(\cdot\,|\,\tau_h^t)$.
    Send the subgoal g_h^t to the Actor.
  end for
  Update $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{\omega^t, \tau_H^t\}$.
end for

Proposition 4.2 (LLM Performs BAIL)**.** *Assume that the pretraining data distribution is given by* (4)*. Under the perfect setting in Definition 4.1, for all* $(h,t) \in [H]\times[T]$*, the LLM conducts task planning via BAIL, following that*

$$\pi^t_{h,\texttt{LLM}}\left(\cdot\,|\,\tau_h^t, \omega^t\right) = \sum_{z\in\mathcal{Z}} \pi^*_{z,h}\left(\cdot\,|\,\tau_h^t, \omega^t\right)\mathbb{P}_{\mathcal{D}}\left(z\,|\,\texttt{pt}_h^t\right),$$

where the prompt is defined in (1)*.*

Proposition 4.2 suggests that LLMs provide recommendations following a two-fold procedure: Firstly, LLMs compute the posterior belief of each latent variable $z \in \mathcal{Z}$ from $\texttt{pt}_h^t$. Secondly, LLMs aggregate the optimal policies over posterior probability and provide recommendations.

4.2. LLM-Empowered Planning Algorithm

Following the arguments above, we propose a planning algorithm for the Planner within a perfect PAR system. From a high level, the process of task planning is an implementation of policies from imitation learning (Ross & Bagnell, 2010; Ross et al., 2011) with two key distinctions: (i) Planner collaborates with LLM, a "nascent" expert that learns the hidden intricacies of the external world from updating prompts; (ii) different from behavior cloning or inverse RL, Planner does not aim to comprehend LLM's behaviors. Instead, the imitation is accomplished during the offline pretraining, and Planner shall selectively adhere to LLM's suggestions during online planning. Next, we show that task planning solely guided by LLMs fails to achieve sample efficiency in the worst case.

Proposition 4.3 (Hard-to-Distinguish Example)**.** *Suppose that Definition 4.1 holds. Given any* $T \in \mathbb{N}$*, there exists an*

HMDP and specific latent varibale $z \in \mathcal{Z}$ such that if Planner strictly follows LLM's recommended policies in Proposition 4.2, it holds that $\mathrm{Reg}_z(T) \geq 0.5T \cdot (1 - 1/|\mathcal{Z}|)^T$.

Proposition 4.3 indicates that relying solely on LLMs for task planning can result in a suboptimal $\Omega(T)$ regret in the worst case when $|Z| = T$. Thus, additional exploration is essential to discern the latent information about the external world, a parallel to the practical implementations in latent imitation learning (Edwards et al., 2019; Kidambi et al., 2021) and LLM-based reasoning (Hao et al., 2023; Nottingham et al., 2023). In practice, while the language model can guide achieving a goal, it's important to note that *this guidance is not grounded in real-world observations*. Thus, as pointed out by (Grigsby et al., 2023), the information provided in narratives might be arbitrarily wrong, which highlights the need for exploration to *navigate new environments effectively*. Similar to ϵ-greedy algorithms (Tokic & Palm, 2011; Dann et al., 2022), we provide a simple but efficient algorithm for LLM-empowered task planning. Algorithm 1 gives the pseudocode. In each episode, the Planner performs two main steps:

- **Policy Decision** (`Line 5`): Randomly decide whether to execute the exploration policy $\pi_{\texttt{exp}}$ or follow the LLM's recommendations within this episode with probability ϵ.
- **Planning with LLMs** (`Line` $7-10$): If Planner decides to follow the LLM's recommendations, it is obtained by prompting LLMs with $\texttt{pt}_h^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t\}$, equivalently sampling from $\texttt{LLM}(\cdot \,|\, \texttt{pt}_h^t)$. Otherwise, the Planner takes sub-goal from $\pi_{h,\texttt{exp}}(\cdot \,|\, \tau_h^t)$.

In conventional ϵ-greedy algorithms, explorations are taken uniformly over the action space $\mathcal{G}$, i.e., $\pi_{\texttt{exp}} = \mathrm{Unif}_{\mathcal{G}}$. Recent work has extended it to a collection of distributions (e.g., softmax, Gaussian noise) for function approximation (Dann et al., 2022). Following this, we instead consider a broader class of exploration strategies that satisfy the η-distinguishability property below.

Definition 4.4 (η-distinguishability)**.** We say an exploration policy $\pi_{\texttt{exp}} = \{\pi_{h,\texttt{exp}}\}_{h\in[H]}$ is η-distinguishable if there exists aconstant $\eta > 0$ such that for all $z, z' \in \mathcal{Z}$ with $z \neq z'$, it holds that $D^2_{\mathrm{H}}\left(\mathbb{P}_z^{\pi_{\texttt{exp}}}(\tau_H), \mathbb{P}_{z'}^{\pi_{\texttt{exp}}}(\tau_H)\right) \geq \eta$.

η-distinguishability implies the existence of an exploration policy $\pi_{\texttt{exp}}$ that could well-distinguish the models with an η-gap in Hellinger distance concerning the distribution of whole trajectory, which also impose condition over model seperation. Next, we introduce the assumption over priori.

Assumption 4.5 (Prior coverage)**.** There exists a constant $c_{\mathcal{Z}} > 0$ such that $\sup_{z,z'} \frac{\mathcal{P}_{\mathcal{Z}}(z')}{\mathcal{P}_{\mathcal{Z}}(z)} \leq c_{\mathcal{Z}}$.

The assumption asserts a bounded ratio of priors, implying that each $z \in \mathcal{Z}$ has a non-negligible prior probability. The assumption is intuitive, as a negligible priori suggests such a scenario almost surely does not occur, rendering the planning in such scenarios unnecessary. Now, we are ready to present the main theorem of Planner under perfect setting.

Theorem 4.6 (Regret under Perfect Setting)**.** *Suppose that Definition 4.1 and Assumption 4.5 hold. Given an η-distinguishable exploration policy $\pi_{\texttt{exp}}$ and $T \leq T_{\mathrm{p}}$, Algorithm 1 ensures*

$$\begin{aligned}\mathrm{Reg}_z(T) &:= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left[\mathcal{J}_z(\pi_z^*, \omega^t) - \mathcal{J}_z(\pi^t, \omega^t)\right] \\ &\leq \tilde{\mathcal{O}}\left(H^{\frac{3}{2}}\sqrt{T/\eta \cdot \log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})}\right),\end{aligned}$$

for any $z \in \mathcal{Z}$ and $\{\omega^t\}_{t\in[T]}$, if the Planner explores with probability $\epsilon = (H\log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.

Theorem 4.6 states that the Planner's algorithm can attain a $\tilde{\mathcal{O}}(\sqrt{T})$ regret for planning facilitated by LLMs. The multiplicative factor of the regret depends on the horizon of the interactive process H, the reciprocal of coverage rate η in Definition 4.4, and the logarithmic term $\log(c_{\mathcal{Z}}|\mathcal{Z}|)$ including both the cardinality of candidate models and the prior coverage in Assumption 4.5, which jointly characterizes the complexity of the physical world.

Remark 4.7. Lee et al. (2023) has demonstrated that a perfect decision-pretrained transformer, similar to the role of LLM in ours, can attain a $\tilde{\mathcal{O}}(H^{\frac{3}{2}}\sqrt{T})$ Bayesian regret, i.e., $\mathbb{E}_{z\sim\mathcal{P}_{\mathcal{Z}}}[\mathrm{Reg}(T)]$, via ICL. In comparison, we focus on a more challenging setting that aims to control the frequentist regret, which is closer to applications, and attain a comparable result with additional exploration.

5. Performance under Practical Setting

5.1. Pretraining Large Language Model

In this subsection, we elaborate on the pretraining of LLMs using transformer architecture. We employ a supervised learning algorithm minimizing the cross-entropy loss, i.e., $\widehat{\theta} = \mathrm{argmin}_{\theta\in\Theta}\, \mathcal{L}_{\mathrm{CE}}(\theta; \mathcal{D}_{\texttt{LLM}})$, as detailed in (6). Following this, the population risk follows that

$$\mathbb{E}_t[\mathbb{E}_{S_t}[D_{\mathrm{KL}}(\texttt{LLM}(\cdot|S_t) \,\|\, \texttt{LLM}_\theta(\cdot|S_t)) + \mathrm{Ent}(\texttt{LLM}(\cdot|S_t))]],$$

where $t \sim \mathrm{Unif}([\bar{T}_{\mathrm{p}}])$, S_t is distributed as the pretraining distribution, and $\mathrm{Ent}(\mathbb{P}) = \mathbb{E}_{x\sim\mathbb{P}}[\log \mathbb{P}(x)]$ is the Shannon entropy. As the minimum is achieved at $\texttt{LLM}_\theta(\cdot|S) = \texttt{LLM}(\cdot|S)$, estimated $\texttt{LLM}_{\widehat{\theta}}$ and $\texttt{LLM}$ are expected to converge under the algorithm with a sufficiently large dataset. Specifically, our design adopts a transformer function class to stay consistent with the architectural choices of language models like BERT and GPT. Specifically, a transformer model comprises D sub-modules, with each sub-module

incorporating a Multi-Head Attention (MHA) mechanism and a fully connected Feed-Forward (FF) layer. Refer to §A.3 for further details, and we specify two widely adopted assumptions in the theoretical analysis of LLM pretraining (Wies et al., 2023; Zhang et al., 2023).

Assumption 5.1 (Boundedness)**.** For all $z \in \mathcal{Z}$ and $t \leq \bar{T}_{\mathrm{p}}$, there exists an absolute constant $R > 0$ such that all $S_t = (\ell_1, \ldots, \ell_t) \sim \mathbb{P}_{\mathcal{D}}(\cdot \,|\, z)$ with $S_t \in \mathfrak{L}^*$ satisfies that $\|S_t\|_{2,\infty} \leq R$ almost surely.

The boundedness assumption requires that the ℓ_2-norm of the magnitude of each token is upper bounded by $R > 0$, and such an assumption holds in most settings.

Assumption 5.2 (Ambiguity)**.** For all latent variable $z \in \mathcal{Z}$, there exists a constant $c_0 > 0$ such that for all $\ell_{t+1} \in \mathfrak{L}$ and $S_t = (\ell_1, \ldots, \ell_t) \in \mathfrak{L}^*$ with length $t < \bar{T}_{\mathrm{p}}$, it holds $\mathbb{P}_{\mathcal{D}}(\ell_{t+1} \,|\, S_t, z) \geq c_0$.

The ambiguity assumption states that the generating distribution is lower bounded, and the assumption is grounded in reasoning as there may be multiple plausible choices for the subsequent words to convey the same meaning. Next, we present the performance of the pretrained LLMs.

Theorem 5.3 (Zhang et al. (2023))**.** *Suppose that Assumptions 5.1 and 5.2 hold. With probability at least $1 - \delta$, the pretrained model $\mathtt{LLM}_{\widehat{\theta}}$ by the algorithm in (5) satisfies that*

$$\begin{aligned}
&\bar{\mathbb{E}}_{\mathcal{D}_{\mathtt{LLM}}}\left[D_{\mathrm{TV}}\left(\mathtt{LLM}(\cdot \,|\, S), \mathtt{LLM}_{\widehat{\theta}}(\cdot \,|\, S)\right)\right] \\
&\leq \mathcal{O}\Bigg(\inf_{\theta^* \in \Theta} \sqrt{\bar{\mathbb{E}}_{\mathcal{D}_{\mathtt{LLM}}}\left[D_{\mathrm{KL}}\left(\mathtt{LLM}(\cdot \,|\, S), \mathtt{LLM}_{\theta^*}(\cdot \,|\, S)\right)\right]} + \\
&\frac{t_{\mathrm{mix}}^{1/4} \log \frac{1}{\delta}}{(N_{\mathrm{p}} \bar{T}_{\mathrm{p}})^{1/4}} + \sqrt{\frac{t_{\mathrm{mix}}}{N_{\mathrm{p}} \bar{T}_{\mathrm{p}}}} \left(\bar{D} \log\big(1 + \bar{B} N_{\mathrm{p}} \bar{T}_{\mathrm{p}}\big) + \log \frac{1}{\delta}\right)\Bigg)
\end{aligned}$$

where $\bar{B}$ and $\bar{D}$ features the tranformer's architecture, t_{mix} denotes the mixing time of Markov chain $\{S_t\}_{t \in [T]}$[2], and $N_{\mathrm{p}} \bar{T}_{\mathrm{p}}$ is the size of dataset $\mathcal{D}_{\mathtt{LLM}}$. See §A.3 for definitions.

Theorem 5.3 states that the total variation of the conditional distribution, with expectation taken over the average distribution of context S in $\mathcal{D}_{\mathtt{LLM}}$ (see Table 1 for definition), converges at $\mathcal{O}\left((N_{\mathrm{p}} \bar{T}_{\mathrm{p}})^{-1/2}\right)$. Note that the first two terms represent the approximation error and deep neural networks act as a universal approximator (Yarotsky, 2017) such that the error would vanish with increasing volume of network (Proposition C.4, Zhang et al., 2023). For notational simplicity, we denote the right-hand side of theorem as $\Delta_{\mathtt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)$.

[2]Note that $\{S_t^n\}_{t \in [T]}$ directly satisfies Markov property since $S_t^n = (\ell_1^n, \ldots, \ell_t^n)$ and thus $S_i^n \subseteq S_t^n$ for all $i \leq t$.

5.2. Pretraining Observation-to-Language Translator

In this subsection, we work on pretraining of observation-to-language translators under a self-supervised learning architecture using the contrastive loss. Given function class

$$\mathcal{F}_\gamma = \{f_\gamma(\cdot, \cdot) : \gamma \in \Gamma, \|f_\gamma\|_\infty \leq B_{\mathcal{F}}, \|1/f_\gamma\|_\infty \leq B_{\mathcal{F}}^-\},$$

with finite elements, and the contrastive loss $\mathcal{L}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\mathtt{Rep}})$ in (6) is defined over $\mathcal{F}_\gamma$. Note that the contrastive loss can be equivalently written as the negative log-likelihood loss of a binary discriminator, following that $\mathcal{L}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\mathtt{Rep}}) = \widehat{\mathbb{E}}_{\mathcal{D}_{\mathtt{Rep}}}\left[-\mathbb{D}_\gamma(y \,|\, o, s)\right]$, where we define

$$\mathbb{D}_\gamma(y \,|\, o, s) = \left(\frac{f_\gamma(o, s)}{1 + f_\gamma(o, s)}\right)^y \left(\frac{1}{1 + f_\gamma(o, s)}\right)^{1-y} \tag{7}$$

Based on the algorithm $\widehat{\gamma} = \mathrm{argmin}_{\gamma \in \Gamma} \mathcal{L}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\mathtt{Rep}})$, the population loss follows that

$$\mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbb{D}_\gamma(\cdot|o, s) \,\|\, \mathbb{D}(\cdot|o, s)\right) + \mathrm{Ent}(\mathbb{D}(\cdot|o, s))\right]. \tag{8}$$

As the minimum is attained at $\mathbb{D}_\gamma(\cdot \,|\, o, s) = \mathbb{D}(\cdot \,|\, o, s)$, where $\mathbb{D}(\cdot \,|\, o, s) := \mathbb{P}_{\mathcal{C}}(\cdot \,|\, o, s)$ is the distribution of the label conditioned on the (o, s) pair in contrastive data collection, estimated $\mathbb{D}_{\widehat{\gamma}}(\cdot \,|\, o, s)$ and $\mathbb{D}(\cdot \,|\, o, s)$ are expected to converge, and thus the learning target is the ground-truth likelihood ratio $f^*(o, s) = \mathbb{O}(o \,|\, s)/\mathcal{P}^-(o)$ (see Lemma D.2). We assume the learning target $f^*(o, s)$ is realizable in $\mathcal{F}_\gamma$, which is standard in literature (Qiu et al., 2022).

Assumption 5.4 (Realizability)**.** Given a designated negative distribution $\mathcal{P}^- \in \Delta(\mathcal{O})$, there exists $\gamma^* \in \Gamma$ such that $f_{\gamma^*}(o, s) = \mathbb{O}(o \,|\, s)/\mathcal{P}^-(o)$ for all $(o, s) \in \mathcal{O} \times \mathcal{S}$.

Next we present the performance of pretrained translator.

Theorem 5.5 (Pretrained Translator)**.** *Suppose that Assumption 5.4 holds. With probability at least $1 - \delta$, the pretrained model $\mathbb{O}_{\widehat{\gamma}}$ by the algorithm in (7) satisfies that*

$$\begin{aligned}
&\bar{\mathbb{E}}_{\mathcal{D}_{\mathtt{Rep}}}\left[D_{\mathrm{TV}}\left(\mathbb{O}(\cdot \,|\, s), \mathbb{O}_{\widehat{\gamma}}(\cdot \,|\, s)\right)\right] \\
&\quad \leq \mathcal{O}\left(\frac{B_{\mathcal{F}}(B_{\mathcal{F}}^-)^{1/2}}{(N_{\mathrm{p}} T_{\mathrm{p}} H)^{1/2}} \sqrt{\log(N_{\mathrm{p}} T_{\mathrm{p}} H |\mathcal{F}_\gamma| / \delta)}\right),
\end{aligned}$$

where let $\mathbb{O}_{\widehat{\gamma}}(\cdot \,|\, s) = f_{\widehat{\gamma}}(\cdot \,|\, s) \cdot \mathcal{P}^-(\cdot)$ and $|\mathcal{F}_\gamma|$ denotes the cardinality of the function class $\mathcal{F}_\gamma$.

Theorem 5.5 posits that the average expectation of the total variation of the translation distribution regarding $\mathcal{D}_{\mathtt{Rep}}$ converges at $\mathcal{O}\left((N_{\mathrm{p}} T_{\mathrm{p}})^{-1/2}\right)$. For notational simplicity, write the right-hand side of the theorem as $\Delta_{\mathtt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)$. Furthermore, the algorithm also ensures a more stringent convergence guarantee concerning χ^2-divergence:

$$\bar{\mathbb{E}}_{\mathcal{D}_{\mathtt{Rep}}}[\chi^2(\mathbb{O}_{\widehat{\gamma}}(\cdot \,|\, s) \,\|\, \mathbb{O}(\cdot \,|\, s))] \leq \Delta_{\mathtt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2.$$

5.3. Performance with Pretrained PAR System

In this subsection, we delve into the performance of task planning with pretrained PAR system. We first introduce the online coverage assumption, which pertains to the distribution of online planning trajectories under practical scenarios and trajectories in pretraining datasets.

Assumption 5.6 (Coverage)**.** There exists absolute constants $\lambda_S > 0$ and $\lambda_R > 0$ such that for all latent variable $z \in \mathcal{Z}$, $t < \bar{T}_{\mathrm{p}}$ and policy sequence $\{\pi^i\}_{i\le\lceil t/2H\rceil}$ from the Planner, it holds that (i) $\prod_{i=1}^{\lceil t/2H\rceil} \widehat{\mathbb{P}}_z^{\pi_i}(\tilde{S}_i) \le \lambda_S \cdot \bar{\mathbb{P}}_{\mathcal{D}_{\texttt{LLM}}}(S_t)$ for all ordered token sequence $S_t = (\tilde{S}_i)_{i\le\lceil t/2H\rceil} \in \mathfrak{L}^*$, where $|\tilde{S}_i| = 2H$ for all $k < \lceil t/2H\rceil$, and (ii) $\bar{\mathbb{P}}_{\mathcal{D}_{\texttt{Rep}}}(s) \ge \lambda_R$ for all state $s \in \mathcal{S}$.

Here, $\widehat{\mathbb{P}}_z$ denotes the distribution of the dynamic system with the pretrained translator. The assumption asserts that (i) distribution of the ICL prompts induced by policy sequences $\{\pi^i\}_{i\le\lceil t/2H\rceil}$ from the Planner under practical scenarios is covered by the pretraining data, where $\lceil t/2H\rceil$ denotes the number of episodes described in S_t. (ii) all states $s \in \mathcal{S}$ are covered by the average distribution of the Reporter's pretraining dataset. Similar conditions are adopted in ICL analysis (Zhang et al., 2023), decision pretrained transformer (Lee et al., 2023; Lin et al., 2023b) and offline RL (Munos, 2005; Duan et al., 2020). Intuitively, LLM and reporter cannot precisely plan or translate beyond the support of the pretraining dataset. These conditions are achievable if an explorative behavior strategy π^b is deployed with a sufficiently large N_{p} when collecting data. We then present the main theorem regarding the practical performance.

Theorem 5.7 (Regret under Practical Setting)**.** *Suppose that Assumptions 4.5, 5.1, 5.2, 5.4 and 5.6. Given an η-distinguishable exploration policy $\pi_{\texttt{exp}}$ and $T \le T_{\mathrm{p}}$, under the practical setting, the Planner's algorithm in Algorithm 1 ensures that*

$$\mathrm{Reg}_z(T) \le \tilde{\mathcal{O}}\Big(H^{\frac{3}{2}}\sqrt{T/\eta \cdot \log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})} + H^2T \cdot \Delta_{\mathrm{p}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, 1\sqrt{T}, \xi)\Big),$$

for any $z \in \mathcal{Z}$ and $\{\omega_t\}_{t\in[T]}$. The cumulative pretraining error of PAR system follows that

$$\Delta_{\mathrm{p}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi) = (\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 + 2\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) + \lambda_S \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta).$$

where $\xi = (\eta, \lambda_S, \lambda_R)$ are defined in Definition 4.4 and Assumption 5.6, and pretraining errors $\Delta_{\texttt{LLM}}$ and $\Delta_{\texttt{Rep}}$ are defined in Theorem 5.3 and Theorem 5.5. Under the practical setting, Planner should explore with probability

$$\epsilon = \big(H\log\big(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T}\big)/T\eta\big)^{1/2} + H(\eta\lambda_{\min})^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, 1/\sqrt{T})^2.$$

Theorem 5.7 reveals that, in comparison to perfect scenario, the Planner can achieve an approximate $\tilde{\mathcal{O}}(\sqrt{T})$ regret, but incorporating an additional pretraining error term that could diminishe with an increase in the volume of pretraining data. Besides, it underscores the necessity of exploration, where the Planner should explore with an additional $H(\eta\lambda_{\min})^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2$ to handle the mismatch between ground-truth and pretrained environment.

Remark 5.8. The challenge of establishing a performance guarantee in a practical setting arises from the mismatch between the ground-truth environment and the pretrained one, leading to a distributional shift in posterior probability. Besides, BAIL is realized through a pretrained LLM, which introduces its pretraining error inaddition. In response, we propose a novel regret decomposition and provide the convergence rate of posterior probability with bounded pretraining errors, distinguishing ours from the previous results in Lee et al. (2023); Liu et al. (2023).

Extentions. In §B.1, we discuss the design of the Planner by taking LLMs as World Model (WM). Here, the Planner prompts the LLM to predict the next observation rather than subgoals, alleviating the reliance on expert knowledge. By leveraging model-based RL methods like Monte Carlo Tree Search (MCTS) and Real-Time Dynamic Programming (RTDP), the Planner utilizes the LLM-simulated environment to optimize its strategies based on the contextual information. As shown in Proposition B.1, the simulated world model via ICL conforms to Bayesian Aggregated World Model (BAWM). Hence, the LLM Planner achieves $\mathrm{Reg}_z(T) \le \tilde{\mathcal{O}}(H\sqrt{T/\eta}) + H^2T \cdot \Delta_{\mathrm{p,wm}}$ under practical setting with regularity conditions (see Corollary B.3). Besides, we extend the results in §4 to accommodate the scenario of multi-agent collaboration, i.e., K Actors. In §B.2, we formulate the probelm as a cooperative hierarchical Markov Game (HMG) and establish a theoretical guarantee of $\mathrm{Reg}_z(T) \le \tilde{\mathcal{O}}(\sqrt{H^3TK/\eta})$ under perfect scenario (see Corollary B.4). The two extention correponds to recent works on LLM planning as world model (Hu & Shu, 2023) and muti-agent collaboration (Mandi et al., 2023).

6. Conclusion

In this work, we embedded the LLM-empowered decision making problem into a hierarchical RL, where at the high level, the LLM Planner decomposes the user-specified task into subgoals, and at the low level, the Actor(s) translate the linguistic subgoals into physical realizations while providing feedbacks for augmenting the planning through a trained reporter. Under the perfect setting, we characterize the BAIL nature of the LLM-aided planning pipeline and the nessecity of exploration even under the expert guidance. We also shed light on how the training errors of both LLM and reporter enter the ICL error under practical scenarios.

Impact Statement

This work presents a theoretical exploration into the mechanisms underpinning LLM-empowered agents, focusing on their potential to solve decision-making problems in the physical world. The immediate societal impact of our theoretical study may not be directly observable. The insights gained might be crucial for the future development of more effective, efficient, and ethical LLM-empowered agents. By advancing our understanding of how these agents function and interact with their environment, this research lays the groundwork for engineering efforts that can translate these theoretical foundations into practical applications.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Ahuja, K., Panwar, M., and Goyal, N. In-context learning through the bayesian prism. *arXiv preprint arXiv:2306.04891*, 2023.

Barto, A. G. and Mahadevan, S. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1-2):41–77, 2003.

Barto, A. G., Bradtke, S. J., and Singh, S. P. Learning to act using real-time dynamic programming. *Artificial intelligence*, 72(1-2):81–138, 1995.

Başar, T. and Olsder, G. J. *Dynamic noncooperative game theory*. SIAM, 1998.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Bonet, B. and Geffner, H. Planning as heuristic search. *Artificial Intelligence*, 129(1-2):5–33, 2001.

Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pp. 287–318. PMLR, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., McClelland, J., and Hill, F. Data distributional properties drive emergent few-shot learning in transformers. *arXiv preprint arXiv:2205.05055*, 2022.

Chane-Sane, E., Schmid, C., and Laptev, I. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning*, pp. 1430–1440. PMLR, 2021.

Dann, C., Mansour, Y., Mohri, M., Sekhari, A., and Sridharan, K. Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International conference on machine learning*, pp. 4666–4689. PMLR, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large timeiii. *Communications on pure and applied Mathematics*, 29 (4):389–461, 1976.

Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J. B., et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Edwards, A., Sahni, H., Schroecker, Y., and Isbell, C. Imitating latent policies from observation. In *International conference on machine learning*, pp. 1755–1763. PMLR, 2019.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., and Qiao, Y. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 910–919, 2024.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Ghallab, M., Nau, D., and Traverso, P. *Automated Planning: theory and practice*. Elsevier, 2004.

Grigsby, J., Fan, L., and Zhu, Y. Amago: Scalable in-context reinforcement learning for adaptive agents. *arXiv preprint arXiv:2310.09971*, 2023.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Hahn, M. and Goyal, N. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*, 2023.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. Bayesian model averaging: a tutorial. *Statistical science*, 14(4):382–417, 1999.

Honovich, O., Shaham, U., Bowman, S. R., and Levy, O. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*, 2022.

Hu, M., Mu, Y., Yu, X., Ding, M., Wu, S., Shao, W., Chen, Q., Wang, B., Qiao, Y., and Luo, P. Tree-planner: Efficient close-loop task planning with large language models. *arXiv preprint arXiv:2310.08582*, 2023.

Hu, Z. and Shu, T. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

Huang, W., Abbeel, P., Pathak, D., and Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pp. 9118–9147. PMLR, 2022.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jiang, H. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*, 2023.

Kidambi, R., Chang, J., and Sun, W. Mobile: Model-based imitation learning from observation alone. *Advances in Neural Information Processing Systems*, 34: 28598–28611, 2021.

Kim, H. J., Cho, H., Kim, J., Kim, T., Yoo, K. M., and Lee, S.-g. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*, 2022.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Lee, J. N., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., and Brunskill, E. Supervised pretraining can learn in-context reinforcement learning. *arXiv preprint arXiv:2306.14892*, 2023.

Li, B., Wu, P., Abbeel, P., and Malik, J. Interactive task planning with language models. *arXiv preprint arXiv:2310.10645*, 2023a.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., and Gao, J. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2, 2023b.

Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., Chen, T., Huang, D.-A., Akyürek, E., Anandkumar, A., et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022.

Lin, B. Y., Huang, C., Liu, Q., Gu, W., Sommerer, S., and Ren, X. On grounded planning for embodied tasks with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13192–13200, 2023a.

Lin, L., Bai, Y., and Mei, S. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023b.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

Liu, M., Zhu, M., and Zhang, W. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022a.

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., and Tang, J. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, 2022b.

Liu, Z., Hu, H., Zhang, S., Guo, H., Ke, S., Liu, B., and Wang, Z. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency. *arXiv preprint arXiv:2309.17382*, 2023.

Mandi, Z., Jain, S., and Song, S. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*, 2023.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Michel, P., Levy, O., and Neubig, G. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Munos, R. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pp. 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Nottingham, K., Ammanabrolu, P., Suhr, A., Choi, Y., Hajishirzi, H., Singh, S., and Fox, R. Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling. *arXiv preprint arXiv:2301.12050*, 2023.

OpenAI, R. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023.

Pateria, S., Subagdja, B., Tan, A.-h., and Quek, C. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

Paulin, D. Concentration inequalities for markov chains by marton couplings and spectral methods. 2015.

Polyanskiy, Y. and Wu, Y. Information theory: From coding to learning. *Book draft*, 2022.

Qiu, S., Wang, L., Bai, C., Yang, Z., and Wang, Z. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pp. 18168–18210. PMLR, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Reid, M., Yamada, Y., and Gu, S. S. Can wikipedia help offline reinforcement learning? *arXiv preprint arXiv:2201.12122*, 2022.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Schiappa, M. C., Rawat, Y. S., and Shah, M. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023.

Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., and Garg, A. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530. IEEE, 2023.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tokic, M. and Palm, G. Value-difference based exploration: adaptive control between epsilon-greedy and softmax. In *Annual conference on artificial intelligence*, pp. 335–346. Springer, 2011.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van Handel, R. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

Wang, X., Zhu, W., Saxon, M., Steyvers, M., and Wang, W. Y. Large language models are latent variable models: Explaining and finding good demonstrations for incontext learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Wang, Y., Jiao, R., Lang, C., Zhan, S. S., Huang, C., Wang, Z., Yang, Z., and Zhu, Q. Empowering autonomous driving with large language models: A safety perspective. *arXiv preprint arXiv:2312.00812*, 2023b.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022b.

Wies, N., Levine, Y., and Shashua, A. The learnability of in-context learning. *arXiv preprint arXiv:2303.07895*, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Xu, P., Zhu, X., and Clifton, D. A. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023a.

Yao, W., Heinecke, S., Niebles, J. C., Liu, Z., Feng, Y., Xue, L., Murthy, R., Chen, Z., Zhang, J., Arpit, D., et al. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*, 2023b.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Zhang, T. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023.

Appendix for "From Words to Actions: Unveiling the Theoretical Underpinnings of LLM-Driven Autonomous Systems"

A. Additional Background and Related Works

Notations. For some $n \in \mathbb{N}^+$, let $[n] = \{1, \ldots, n\}$. Denote $\Delta(\mathcal{X})$ as the probability simplex over $\mathcal{X}$. Consider two non-negative sequence $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$, if $\limsup a_n/b_n < \infty$, we write it as $a_n = \mathcal{O}(b_n)$ and use $\tilde{\mathcal{O}}$ to omit logarithmic terms. Else if $\liminf a_n/b_n < \infty$, we write $a_n = \Omega(b_n)$. For continuum $\mathcal{S}$, denote $|\mathcal{S}|$ as the cardinality. For matrix $X \in \mathbb{R}^{m\times n}$, the $\ell_{p,q}$-norm is defined as $\|X\|_{p,q} = (\sum_{i=1}^n \|X_{:,i}\|_p^q)^{1/q}$, where $X_{:,i}$ denotes the i-th column of X.

Table 1. Table of Notations.

Notation	Meaning
$\mathcal{J}_z(\cdot,\cdot), \pi_z^*(\cdot)$	value function and optimal policy $\pi_z^*(\cdot) := \arg\max_\pi \mathcal{J}(\pi,\cdot)$ concerning ground-truth $\mathbb{O}$
$\widehat{\mathcal{J}}_z(\cdot,\cdot), \widehat{\pi}_z^*(\cdot)$	value function and optimal policy $\widehat{\pi}_z^*(\cdot) := \arg\max_\pi \widehat{\mathcal{J}}(\pi,\cdot)$ concerning pretrained $\mathbb{O}_{\widehat{\gamma}}$
$\mathbb{P}_{\mathcal{D}}(\cdot), \mathbb{P}_{\mathcal{C}}(\cdot)$	probability induced by the distribution of joint and contrastive data collection
$\pi_{h,\texttt{LLM}}^t, \widehat{\pi}_{h,\texttt{LLM}}^t$	$\pi_{h,\texttt{LLM}}^t(\cdot \mid \tau_h^t, \omega^t) = \texttt{LLM}(\cdot \mid \texttt{pt}_h^t)$ and $\widehat{\pi}_{h,\texttt{LLM}}^t(\cdot \mid \tau_h^t, \omega^t) = \texttt{LLM}_{\widehat{\theta}}(\cdot \mid \texttt{pt}_h^t)$ at step h
$\mathbb{P}_z(\cdot), \widehat{\mathbb{P}}_z(\cdot)$	probability under environment featured by z, ground-truth $\mathbb{O}$ or pretrained $\mathbb{O}_{\widehat{\gamma}}$
$\mathbb{P}_z^\pi(\cdot), \widehat{\mathbb{P}}_z^\pi(\cdot)$	probability under environment featured by z, policy π, ground-truth $\mathbb{O}$ or pretrained $\mathbb{O}_{\widehat{\gamma}}$
$\mathcal{P}_\Omega(\cdot), \mathcal{P}_{\mathcal{Z}}(\cdot)$	prior distributions of high-level tasks and latent variables
$\breve{\tau}_{h/t}^i$	$\breve{\tau}_{h/t}^i = \tau_H$ for all $i < t$ and $\breve{\tau}_{h/t}^t = \tau_h$
$\mathbb{P}_z(\cdot \mid \cdot, \mathbf{do}\,\cdot)$	$\mathbb{P}_z(\cdot \mid o_1, \mathbf{do}\, g_{1:h-1}) = \int_{o_{2:h-1}} \prod_{h'=1}^{h-1} \mathbb{P}_z\left(o_{h'+1} \mid (o,g)_{1:h'}\right) \mathrm{d}o_{2:h-1}$
$\mathbb{P}_{\texttt{LLM}}^t(\cdot \mid \cdot, \mathbf{do}\,\cdot)$	$\mathbb{P}_{\texttt{LLM}}^t(\cdot \mid o_1, \mathbf{do}\, g_{1:h-1}) := \int_{o_{2:h-1}} \prod_{h'=1}^{h-1} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1} \mid (o,g)_{1:h'}, \mathcal{H}_t\right) \mathrm{d}o_{2:h-1}$
$\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\cdot,\cdot), \widehat{\pi}_{\texttt{LLM}}^{t,*}(\cdot)$	value function of environment simulated by $\texttt{LLM}_{\widehat{\theta}}$ and $\widehat{\pi}_{\texttt{LLM}}^{t,*}(\cdot) := \arg\max_\pi \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi,\cdot)$
$\mathcal{J}_{t,\texttt{LLM}}(\cdot,\cdot), \pi_{\texttt{LLM}}^{t,*}(\cdot)$	value function of environment simulated by perfect $\texttt{LLM}$ and $\pi_{\texttt{LLM}}^{t,*}(\cdot) := \arg\max_\pi \mathcal{J}_{t,\texttt{LLM}}(\pi,\cdot)$
$\mathbb{P}_{\texttt{LLM}}^t(\cdot), \widehat{\mathbb{P}}_{\texttt{LLM}}^t(\cdot)$	probability of environment simulated by perfect $\texttt{LLM}$ or pretrained $\texttt{LLM}_{\widehat{\theta}}$ with $\mathcal{H}_t$
$D_{\mathrm{TV}}(P,Q)$	total variation distance, $D_{\mathrm{TV}}(P,Q) := 1/2 \cdot \mathbb{E}_{x\sim P}[\lvert \mathrm{d}Q(x)/\mathrm{d}P(x) - 1\rvert]$
$D_{\mathrm{H}}^2(P,Q)$	Helliger distance, $D_{\mathrm{H}}^2(P,Q) := 1/2 \cdot \mathbb{E}_{x\sim P}\left[\left(\sqrt{\mathrm{d}Q(x)/\mathrm{d}P(x)} - 1\right)^2\right]$
$D_{\mathrm{KL}}(P,Q)$	KL divergence, $D_{\mathrm{KL}}(P\|Q) := \mathbb{E}_{x\sim P}[\log \mathrm{d}P(x)/\mathrm{d}Q(x)]$
$\chi^2(P,Q)$	χ^2-divergence, $\chi^2(P\|Q) := \mathbb{E}_{x\sim P}[(\mathrm{d}Q(x)/\mathrm{d}P(x) - 1)^2]$
$\widehat{\mathbb{E}}_{\mathcal{D}}[f]$	$\bar{\mathbb{E}}[f] := 1/n \cdot \sum_{t=1}^n f(\ell_t)$ given dataset $\mathcal{D} = \{\ell_t\}_{t\in[n]}$
$\bar{\mathbb{P}}_{\mathcal{D}}(\cdot), \bar{\mathbb{E}}_{\mathcal{D}}[f]$	$\bar{\mathbb{P}}_{\mathcal{D}}(\cdot) := \sum_{n=1}^N \sum_{t=0}^{T-1} \mathbb{P}_{\mathcal{D}}(\cdot \mid \ell_{1:t}^n)/NT$ and $\bar{\mathbb{E}}[f] := \mathbb{E}_{\ell\sim\bar{\mathbb{P}}_{\mathcal{D}}}[f(\ell)]$ given $\mathcal{D} = \{\ell_{1:T}^n\}_{n\in[N]}$

Elaboration on Contrastive Learning. As an example, the noise contrastive estimation (NCE, Gutmann & Hyvärinen, 2010) is one of the most widely adopted objectives in contrastive representation learning. From the theoretical lens, to estimate unnormalized model p_d with $x_i \overset{\text{iid}}{\sim} p_d$, additional noise data is sampled from a reference distribution p_n and the estimation is gained by maximizing $\widehat{\mathbb{E}}[y \cdot \log(h_\gamma(x)) + (1-y) \cdot \log(1 - h_\gamma(x))]$, where $y = \mathbb{1}(x \text{ is not noise})$ and $h^*(x) = p_d(x)/(p_d(x)+p_n(x))$. With slight modifications, we use a function class $\mathcal{F}$ to approximate the ratio p_d/p_n rather than the relative probability h itself. In practice, the most commonly used contrastive training objectives are variations of NCE and originated from the NLP domain (Schiappa et al., 2023) by sharing the same idea of *minimizing the distance between the positive pair and maximizing the distance between the negative pairs*.

A.1. More Related Works

Autorgressive Large Language Models and In-Context Learning. Most commercial Large Language Models (LLMs), including ChatGPT (Brown et al., 2020), GPT-4 (OpenAI, 2023), Llama (Touvron et al., 2023), and Gemini (Team et al.,

2023), operate in an autoregressive manner. These LLMs exhibit robust reasoning capabilities, and a crucial aspect of their reasoning prowess is the in-context learning (ICL) ability. This ability is further enhanced through additional training stages (Iyer et al., 2022), careful selection and arrangement of informative demonstrations (Liu et al., 2021; Kim et al., 2022), explicit instruction (Honovich et al., 2022), and the use of prompts to stimulate chain of thoughts (Wei et al., 2022b). Theoretical understanding of ICL is an active area of research. Since the real-world datasets used for LLM pretraining are difficult to model theoretically and are very large, ICL has also been studied in stylized setups (Xie et al., 2021; Garg et al., 2022; Chan et al., 2022; Hahn & Goyal, 2023; Zhang et al., 2023). In this paper, we build upon the framework attributing the ICL capability to Bayesian inference (Xie et al., 2021; Jiang, 2023; Zhang et al., 2023), a series of practical experiments, including Wang et al. (2023a); Ahuja et al. (2023), provide empirical support for this Bayesian statement. Our work leverages the ICL ability of LLM with detailed examination under the Bayesian framework.

LLM-empowered Agents. In the task-planning and decision-making problems, symbolic planners have commonly been employed to transform them into search problems (Bonet & Geffner, 2001; Ghallab et al., 2004) or to design distinct reinforcement learning or control policies for each specific scenario. Recent empirical studies have shifted towards leveraging LLMs as symbolic planners in various domains, including robotic control (Mandi et al., 2023; Brohan et al., 2023; Li et al., 2023a; Du et al., 2023), autonomous driving (Wang et al., 2023b; Fu et al., 2024) and personal decision assistance (Li et al., 2022; Lin et al., 2023a; Hu et al., 2023; Liu et al., 2023; Nottingham et al., 2023). Another recent line of research has been dedicated to devising diverse prompting schemes to enhance the reasoning capability of LLMs (Wei et al., 2022b; Yao et al., 2023a;b; Hao et al., 2023). Despite their considerable empirical success, there is a lack of comprehensive theoretical analysis of LLM Agent. In this paper, we formalize the problem into a general framework with a hierarchical structure and design RL-like prompts to facilitate planning with provable performance. Two recent works by Liu et al. (2023) and Lee et al. (2023) also aim to establish provable algorithms for planning with LLMs or Decision-pretrained Transformers (DPT). In comparison, we discuss both the plausibility of taking LLMs as an action generator (Lee et al., 2023) and simulated world model (Liu et al., 2023). Furthermore, we provide a statistical guarantee for pretrained models and conduct a detailed examination of the algorithm's performance in practical settings, bringing our analysis closer to real-world applications.

A.2. Hierarchical Markov Decision Process

In this subsection, we present a formalized definition of the HMDP model introduced in §3.1.

Low-level MDP. Define $\mathcal{G}$ as the space of high-level actions. For fixed $g \in \mathcal{G}$ and high-level step $h \in [H]$, the low-level MDP is defined as $\mathcal{M}_h(g) = (\mathcal{S}, \mathcal{A}, H_a, \mathbb{T}_h, \bar{r}_g)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the low-level action space, H_a is the number of steps, $\mathbb{T}_h = \{\mathbb{T}_{h,\bar{h}}\}_{\bar{h}\in[H_a]}$ is the transition kernel, and $\bar{r} = \{\bar{r}_{\bar{h}}\}_{\bar{h}\in[H_a]}$ is the reward function with $\bar{r}_{\bar{h}} : \mathcal{S}\times\mathcal{A}\times\mathcal{G} \mapsto \mathbb{R}$. The low-level agent follows policy $\mu = \{\mu_g\}_{g\in\mathcal{G}}$, where $\mu_g = \{\mu_{\bar{h}}\}_{\bar{h}\in[H_a]}$ and $\mu_{\bar{h}} : \mathcal{S}\times\mathcal{G} \mapsto \Delta(\mathcal{A})$.

High-level POMDP. Define Ω be the space of disclosed variables, and we write $z = (\mathbb{T}, \mu)$ to feature the low-level environment. Each low-level episode corresponds to a single high-level action. Given fixed pair $(z, \omega) \in \mathcal{Z} \times \Omega$, the POMDP is characterized by $\mathcal{W}(z, \omega) = (\mathcal{S}, \mathcal{O}, \mathcal{G}, H, \mathbb{P}_z, r_\omega)$, where $\mathcal{O}$ is the observation space, $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H]}$ is the emission distribution with $\mathbb{O}_h : \mathcal{O} \mapsto \Delta(\mathcal{S})$, $r = \{r_h\}_{h\in[H]}$ is the reward function with $r_h : \mathcal{O} \times \Omega \mapsto \mathbb{R}$, and $\mathbb{P}_z = \{\mathbb{P}_{z,h}\}_{h\in[H]}$ is the high-level transition kernel following that

$$\mathbb{P}_{z,h}(s' \,|\, s, g) = \mathbb{P}\big(\bar{s}_{h,H_a+1} = s' \,|\, \bar{s}_{h,1} = s, a_{h,1:\bar{h}} \sim \mu_g, \bar{s}_{h,2:\bar{h}+1} \sim \mathbb{T}_h\big),$$

for all $h \in [H]$. The space of state $\mathcal{S}$ and latent variable z are inherited from the low-level MDP.

Furthermore, for the high-level POMDP, the state value function of policy π, the state value function is defined as

$$V^{\pi}_{z,h}(s, \tau, \omega) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(o_{h'}, \omega) \Big| s_h = s, \tau_h = \tau \right], \tag{9}$$

where trajectory $\tau_h \in (\mathcal{O} \times \mathcal{G})^{h-1} \times \mathcal{O}$, and similarly we define the state-action value function as

$$Q^{\pi}_{z,h}(s, \tau, g, \omega) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(o_{h'}, \omega) \Big| s_h = s, \tau_h = \tau, g_h = g \right], \tag{10}$$

where expectation is taken concerning the policy π, transition kernel $\mathbb{P}_z$, and emission distribution $\mathbb{O}$. Besides, for all $h \in [H]$, denote the probability of observing trajectory τ_h under policy π as

$$\mathbb{P}_z^\pi(\tau_h) = \pi(\tau_h) \cdot \mathbb{P}_z(\tau_h), \quad \mathbb{P}_z(\tau_h) = \prod_{h'=1}^{h-1} \mathbb{P}(o_{h'+1} \,|\, \tau_{h'}, g_{h'}), \quad \pi(\tau_h) = \prod_{h'=1}^{h-1} \pi_h(g_{h'} \,|\, \tau_{h'}), \tag{11}$$

where $\mathbb{P}_z(\tau_h)$ denotes the part of the probability of τ_h that is incurred by the dynamic environment independent of policies, $\pi(\tau_h)$ denotes the part that can be attributed to the randomness of policy.

A.3. LLM Pretraining under Transformer Architecture

Transformer and Attention Mechanism. Consider a sequence of N input vectors $\{\mathbf{h}_i\}_{i=1}^n \subset \mathbb{R}^d$, written as an input matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]^\top \in \mathbb{R}^{n \times d}$, where each $\mathbf{h}_i$ is a row of $\mathbf{H}$ (also a token). Consider $\mathbf{K} \in \mathbb{R}^{n_s \times d}$ and $\mathbf{V} \in \mathbb{R}^{n_s \times d_s}$, then the (softmax) attention mechanism maps these input vectors using the function $\texttt{attn}(\mathbf{H}, \mathbf{K}, \mathbf{V}) = \texttt{Softmax}(\mathbf{H}\mathbf{K}^\top)\mathbf{V} \in \mathbb{R}^{n \times d_s}$, where softmax function is applied row-wisely and normalize each vector via the exponential function such that $[\texttt{Softmax}(\mathbf{h})]_i = \exp(\mathbf{h}_i)/\sum_{j=1}^d \exp(\mathbf{h}_j)$ for all $i \in [d]$. To approximate sophisticated functions, practitioners use Multi-head Attention (MHA) instead, which forwards the input vectors into h attention modules in parallel with $h \in \mathbb{N}$ as a hyperparameter and outputs the sum of these sub-modules. Denote $\mathbf{W} = \{(\mathbf{W}_i^H, \mathbf{W}_i^K, \mathbf{W}_i^V)\}_{i=1}^h$ as the set of weight matrices, the MHA outputs $\texttt{Mha}(\mathbf{H}, \mathbf{W}) = \sum_{i=1}^h \texttt{attn}(\mathbf{H}\mathbf{W}_i^H, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V)$, where $\mathbf{W}_i^H \in \mathbb{R}^{d \times d_h}$, $\mathbf{W}_i^K \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_i^V \in \mathbb{R}^{d \times d}$ for all $i \in [h]$, and d_h is usually set to d/h (Michel et al., 2019). Based on the definitions above, we are ready to present the transformer architecture employed in LLMs like BERT and GPT (Devlin et al., 2018; Brown et al., 2020). Detailedly, the transformer network has D sub-modules, consisting of an MHA and Feed-Forward (FF) fully-connected layer. Given input matrix $\mathbf{H}^{(0)} = \mathbf{H} \in \mathbb{R}^{n \times d}$, in the j-th layer for $j \in [D]$, it first takes the output from the $(t-1)$-th layer $\mathbf{H}^{(t-1)}$ as the input matrix, and forwards it to the MHA module with a projection function $\texttt{Proj}[\cdot]$ and a residual link. After receiving intermediate $\overline{\mathbf{H}}^{(t)} \in \mathbb{R}^{n \times d}$, the FF module maps each row through a same single-hidden layer neural network with d_F neurons such that $\texttt{ReLU}(\overline{\mathbf{H}}^{(t)}\mathbf{A}_1^{(t)})\mathbf{A}_2^{(t)}$, where $\mathbf{A}_1^{(t)} \in \mathbb{R}^{d \times d_F}$, $\mathbf{A}_2^{(t)} \in \mathbb{R}^{d_F \times d}$, and $[\texttt{ReLU}(\mathbf{X})]_{i,j} = \max\{\mathbf{X}_{i,j}, 0\}$. Specifically, the output of the t-th layer with $t \in [D]$ can be summarized as below:

$$\overline{\mathbf{H}}^{(t)} = \texttt{Proj}\left[\texttt{Mha}\left(\mathbf{H}^{(t-1)}, \mathbf{W}^{(t)}\right) + \gamma_1^{(t)}\mathbf{H}^{(t-1)}\right], \quad \mathbf{H}^{(t)} = \texttt{Proj}\left[\texttt{ReLU}(\overline{\mathbf{H}}^{(t)}\mathbf{A}_1^{(t)})\mathbf{A}_2^{(t)} + \gamma_2^{(t)}\overline{\mathbf{H}}^{(t)}\right],$$

where $\gamma_1^{(t)}$ and $\gamma_2^{(t)}$ features the allocation of residual link. The final output of the transformer is the probability of the next token via a softmax distribution such that

$$\mathbf{H}^{(D+1)} = \texttt{Softmax}\left(\mathbf{1}^\top \mathbf{H}^{(D)}\mathbf{A}^{(D+1)}/N\gamma^{(D+1)}\right),$$

where $\mathbf{A}^{(D+1)} \in \mathbb{R}^{d \times d_E}$ denotes the weight matrix with dimension $d_E \in \mathbb{N}$ and $\gamma^{(D+1)} \in (0, 1]$ is the fixed temperature parameter. Let $\boldsymbol{\theta}^{(t)} = \left(\mathbf{W}^{(t)}, \mathbf{A}^{(t)}, \boldsymbol{\gamma}^{(t)}\right)$ for all $t \in [D]$, where $\mathbf{A}^{(t)} = (\mathbf{A}_1^{(t)}, \mathbf{A}_2^{(t)})$ and $\boldsymbol{\gamma}^{(t)} = (\gamma_1^{(t)}, \gamma_2^{(t)})$, and denote $\boldsymbol{\theta}^{(D+1)} = (\mathbf{A}^{(D+1)}, \gamma)$. Hence, the parameter of the whole transformer architecture is the concatenation of parameters in each layer such that $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(D+1)})$, and we consider a bounded parameter space, which is defined as below

$$\begin{aligned}\boldsymbol{\Theta} := \{\boldsymbol{\theta} \,|\, &\|\mathbf{A}_1^{(t)}\|_F \le B_{A,1}, \|\mathbf{A}_2^{(t)}\|_F \le B_{A,2}, \|\mathbf{A}^{(D+1),\top}\|_{1,2} \le B_A, |\gamma_1^{(t)}| \le 1, |\gamma_2^{(t)}| \le 1, \\ &|\gamma^{(D+1)}| \le 1, \|\mathbf{W}_i^{H,(t)}\| \le B_H, \|\mathbf{W}_i^{K,(t)}\| \le B_K, \|\mathbf{W}_i^{V,(t)}\| \le B_V, \forall (i,t) \in [h] \times [D]\}.\end{aligned}$$

To facilitate the expression of Theorem 5.3, we further define $\bar{D} = D^2 d \cdot (d_h + d_F + d) + d_E \cdot d$ and $\bar{B} = \gamma^{-1} R h B_{A,1} B_{A,2} B_A B_H B_K B_V$, where R is (almost surely) the upper bound of the magnitude of each token $\ell \in \mathfrak{L}$ in token sequence $S_t \in \mathfrak{L}^*$, which is defined in Assumption 5.1.

Markov Chains. We follow the notations used in Paulin (2015); Zhang et al. (2023). Let Ω be a Polish space. The transition kernel for a time-homogeneous Markov chain $\{X_i\}_{i=1}^\infty$ supported on Ω is a probability distribution $\mathbb{P}(x, y)$ for every $x \in \Omega$. Given $X_1 = x_1, \cdots, X_{t-1} = x_{t-1}$, the conditional distribution of X_t equals $\mathbb{P}(x_{t-1}, y)$. A distribution π is said to be a stationary distribution of this Markov chain if $\int_{x \in \Omega} \mathbb{P}(x, y) \cdot \pi(x) = \pi(y)$. We adopt $\mathbb{P}_t(x, \cdot)$ to denote the distribution of X_t conditioned on $X_1 = x$. The mixing time of the chain is defined by

$$d(t) = \sup_{x \in \Omega} D_{\mathrm{TV}}\big(\mathbb{P}_t(x, \cdot), \pi\big), \quad t_{\mathrm{mix}}(\varepsilon) = \min\{t \,|\, d(t) \le \varepsilon\}, \quad t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4). \tag{12}$$

Algorithm 2 Planning with PAR System - Planner with LLM as WM

Input: Policy $\pi_{\texttt{exp}}$ with $\eta \in (0,1)$, parameter $c_{\mathcal{Z}} > 0$, $N_{\mathrm{s}} \in \mathbb{N}$, and $|\mathcal{Z}| \in \mathbb{N}$,
  and reward function $\{r_h\}_{h\in[H]}$ specified by the human user.
Initialize: $\mathcal{H}_0 \leftarrow \emptyset$, $\mathcal{D}_t^{\mathrm{s}} \leftarrow \emptyset, \forall t \in [T]$, and $\epsilon \leftarrow (H\log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.
for episode t from 1 to T **do**
  Receive the high-level task ω^t from the human user.
  Sample $\mathcal{I}_t \sim \text{Bernuolli}(\epsilon)$.
  for stimulation n from 1 to N_{s} **do**
    Sample $\mathbf{g}_n^{t,\mathrm{s}} \sim \text{Unif}(\mathcal{G}^H)$ and set $\texttt{pt}_{1,n}^t \leftarrow \mathcal{H}_t \cup \{o_1^t, g_{1,n}^{t,\mathrm{s}}\}$.
    for step h from 1 to H **do**
      Set $\texttt{pt}_{h,n}^t \leftarrow \mathcal{H}_t \cup \{o_{1,n}^t, g_{1,n}^{t,\mathrm{s}}, \ldots, o_{h,n}^{t,\mathrm{s}}, g_{h,n}^{t,\mathrm{s}}\}$.
      Predict $\tilde{o}_{h+1,n}^t \sim \texttt{LLM}(\cdot \,|\, \texttt{pt}_{h,n}^t)$ via prompting LLM.
    end for
    Update $\mathcal{D}_t^{\mathrm{s}} \leftarrow \mathcal{D}_t^{\mathrm{s}} \cup \{o_{1,n}^t, g_{1,n}^{t,\mathrm{s}}, \ldots, o_{H-1,n}^{t,\mathrm{s}}, g_{H-1,n}^{t,\mathrm{s}}, o_{H,n}^{t,\mathrm{s}}\}$.
  end for
  for step h from 1 to H **do**
    Collect the observation o_h^t from the Reporter.
    Calculate $\pi_{\texttt{LLM}}^t \leftarrow \textsc{Optimal-planning}(\omega^t, \mathcal{D}_t^{\mathrm{s}}, \{r_h\})$
    Sample $g_h^t \sim (1-\mathcal{I}_t)\cdot \pi_{h,\texttt{LLM}}^t(\cdot \,|\, \omega^t, \tau_h^t) + \mathcal{I}_t \cdot \pi_{h,\texttt{exp}}^t(\cdot \,|\, \tau_h^t)$.
    Send the subgoal g_h^t to the Actor.
  end for
  Update $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{\omega^t, \tau_H^t\}$.
end for

B. Extentions

B.1. LLM Planning via Bayesian Aggregated World Model

Recall that the pretraining algorithm in §3.2 also equips LLM with the capability to predict observation generation, i.e., $\mathbb{P}_h(o_h \,|\, (o,g)_{1:h-1})$. Existing literature has shown the benefits of augmenting the reasoning process with predicted world states, as it endows LLMs with a more grounded inference without reliance on expert knowledge (Hu & Shu, 2023). Specifically, the Planner interactively prompts LLM to internally simulate entire trajectories grounded on historical feedback. By leveraging model-based RL methods such as Monte Carlo Tree Search (Browne et al., 2012) and Real-Time Dynamic Programming (Barto et al., 1995), the Planner utilizes the LLM-simulated environment to optimize its strategies. The planning protocol is as follows: at the beginning of t-th episode, Planner iteratively prompts LLM with initial observation o_1, history $\mathcal{H}_t$, and subgoals $g_{1:H}$ sequentially to predict observations $o_{1:H}$. Subsequently, a simulation dataset $\mathcal{D}_t^{\mathrm{s}}$ is collected, allowing the Planner to compute the optimal policy with rewards specified by the human users, using methods such as MCTS. We first show that the LLM-simulated environment conforms to a Bayesian Aggregated World Model (BAWM), and is formalized as follows.

Proposition B.1 (LLM as BAWM)**.** *Assume that the distribution of pretraining data is given by (4). Under the perfect setting in Definition 4.1, for each $(h,t) \in [H]\times[T]$, the LLM serves as a Bayesian aggregated world model, following that*

$$\mathbb{P}_{\texttt{LLM}}^t(\cdot \,|\, o_1, \mathbf{do}\, g_{1:h-1}) = \sum_{z\in\mathcal{Z}} \mathbb{P}_z\left(\cdot \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \mathcal{H}_t\right), \tag{13}$$

where the marginal distributions are defined as $\mathbb{P}_z(\cdot \,|\, o_1, \mathbf{do}\, g_{1:h-1}) = \int_{o_{2:h-1}} \prod_{h'=1}^{h-1} \mathbb{P}_z\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) \mathrm{d}o_{2:h-1}$ and $\mathbb{P}_{\texttt{LLM}}^t(\cdot \,|\, o_1, \mathbf{do}\, g_{1:h-1}) = \int_{o_{2:h-1}} \prod_{h'=1}^{h-1} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) \mathrm{d}o_{2:h-1}$.

Proof of Propoition B.1. Please refer to §E.1 for a detailed proof. □

Note that the generation distribution $\mathbb{P}_{\texttt{LLM}}^t(\cdot \,|\, (o,g)_{1:h}) = \texttt{LLM}(\cdot \,|\, (o,g)_{1:h}, \mathcal{H}_t)$ is non-stationary, since $\mathbb{P}_{\mathcal{D}}(z \,|\, (o,g)_{1:h}, \mathcal{H}_t)$ fluctuates with simulated part $(o,g)_{1:h}$ due to the autoregressive manner of LLMs. Instead, Proposition B.1 posits that the

Algorithm 3 Multi-Agent Planning with PAR System - Planner

Input: Policy $\pi_{\texttt{exp}}$ with $\eta \in (0,1)$, parameter $c_{\mathcal{Z}} > 0$, and $|\mathcal{Z}| \in \mathbb{N}$.
Initialize: $\mathcal{H}_0 \leftarrow \emptyset$, and $\epsilon \leftarrow (HK\log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.
for episode t from 1 to T **do**
  Receive the high-level task ω^t from the human user.
  Sample $\mathcal{I}_t \sim$ Bernuolli(ϵ).
  for step h from 1 to H **do**
    Collect the observation o_h^t from Reporter.
    for Actor k from 1 to K **do**
      Set $\texttt{pt}_{h,k}^t \leftarrow \mathcal{H}_t \cup \{\omega^t, o_1^t, \mathbf{g}_1^t, \ldots, o_h^t, k\}$.
      Sample $g_{h,k,\texttt{LLM}}^t \sim \texttt{LLM}(\cdot \,|\, \texttt{pt}_{h,k}^t)$ via prompting LLM.
    end for
    If $\mathcal{I}_t = 1$ **then** $\mathbf{g}_h^t \leftarrow \mathbf{g}_{h,\texttt{LLM}}^t$, **else** sample $\mathbf{g}_h^t \sim \pi_{h,\texttt{exp}}(\cdot \,|\, \tau_h^t)$.
  end for
  Send the subgoal $\mathbf{g}_h^t$ to the Actors.
  Update $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{\omega^t, \tau_H^t\}$.
end for

marginal distribution has a stationary expression based on posterior aggregation. Akin to Assumption 5.6, we introduce the coverage assumption.

Assumption B.2 (Strong Coverage)**.** There exists absolute constants $\lambda_{S,1}, \lambda_{S,2}$ and λ_R such that for all $z \in \mathcal{Z}$, length $t < \bar{T}_{\mathrm{p}}$ and policy sequence $\{\pi^i\}_{i \le \lfloor t/2H \rfloor}$ from the Planner, it holds that (i) $\prod_{i=1}^{\lfloor t/2H \rfloor} \widehat{\mathbb{P}}_z^{\pi_i}(\tilde{S}_i) \le \lambda_{S,1} \cdot \bar{\mathbb{P}}_{\mathcal{D}_{\texttt{LLM}}}((\tilde{S}_i)_{i \le \lfloor t/2H \rfloor})$ and $\bar{\mathbb{P}}_{\mathcal{D}_{\texttt{LLM}}}(\tilde{S}_{\lceil t/2H \rceil} | (\tilde{S}_i)_{i \le \lfloor t/2H \rfloor}) \ge \lambda_{S,2}$ for all ordered token sequence $S_t = (\tilde{S}_i)_{i \le \lceil t/2H \rceil} \in \mathfrak{L}^*$, where $|\tilde{S}_i| = 2H$ for all $k < \lceil t/2H \rceil$, (ii) $\bar{\mathbb{P}}_{\mathcal{D}_{\texttt{Rep}}}(s) \ge \lambda_R$ for all $s \in \mathcal{S}$.

We remark that Assumption B.2 imposes a stronger condition over the coverage, particularly on the in-episode trajectory $\tilde{S}_{\lceil t/2H \rceil}$, Here, $\lceil t/2H \rceil$ denotes the number of episodes described in S_t. The demand of the stronger assumption arises from LLM now serving as a WM, necessitating more extensive information across all kinds of scenarios. Suppose that the Planner can learn optimal policy $\widehat{\pi}_{\texttt{LLM}}^{t,*} = \mathrm{argmax}_{\pi \in \Pi} \widehat{\mathcal{J}}_{\texttt{LLM}}^t(\pi, \omega)$ with sufficiently large simulation steps $|\mathcal{D}_t^{\mathrm{s}}|$, where $\widehat{\mathcal{J}}_{\texttt{LLM}}^t$ denotes the value function concerning $\texttt{LLM}_{\widehat{\theta}}$ and history $\mathcal{H}_t$. Akin to Algorithm 1, the planning algorithm by taking LLM as WM includes an ϵ-greedy exploration with η-distinguishable $\pi_{\texttt{exp}}$. The pseudocode is in Algorithm 2. The following corollary presents the performance under practical settings.

Corollary B.3 (Regret under Practical Setting with LLM as World Model)**.** *Suppose that Assumptions 4.5, 5.1, 5.2, 5.4 and 5.6. Given an η-distinguishable exploration policy $\pi_{\texttt{exp}}$ and $T \le T_{\mathrm{p}}$, under the practical setting, the Planner's algorithm in Algorithm 2 ensures that*

$$\mathrm{Reg}_z(T) \le \tilde{\mathcal{O}}\Big(H\sqrt{T/\eta \cdot \log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})} + H^2 T \cdot \Delta_{\mathrm{p,wm}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, 1/\sqrt{T}, \xi)\Big),$$

for any $z \in \mathcal{Z}$ and $\{\omega_t\}_{t \in [T]}$. The cumulative pretraining error of the PAR system follows

$$\begin{aligned}\Delta_{\mathrm{p,wm}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi) &= 2(\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \\ &\quad + 2\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) + 2\lambda_{S,1}\lambda_{S,2}^{-1} \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta).\end{aligned}$$

where $\xi = (\eta, \lambda_{S,1}, \lambda_{S,2}, \lambda_R)$ are defined in Definition 4.4 and Assumption 5.6, and the errors $\Delta_{\texttt{LLM}}$ and $\Delta_{\texttt{Rep}}$ are defined in Theorem 5.3 and Theorem 5.5. Under the practical setting, the Planner should explore with probability $\epsilon = (\log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2} + H(\eta\lambda_{\min})^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, 1/\sqrt{T})^2$.

Proof of Corollary B.3. Please refer to §E.2 for a detailed proof. □

B.2. LLM-Empowered Multi-Agent Collaboration

To characterize the multi-agent interactive process, i.e., several Actors, of task planning, we consider a turn-based *cooperative* hierarchical Markov Game (HMG), corresponding to HMDP in §3.1. Instead, HMG consists of a low-level language-conditioned Markov Game (MG) and a high-level language-conditioned cooperative Partially Observable Markov Game

(POMG). To extend this framework, we introduce the following modifications: (i) low-level MG: let $\mathcal{K} = [K]$ be the set of Actors, and $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_K$ and $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ be the space of subgoals and low-level actions. Low-level Actors conduct planning following a joint policy $\mu = \{\mu_h\}_{h\in[H]}$ with $\mu_h : \mathcal{S} \times \mathcal{G} \mapsto \Delta(\mathcal{A})$, where $\{\mu_{h,k}\}_{k\in\mathcal{K}}$ can be correlated, e.g., within zero-sum game, Stackelberg game (Başar & Olsder, 1998). (ii) high-level POMG: under cooperation, assume that policies can be factorized as

$$\pi_h(\mathbf{g}_h \,|\, \tau_{h-1}, \omega) = \prod_{k=1}^{K} \pi_{h,k}(g_{h,k} \,|\, \tau_{h-1}, \omega), \quad \forall h \in [H].$$

The remaining concepts are consistent with HMDP. Here, the Planner assumes the role of *central controller* and solves a fully-cooperative POMG that aims to maximize a shared value function. Thus, the Planner should infer both the Actors' intentions, i.e., joint policy μ, and the environment, i.e., transition kernel $\mathbb{T}$, from the historical context, and then assign subgoal for each Actor.

Specifically, the LLM's recommendations are obtained by invoking the ICL ability of LLMs with the history-dependent prompt akin to (1) sequentially for each Actor. For the k-th Actor, prompt LLM with $\mathtt{pt}_{h,k}^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t, k\}$, where denote $\mathcal{H}_t = \bigcup_{i=1}^{t-1}\{\omega^i, \tau_H^i\}$ and $\tau_h^t = \{o_h^1, \mathbf{g}_h^1, \ldots, o_h^t\}$. Under the perfect setting (see Definition 4.1), LLM's joint policy for recommendations follows:

$$\pi_{h,\mathtt{LLM}}^t\big(\mathbf{g}_h^t \,|\, \tau_h^t, \omega^t\big) = \prod_{k\in\mathcal{K}} \left(\sum_{z\in\mathcal{Z}} \pi_{z,h,k}^* \left(g_{h,k}^t \,|\, \tau_h^t, \omega^t\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \mathtt{pt}_h^t\right) \right), \tag{14}$$

which is akin to Proposition 4.2 and the proof of the statement is provided in §E.3. The pseudocode is presented in Algorithm 3. Then, we give the performance guarantee under multi-agent scenarios with the perfect PAR system.

Corollary B.4 (Multi-agent Collaboration Regret under Perfect Setting)**.** *Suppose that Assumptions 4.1 and 4.5 hold. Given an η-distinguishable exploration policy $\pi_{\mathtt{exp}}$ and $T \le T_{\mathrm{p}}$, the Planner's algorithm in Algorithm 3 guarantees that*

$$\mathrm{Reg}_z(T) \le \tilde{\mathcal{O}}\left(H^{\frac{3}{2}} \sqrt{TK/\eta \cdot \log\left(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T}\right)} \right),$$

for any $z \in \mathcal{Z}$ and $\{\omega^t\}_{t\in[T]}$, if Planner explores with $\epsilon = \left(HK \log\left(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T}\right)/T\eta\right)^{1/2}$.

Proof of Corollary B.4. Please refer to §E.3 for a detailed proof. □

Corollary B.4 is akin to Theorem 4.6 with an additional $\sqrt{K}$ in regret. Besides, the multi-agent space of latent variable $|\mathcal{Z}| = |\mathcal{Z}_{\mathbb{T}}| \times |\mathcal{Z}_{\mu,\mathrm{m}}|$, where $\mathcal{Z}_{\mu,\mathrm{m}}$ is the space of joint policy, is generally larger than the single-agent space. Specifically, if responses are uncorrelated, then we have $\log|\mathcal{Z}_{\mu,\mathrm{m}}| = K \log|\mathcal{Z}_{\mu,\mathrm{s}}|$, resulting in a $\sqrt{K}$ times larger regret. The proof of extension to practical setting is akin to Corollary B.4 based on derivations in Theorem 5.7, and is omitted.

C. Proof for Section 4: Perfect Setting

C.1. Proof of Proposition 4.2

Proof of Proposition 4.2. Note that for all $h \in [H]$ and $t \in [T]$, we have

$$\begin{aligned}
\pi_{h,\mathtt{LLM}}^t\left(g_h^t \,|\, \tau_h^t, \omega^t\right) &= \sum_{z\in\mathcal{Z}} \mathbb{P}_{\mathcal{D}}\left(g_h^t \,|\, \mathtt{pt}_h^t, z\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \mathtt{pt}_h^t\right) \\
&= \sum_{z\in\mathcal{Z}} \mathbb{P}_{\mathcal{D}}\left(g_h^t \,|\, \mathcal{H}_t, \tau_h^t, \omega^t, z\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \mathtt{pt}_h^t\right) \\
&= \sum_{z\in\mathcal{Z}} \pi_{z,h}^*\left(\cdot \,|\, \tau_h^t, \omega^t\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \mathtt{pt}_h^t\right),
\end{aligned} \tag{15}$$

where the second equation results from the law of total probability, the third equation follows the definition of prompts in (1), and the last equation results from the generation distribution. □

C.2. Proof of Theorem 4.6

Proof of Thereom 4.6. Recall that the Planner takes a mixture policy of $\pi_{\texttt{exp}}$ and $\pi_{\texttt{LLM}}$ such that

$$\pi_h^t(\cdot\,|\,\tau_h^t,\omega^t) \sim (1-\epsilon)\cdot\pi_{h,\texttt{LLM}}^t(\cdot\,|\,\tau_h^t,\omega^t) + \epsilon\cdot\pi_{h,\texttt{exp}}(\cdot\,|\,\tau_h^t), \tag{16}$$

and Proposition 4.2 indicates that LLM's recommended policies take the form:

$$\begin{aligned}\pi_{h,\texttt{LLM}}^t\left(\cdot\,|\,\tau_h^t,\omega^t\right) &= \sum_{z\in\mathcal{Z}}\pi_{z,h}^*\left(\cdot\,|\,\tau_h^t,\omega^t\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z\,|\,\texttt{pt}_h^t\right),\\ \text{where } \texttt{pt}_h^t &= \mathcal{H}_t\cup\tau_h^t \text{ with } \mathcal{H}_t=\left\{\omega^i,\tau_H^i\right\}_{i\in[t-1]},\end{aligned} \tag{17}$$

for all $(h,t)\in[H]\times[T]$. Following (16), given $z\in\mathcal{Z}$ and $\{\omega^t\}_{t\in[T]}$, the regret is decomposed as

$$\begin{aligned}\mathrm{Reg}(T) &= \underbrace{\sum_{t=1}^T\sum_{h=1}^H\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}}\mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}}\left[\left(\pi_{z,h}^*-\pi_{h,\texttt{exp}}\right)Q_{z,h}^*(s_h^t,\tau_h^t,\omega^t)\right]\cdot\epsilon}_{\textbf{(i)}}\\ &\quad+\underbrace{\sum_{t=1}^T\sum_{h=1}^H\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}}\mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}}\left[\left(\pi_{z,h}^*-\pi_{h,\texttt{LLM}}^t\right)Q_{z,h}^*(s_h^t,\tau_h^t,\omega^t)\right]\cdot(1-\epsilon)}_{\textbf{(ii)}}\\ &\le\sum_{t=1}^T\sum_{h=1}^H\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}}\mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}}\left[\left(\pi_{z,h}^*-\pi_{h,\texttt{LLM}}^t\right)Q_{z,h}^*(s_h^t,\tau_h^t,\omega^t)\right]+HT\epsilon,\end{aligned} \tag{18}$$

where the second equation results from performance difference lemma (PDL, see Lemma F.4), and we write $\pi_hQ_h(s_h,\tau_h,\omega)=\langle\pi_h(\cdot|\tau_h,\omega),Q_h(s_h,\tau_h,\cdot,\omega)\rangle_{\mathcal{G}}$, and $\mathbb{P}_z^{\pi}(\tau_h)$ is defined in (11). Based on Lemma C.1, with probability at least $1-\delta$, the following event $\mathcal{E}_1$ holds: for all $(h,t)\in[H]\times[T]$,

$$\sum_{z'\in\mathcal{Z}}\sum_{i\in[t]}D_{\mathrm{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i),\mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big)\cdot\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{pt}_h^t)\le 2\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right), \tag{19}$$

where the randomness is incurred by $\texttt{pt}_h^t$ and define $\breve{\tau}_{h/t}^i=\tau_H$ for all $i\in[t-1]$ and $\breve{\tau}_{h/t}^t=\tau_h$ for notational simplicity. Suppose that event $\mathcal{E}_1$ in (19) holds, and denote $\mathcal{X}_{\texttt{exp}}^t=\{i\in[t]:\pi^i=\pi_{\texttt{exp}}\}$ as the set of exploration episodes. Note that for all $(h,t,z')\in[H]\times[T]\times\mathcal{Z}$, it holds that

$$\sum_{i\in[t]}D_{\mathrm{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i),\mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big)\ge\sum_{i\in\mathcal{X}_{\texttt{exp}}^{t-1}}D_{\mathrm{H}}^2\left(\mathbb{P}_z^{\pi_{\texttt{exp}}}(\tau_H),\mathbb{P}_{z'}^{\pi_{\texttt{exp}}}(\tau_H)\right)\ge\eta\cdot|\mathcal{X}_{\texttt{exp}}^{t-1}|, \tag{20}$$

where the last inequality results from $\pi_{\texttt{exp}}$ is η-distinguishable (see Definition 4.4) and the fact that $D_{\mathrm{H}}^2(P,Q)\le 1$ for all $P,Q\in\Delta(\mathcal{X})$. Combine (19) and (20), we can get

$$\sum_{z'\ne z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{pt}_h^t)\le\min\left\{2\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)\eta^{-1}/|\mathcal{X}_{\texttt{exp}}^{t-1}|,1\right\}, \tag{21}$$

for all $(h,t)\in[H]\times[T]$. Recall that (17) indicates that for all $(h,t)\in[H]\times[T]$, we have

$$\left(\pi_{z,h}^*-\pi_{h,\texttt{LLM}}^t\right)(\cdot\,|\,\tau_h,\omega)=\sum_{z'\ne z}\left(\pi_{z,h}^*-\pi_{z',h}^*\right)(\cdot\,|\,\tau_h,\omega)\cdot\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{pt}_h^t).$$

Based on Proposition 4.2 and conditioned on $\mathcal{E}_1$, it holds that

$$\begin{aligned}&\sum_{t=1}^T\sum_{h=1}^H\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}}\mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}}\left[\left(\pi_{z,h}^*-\pi_{h,\texttt{LLM}}^t\right)Q_{z,h}^*(s_h^t,\tau_h^t,\omega^t)\right]\\ &\quad\le H\cdot\sum_{t=1}^T\sum_{h=1}^H\sum_{z'\ne z}\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}}\mathbb{E}_{\tau_h^t\sim\mathbb{P}_z^{\pi^t}}\left[\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{pt}_h^t)\right]\\ &\quad\le 2\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)H\eta^{-1}\cdot\sum_{t=1}^T\sum_{h=1}^H\mathbb{E}\left[\min\left\{1/|\mathcal{X}_{\texttt{exp}}^{t-1}|,1\right\}\right],\end{aligned} \tag{22}$$

Note that $\mathbb{1}(\pi^t = \pi_{\texttt{exp}}) \overset{\text{iid}}{\sim} \text{Bernuolli}(\epsilon)$ for all $t \in [T]$. Besides,the following event $\mathcal{E}_2$ holds:

$$\sum_{t=1}^{T} \min\left\{1/|\mathcal{X}_{\texttt{exp}}^{t-1}|, 1\right\} \leq \mathcal{O}(\epsilon^{-1}\log(T\log T/\delta)). \tag{23}$$

with probability at least $1-\delta$ based on Lemma F.5. Combine (18), (22) and (23), we have

$$\begin{aligned}
\text{Reg}_z(T) &\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}} \mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t\right) Q_{z,h}^*(s_h,\tau_h,\omega^t)\mathbb{1}\left(\mathcal{E}_1\cap\mathcal{E}_2 \text{ holds}\right)\right] \\
&\quad + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}} \mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t\right) Q_{z,h}^*(s_h,\tau_h,\omega^t)\mathbb{1}\left(\mathcal{E}_1\cap\mathcal{E}_2 \text{ fails}\right)\right] + HT\epsilon \\
&\leq \mathcal{O}\Big(\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)H^2\log(T\log T/\delta)\cdot(\eta\epsilon)^{-1} + HT\epsilon + 2HT\delta\Big) \\
&\leq \tilde{\mathcal{O}}\left(H^{\frac{3}{2}}\sqrt{\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)T/\eta}\right),
\end{aligned}$$

where we choose to expolre with probability $\epsilon = \left(H\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)/T\eta\right)^{1/2}$. If we take $\delta = 1/\sqrt{T}$ in the arguments above, then we can conclude the proof of Theorem 4.6. □

C.3. Proof of Lemma C.1

Lemma C.1. *Suppose that Assumptions 4.1 and 4.5 hold. Given $\delta \in (0,1)$ and ground-truth $z \in \mathcal{Z}$, for all $(h,t) \in [H]\times[T]$, with probability at least $1-\delta$, it holds*

$$\sum_{z'\in\mathcal{Z}}\sum_{i\in[t]} D_{\text{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big)\cdot\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{pt}_h^t) \leq 2\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right),$$

where denote $\breve{\tau}_{h/t}^i = \tau_H$ for all $i < t$ and $\breve{\tau}_{h/t}^t = \tau_h$, and $\mathbb{P}_z^{\pi}(\tau_h)$ is defined in (11).

Proof of Lemma C.1. The proof is rather standard (e.g., see (Geer, 2000)). Let $\mathfrak{F}_t$ be the filtration induced by $\{\omega^i, \tau_H^i\}_{i<t}\cup\{\mathbb{1}(\pi^i = \pi_{\text{exp}})\}_{i\in[t]}$. For all $(h,t,z') \in [H]\times[T]\times\mathcal{Z}$, with probability at least $1-\delta$, the information gain concerning z' satisfies that

$$L_{h,t}(z') = \sum_{i=1}^{t}\log\left(\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right) \leq 2\log\mathbb{E}_{\mathfrak{F}_{1:t}}\left[\exp\left(\frac{1}{2}\sum_{i=1}^{t}\log\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right)\right] + 2\log(|\mathcal{Z}|/\delta), \tag{24}$$

where the inequality follows Lemma F.1 with $\lambda = 1/2$ and a union bound taken over $\mathcal{Z}$. Besides,

$$\mathbb{E}_{\mathfrak{F}_{1:t}}\left[\exp\left(\frac{1}{2}\sum_{i=1}^{t}\log\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right)\right] = \prod_{i=1}^{t}\left(1 - D_{\text{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big)\right). \tag{25}$$

Combine (24), (25) and fact that $\log(1-x) \leq -x$ for all $x \leq 1$, it holds that

$$L_{h,t}(z') \leq -2\sum_{i=1}^{t} D_{\text{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big) + 2\log(|\mathcal{Z}|/\delta), \tag{26}$$

with probability greater than $1-\delta$. Based on the Donsker-Varadhan representation in Lemma F.2 and duality principle, we have $\log\mathbb{E}_Q[e^f] = \sup_{P\in\Delta(\mathcal{X})}\{\mathbb{E}_P[f] - D_{\text{KL}}(P\,\|\,Q)\}$, where the supremum is taken at $P(x) \propto \exp(f(x))\cdot Q(x)$. Please refer to Lemma 4.10 in (Van Handel, 2014) for detailed proof. Based on the arguments above, for all $(h,t,P) \in [H]\times[T]\times\Delta(\mathcal{Z})$, it holds

$$\sum_{z'\in\mathcal{Z}} L_{h,t}(z')\cdot P(z') - D_{\text{KL}}\big(P\,\|\,\mathcal{P}_{\mathcal{Z}}\big) \leq \sum_{z'\in\mathcal{Z}} L_{h,t}(z')\cdot\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{pt}_h^t) - D_{\text{KL}}\big(\mathbb{P}_{\mathcal{D}}(\cdot\,|\,\texttt{pt}_h^t)\,\|\,\mathcal{P}_{\mathcal{Z}}\big). \tag{27}$$

since $\mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{pt}_h^t) \propto \exp\left(L_{h,t}(z')\right) \cdot \mathcal{P}_{\mathcal{Z}}(z')$ for all $(h,t) \in [H] \times [T]$. Denote $\delta_z(\cdot)$ as the Dirac distribution over the singleton z. Following this, by taking $P = \delta_z$ in (27), we have

$$\sum_{z' \in \mathcal{Z}} L_{h,t}(z') \cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{pt}_h^t) \geq D_{\mathrm{KL}}\big(\mathbb{P}_{\mathcal{D}}(\cdot \,|\, \mathtt{pt}_h^t) \,\|\, \mathcal{P}_{\mathcal{Z}}\big) + \log \mathcal{P}_{\mathcal{Z}}(z) \geq \log \mathcal{P}_{\mathcal{Z}}(z), \tag{28}$$

where the first inequality uses $D_{\mathrm{KL}}(\delta_z(\cdot) \,\|\, \mathcal{P}_{\mathcal{Z}}(\cdot)) = -\log \mathcal{P}_{\mathcal{Z}}(z)$ based on the definitions. Therefore, for all $(h,t) \in [H] \times [T]$, with probability at least $1-\delta$, it holds that

$$\sum_{z' \in \mathcal{Z}} \sum_{i \in [t]} D_{\mathrm{H}}^2\big(\mathbb{P}_z^{\pi^i}(\check{\tau}_{h/t}^i), \mathbb{P}_{z'}^{\pi^i}(\check{\tau}_{h/t}^i)\big) \cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{pt}_h^t) \leq -\frac{1}{2} \sum_{z' \in \mathcal{Z}} L_{h,t}(z') \cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{pt}_h^t) + \log\left(|\mathcal{Z}|/\delta\right) \leq 2 \log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right),$$

where the first inequality results from (25), and the last inequality follows (28) and Assumption 4.5, which indicates that $1/\mathcal{P}_{\mathcal{Z}}(z) \leq c_{\mathcal{Z}}|\mathcal{Z}|$. Thus, we conclude the proof of Lemma C.1. □

C.4. Proof of Proposition 4.3

Our construction of the hard-to-distinguish example is a natural extension to the hard instance for the contextual bandit problem in Proposition 1 (Zhang, 2022).

Proof of Proposition 4.3. Suppose that the high-level POMDP is fully observable, i.e., $\mathbb{O}(s) = s$, with $H = 2$ and $|\Omega|$=1. Consider $\mathcal{S} = \{s_1, s_2, s_3\}$ with rewards $r(s_1) = 0.5$, $r(s_2) = 1$, $r(s_3) = 0$, $\mathcal{G} = \{g_1, g_2\}$, and $\mathcal{Z} = \{z_1, \ldots, z_N\}$. Starting from initial state s_1, the transition kernel follows

$$\begin{cases} \mathbb{P}_{z_i}(s_1 \,|\, s_1, g_1) = 1, & \mathbb{P}_{z_i}(s_2 \,|\, s_1, g_1) = 0, & \mathbb{P}_{z_i}(s_3 \,|\, s_1, g_1) = 0, & \forall i \in [N], \\ \mathbb{P}_{z_1}(s_1 \,|\, s_1, g_2) = 0, & \mathbb{P}_{z_1}(s_1 \,|\, s_1, g_2) = 1, & \mathbb{P}_{z_1}(s_3 \,|\, s_1, g_2) = 0, & \text{if } i = 1, \\ \mathbb{P}_{z_i}(s_1 \,|\, s_1, g_2) = 0, & \mathbb{P}_{z_i}(s_2 \,|\, s_1, g_2) = p_i, & \mathbb{P}_{z_i}(s_3 \,|\, s_1, g_2) = 1 - p_i, & \text{if } i \neq 1, \end{cases}$$

where $p_i = 0.5(1 - \frac{i}{N})$ for all $i \in [N]$. For latent environment z_1, the optimal policy is $\pi_{z_1,1}^*(s_1) = g_2$ and $\pi_{z_i,1}^*(s_1) = g_1$ if $i \neq 1$. Suppose that prior distribution $\mathcal{P}_{\mathcal{Z}}$ is uniform. At $t = 1$, without any information, the posterior $\mathbb{P}(\cdot \,|\, \mathtt{pt}_1)$ degenerates to prior $\mathcal{P}_{\mathcal{Z}}(\cdot) = \mathrm{Unif}_{\mathcal{Z}}(\cdot)$. Hence, the LLM's policy at first step follows that $\pi_{\mathtt{LLM}}(\cdot \,|\, s_1) = \big(1 - \frac{1}{N}\big) \cdot \delta_{g_1}(\cdot) + \frac{1}{N} \cdot \delta_{g_2}(\cdot)$. Since $\mathbb{P}_{z_i}(s_1 \,|\, s_1, g_1) = 1$ and $\mathbb{P}_{z_i}(s_2 \,|\, s_1, g_1) = \mathbb{P}_{z_i}(s_3 \,|\, s_1, g_1) = 0$ for all $i \in [N]$, taking subgoal g_1 provides no information to differentiate z_i from others, and the posterior remains uniform. Such situation, i.e., $\mathbb{P}(\cdot \,|\, \mathtt{pt}_t) = \mathrm{Unif}_{\mathcal{Z}}(\cdot)$, ends only if the LLM suggests taking g_2 at some epsiode t. Consider the hard trajectory $\tau_{\mathrm{hard}} = \{s_1, g_1, s_1\}_{t \in [T]}$, where LLM consistently adheres to the initial $\pi_{\mathtt{LLM}}$ and keeps recommending subgoal g_1. Thus, we have $\mathbb{P}_{z_1}(\tau_{\mathrm{hard}}) = (1 - 1/N)^T$, indicating $\mathrm{Reg}_{z_1}(T) \geq 0.5T \cdot (1 - 1/N)^T$. □

D. Proof for Section 5: Practical Setting

D.1. Proof of Theorem 5.5

Proof of Theorem 5.5. Recall that the binary discriminator for label $y \in \{0, 1\}$ is defined as

$$\mathbb{D}_\gamma(y \,|\, o, s) := \left(\frac{f_\gamma(o,s)}{1 + f_\gamma(o,s)}\right)^y \left(\frac{1}{1 + f_\gamma(o,s)}\right)^{1-y},$$

and the contrastive learning algorithm in (6) follows $\widehat{\gamma} = \mathrm{argmax}_{\gamma \in \Gamma} \widehat{\mathbb{E}}_{\mathcal{D}_{\mathtt{Rep}}}\big[\log \mathbb{D}_\gamma(y \,|\, o, s)\big]$, and thus $f_{\widehat{\gamma}}$ is the maximum likelihood estimator (MLE) concerning the dataset $\mathcal{D}_{\mathtt{Rep}}$. Based on Lemma F.3, the MLE-type algorithm ensures that, with probability at least $1 - \delta$, it holds that

$$\bar{\mathbb{E}}_{(o,s) \sim \mathcal{D}_{\mathtt{Rep}}}\left[D_{\mathrm{TV}}^2\left(\mathbb{D}_{\widehat{\gamma}}(\cdot \,|\, o, s), \mathbb{D}(\cdot \,|\, o, s)\right)\right] \leq 2 \log(N_{\mathrm{p}} T_{\mathrm{p}} H |\mathcal{F}_\gamma|/\delta)/N_{\mathrm{p}} T_{\mathrm{p}} H, \tag{29}$$

where $\mathbb{D}(\cdot\,|\,o,s) = \mathbb{D}_{\gamma^*}(\cdot\,|\,o,s)$ with $f_{\gamma^*} = f^* \in \mathcal{F}_\gamma$ denotes the ground-truth discriminator based on the realizability in Assumption 5.4. Based on the definition of total variation, it holds that

$$\begin{aligned} &D_{\mathrm{TV}}^2\left(\mathbb{D}_{\widehat{\gamma}}(\cdot\,|\,o,s), \mathbb{D}(\cdot\,|\,o,s)\right) \\ &= \left(\frac{f_{\widehat{\gamma}}(o,s) - f^*(o,s)}{(1+f_{\widehat{\gamma}}(o,s))(1+f^*(o,s))}\right)^2 \leq \frac{1}{(1+R_{\mathcal{F}})^2}\left(\frac{f_{\widehat{\gamma}}(o,s) - f^*(o,s)}{1+f^*(o,s)}\right)^2 \\ &= \frac{1}{(1+R_{\mathcal{F}})^2}\left(\frac{\mathbb{O}_{\widehat{\gamma}}(o\,|\,s) - \mathbb{O}(o\,|\,s)}{\mathcal{P}^-(o) + \mathbb{O}(o\,|\,s)}\right)^2 = \frac{1}{(1+R_{\mathcal{F}})^2}\left(\frac{\bar{\mathbb{O}}_{\widehat{\gamma}}(o\,|\,s) - \bar{\mathbb{O}}(o\,|\,s)}{\bar{\mathbb{O}}(o\,|\,s)}\right)^2, \end{aligned} \tag{30}$$

where the first inequality results from $\|f\|_\infty \leq R_{\mathcal{F}}$ for all $f \in \mathcal{F}_\gamma$, the third equation arise from the definition that $\mathbb{O}_\gamma(\cdot|s) = f_\gamma(\cdot, s)\cdot\mathcal{P}^-(\cdot)$, and we write $\bar{\mathbb{O}}(\cdot\,|\,s) = \frac{1}{2}\left(\mathbb{O}(\cdot\,|\,s) + \mathcal{P}^-(\cdot)\right), \bar{\mathbb{O}}_\gamma(\cdot\,|\,s) = \frac{1}{2}\left(\mathbb{O}_\gamma(\cdot\,|\,s) + \mathcal{P}^-(\cdot)\right)$. Moreover, $\bar{\mathbb{O}}(\cdot|s)$ represents the marginal distribution derived from the joint distribution $\mathbb{P}_{\mathcal{C}}$ of collected dataset $\mathcal{D}_{\texttt{Rep}}$ (see data collection process in §3.2), as follows:

$$\begin{aligned} \mathbb{P}_{\mathcal{C}}(o\,|\,s) &= \mathbb{P}_{\mathcal{C}}(o\,|\,s, y=0)\cdot\mathbb{P}_{\mathcal{C}}(y=0\,|\,s) + \mathbb{P}_{\mathcal{C}}(o\,|\,s,y=1)\cdot\mathbb{P}_{\mathcal{C}}(y=1\,|\,s) \\ &= \mathbb{P}_{\mathcal{C}}(o\,|\,s,y=0)\cdot\mathbb{P}_{\mathcal{C}}(y=0) + \mathbb{P}_{\mathcal{C}}(o\,|\,s,y=1)\cdot\mathbb{P}_{\mathcal{C}}(y=1) := \bar{\mathbb{O}}(o\,|\,s), \end{aligned} \tag{31}$$

where the second equation results from the fact that contrastive data are labeled independent of data itself such that $\mathbb{P}_{\mathcal{C}}(s\,|\,y) = \mathbb{P}_{\mathcal{C}}(s)$ for all $y \in \{0,1\}$. Based on (31), we can get

$$\bar{\mathbb{E}}_{(o,s)\sim\mathcal{D}_{\texttt{Rep}}}\left[\left(\frac{\bar{\mathbb{O}}_{\widehat{\gamma}}(o\,|\,s) - \bar{\mathbb{O}}(o\,|\,s)}{\bar{\mathbb{O}}(o\,|\,s)}\right)^2\right] = \bar{\mathbb{E}}_{s\sim\mathcal{D}_{\texttt{Rep}}}\left[\mathbb{E}_{o\sim\bar{\mathbb{O}}(\cdot\,|\,s)}\left[\left(\frac{\bar{\mathbb{O}}_{\widehat{\gamma}}(\cdot\,|\,s) - \bar{\mathbb{O}}(\cdot\,|\,s)}{\bar{\mathbb{O}}(\cdot\,|\,s)}\right)^2\right]\right], \tag{32}$$

where equations results from the fact that $\mathbb{P}_{\mathcal{C}}(o,s) = \bar{\mathbb{O}}(o\,|\,s)\cdot\mathbb{P}_{\mathcal{C}}(s)$ and definition of χ^2-divergence. Therefore, combine (30) and (32), it holds that

$$\bar{\mathbb{E}}_{(o,s)\sim\mathcal{D}_{\texttt{Rep}}}\left[D_{\mathrm{TV}}^2\left(\mathbb{D}_{\widehat{\gamma}}(\cdot\,|\,o,s), \mathbb{D}(\cdot\,|\,o,s)\right)\right] \leq \frac{1}{(1+R_{\mathcal{F}})^2}\cdot\bar{\mathbb{E}}_{s\sim\mathcal{D}_{\texttt{Rep}}}\left[\chi^2\left(\bar{\mathbb{O}}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\bar{\mathbb{O}}(\cdot\,|\,s)\right)\right]. \tag{33}$$

Based on the variational representation of f-divergenve (§7.13, Polyanskiy & Wu, 2022), we have

$$\begin{aligned} \chi^2\left(\bar{\mathbb{O}}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\bar{\mathbb{O}}(\cdot\,|\,s)\right) &= \sup_{g:\mathcal{O}\mapsto\mathbb{R}}\left\{\frac{\left(\mathbb{E}_{\bar{\mathbb{O}}_{\widehat{\gamma}}}[g(o)\,|\,s] - \mathbb{E}_{\bar{\mathbb{O}}}[g(o)\,|\,s]\right)^2}{\mathrm{Var}_{\bar{\mathbb{O}}}[g(o)\,|\,s]}\right\} \\ &= \sup_{g:\mathcal{O}\mapsto\mathbb{R}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\gamma}}}[g(o)\,|\,s] - \mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{4\cdot\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\cdot\frac{\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}{\mathrm{Var}_{\bar{\mathbb{O}}}[g(o)\,|\,s]}\right\} \\ &\geq \sup_{\substack{g:\mathcal{O}\mapsto\mathbb{R},\\ \mathbb{E}_{\mathbb{O}}[g(o)|s]=0}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\gamma}}}[g(o)|s] - \mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{4\cdot\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\cdot\frac{\mathbb{E}_{\mathbb{O}}[g(o)^2\,|\,s]}{\mathbb{E}_{\bar{\mathbb{O}}}[g(o)^2\,|\,s]}\right\}, \end{aligned} \tag{34}$$

where the second equation follows the defintions of $\bar{\mathbb{O}}(\cdot\,|\,s)$ and $\bar{\mathbb{O}}_{\widehat{\gamma}}(\cdot\,|\,s)$, and the inequality results from $\mathrm{Var}_{\bar{\mathbb{O}}}[g(o)|s] = \mathbb{E}_{\bar{\mathbb{O}}}[g(o)^2|s]$ if $\mathbb{E}_{\mathbb{O}_{\widehat{\gamma}}}[g(o)|s]$=0. Furthermore, note that

$$\frac{\mathbb{E}_{\mathbb{O}}[g(o)^2\,|\,s]}{\mathbb{E}_{\bar{\mathbb{O}}}[g(o)^2\,|\,s]} = 2\left(1 + \frac{\mathbb{E}_{\mathcal{P}^-}[g(o)^2\,|\,s]}{\mathbb{E}_{\mathbb{O}}[g(o)^2\,|\,s]}\right)^{-1} \leq 2\left(1 + \left\|\frac{\mathcal{P}^-(\cdot)}{\mathbb{O}(\cdot\,|\,s)}\right\|_\infty\right)^{-1} \leq 2(1+B_{\mathcal{F}}^-)^{-1}, \tag{35}$$

as $\mathcal{P}^-(\cdot)/\mathbb{P}(\cdot|s) = f^* \in \mathcal{F}_\gamma$ and $\|1/f\|_\infty \leq B_{\mathcal{F}}^-$ for all $f \in \mathcal{F}$ under the realizability in Assumption 5.4. Besides, it holds that

$$\begin{aligned} \sup_{\substack{g:\mathcal{O}\mapsto\mathbb{R},\\ \mathbb{E}_{\mathbb{O}}[g(o)|s]=0}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\gamma}}}[g(o)\,|\,s] - \mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\right\} &= \sup_{g:\mathcal{O}\mapsto\mathbb{R}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\gamma}}}[g(o)\,|\,s] - \mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\right\} \\ &= \chi^2\left(\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right), \end{aligned} \tag{36}$$

Based on (29), (33), (34), (35) and (36), then we have

$$\bar{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}\left[\chi^2\left(\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right)\right] \leq \mathcal{O}\left(\frac{(1+B_{\mathcal{F}}^{-})(1+B_{\mathcal{F}})^2}{N_{\mathrm{p}}T_{\mathrm{p}}H}\cdot\log(N_{\mathrm{p}}T_{\mathrm{p}}H|\mathcal{F}|/\delta)\right). \tag{37}$$

Combine (37) and the divergence inequalities (§7.6, Polyanskiy & Wu, 2022), we have

$$\begin{aligned}\bar{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}\left[D_{\mathrm{TV}}\left(\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right)\right] &\leq \frac{1}{2}\cdot\bar{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}\left[\sqrt{\chi^2\left(\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right)}\right]\\ &\leq \frac{1}{2}\cdot\sqrt{\bar{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}\left[\chi^2\left(\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right)\right]} \leq \mathcal{O}\left(\frac{B_{\mathcal{F}}(B_{\mathcal{F}}^{-})^{1/2}}{(N_{\mathrm{p}}T_{\mathrm{p}}H)^{1/2}}\sqrt{\log(N_{\mathrm{p}}T_{\mathrm{p}}H|\mathcal{F}_\gamma|/\delta)}\right),\end{aligned}$$

where the second inequality follows $\mathbb{E}[X]\leq\sqrt{\mathbb{E}[X^2]}$ and we finish the proof of Theorem 5.5. □

D.2. Proof of Theorem 5.7

Notations. Denote $(\mathcal{J},\widehat{\mathcal{J}})$, $(\pi_z^*,\widehat{\pi}_z^*)$, and $(\mathbb{P}_{z,h},\widehat{\mathbb{P}}_{z,h})$ as the value functions, optimal policies, and probability distributions under the environment concerning the ground-truth $\mathbb{O}$ and the pretrained $\mathbb{O}_{\widehat{\gamma}}$. Furthermore, $(\pi^t,\widehat{\pi}^t)$ are the Planner's policy empowered by perfect $\texttt{LLM}$ or pretrained $\texttt{LLM}_{\widehat{\theta}}$.

Proof of Theorem 5.7. Conditioned on the event $\mathcal{E}_1$ that both Theorem 5.3 and 5.5 hold, the regret under the practical setting can be decomposed as

$$\begin{aligned}\mathrm{Reg}_z(T) \leq &\underbrace{\sum_{t=1}^{T}\widehat{\mathcal{J}}_z(\widehat{\pi}_z^*,\omega^t)-\mathcal{J}_z(\widehat{\pi}_z^*,\omega^t)}_{\textbf{(i)}}+\underbrace{\sum_{t=1}^{T}\mathcal{J}_z(\widehat{\pi}_z^*,\omega^t)-\mathcal{J}_z(\pi_z^*,\omega^t)}_{\textbf{(ii)}}\\ &+\underbrace{\sum_{t=1}^{T}\mathcal{J}_z(\pi_z^*,\omega^t)-\widehat{\mathcal{J}}_z(\pi_z^*,\omega^t)}_{\textbf{(iii)}}+\underbrace{\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_z(\pi_z^*,\omega^t)-\widehat{\mathcal{J}}_z(\widehat{\pi}^t,\omega^t)\right]}_{\textbf{(iv)}},\end{aligned} \tag{38}$$

and **(ii)** ≤ 0 results from the optimality such that $\mathcal{J}_z(\widehat{\pi}_z^*,\omega^t)\leq\mathcal{J}_z(\pi_z^*,\omega^t)$ for all $t\in[T]$.

Step 1. Bound (i) and (iii) with Translator's Pretraining Error.

For any policy sequence $\{\pi_t\}_{t\leq T}\subseteq\Pi$ and length $T\in\mathbb{N}$, based on PDL in Lemma F.4, we have

$$\begin{aligned}&\sum_{t=1}^{T}\widehat{\mathcal{J}}_z(\pi_t,\omega^t)-\mathcal{J}_z(\pi_t,\omega^t)\\ &=\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{(s_h^t,\tau_h^t,g_h^t)\sim\mathbb{P}_z^{\pi_t}}\left[(\mathbb{P}_{z,h}\widehat{V}_h^{\pi_t}-\widehat{\mathbb{P}}_{z,h}\widehat{V}_h^{\pi_t})(s_h^t,\tau_h^t,g_h^t,\omega^t)\right]\\ &\leq H\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{(s_h^t,\tau_h^t,g_h^t)\sim\mathbb{P}_z^{\pi_t}}\left[D_{\mathrm{TV}}\left(\mathbb{P}_{z,h}(\cdot,\cdot\,|\,s_h^t,\tau_h^t,g_h^t),\widehat{\mathbb{P}}_{z,h}(\cdot,\cdot\,|\,s_h^t,g_h^t,\tau_h^t)\right)\right]\\ &\leq H\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{(s_h^t,g_h^t)\sim\mathbb{P}_z^{\pi_t}}\mathbb{E}_{s_{h+1}^t\sim\mathbb{P}_{z,h}(\cdot\,|\,s_h^t,g_h^t)}\left[D_{\mathrm{TV}}\left(\mathbb{O}(\cdot|s_{h+1}^t),\mathbb{O}_{\widehat{\gamma}}(\cdot|s_{h+1}^t)\right)\right],\end{aligned} \tag{39}$$

where the last inequality results from the fact that for any f-divergence, it holds that

$$D_f(\mathbb{P}_{Y|X}\otimes\mathbb{P}_X,\mathbb{Q}_{Y|X}\otimes\mathbb{P}_X)=\mathbb{E}_{X\sim\mathbb{P}_X}[D_f(\mathbb{P}_{Y|X},\mathbb{Q}_{Y|X})],$$

Based on (39), by taking policies $\pi = \widehat{\pi}_z^*$ and $\pi = \pi_z^*$ respectively, we have

$$\begin{aligned}\textbf{(i)} + \textbf{(iii)} &= \sum_{t=1}^{T} \widehat{\mathcal{J}}_z(\widehat{\pi}_z^*, \omega^t) - \mathcal{J}_z(\widehat{\pi}_z^*, \omega^t) + \sum_{t=1}^{T} \mathcal{J}_z(\pi_z^*, \omega^t) - \widehat{\mathcal{J}}_z(\pi_z^*, \omega^t) \\ &\leq 2H^2 T \cdot \max_{s \in \mathcal{S}} \left\{ D_{\mathrm{TV}} \left(\mathbb{O}(\cdot \,|\, s), \mathbb{O}_{\widehat{\gamma}}(\cdot \,|\, s) \right) \right\} \leq 2H^2 T \lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta), \end{aligned} \tag{40}$$

where the last inequality results from Assumption 5.6 and Theorem 5.5.

Step 2. Bound (iv) with LLM's and Translator's Pretraining Errors

Recall that the Planner follows a mixture policy of $\pi_{\texttt{exp}}$ and $\widehat{\pi}_{\texttt{LLM}}$ as

$$\pi_h^t(\cdot \,|\, \tau_h^t, \omega^t) \sim (1-\epsilon) \cdot \widehat{\pi}_{h,\texttt{LLM}}^t(\cdot \,|\, \tau_h^t, \omega^t) + \epsilon \cdot \pi_{h,\texttt{exp}}(\cdot | \tau_h^t). \tag{41}$$

Based on PDL in Lemma F.4, the performance difference in term **(iv)** can be decomposed as

$$\begin{aligned}\textbf{(iv)} &= \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[(\pi_{z,h}^* - \widehat{\pi}_h^t) \widehat{Q}_h^{\pi_z^*}(s_h^t, \tau_h^t, \omega^t) \right] \\ &= \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[\left(\pi_{z,h}^* - \widehat{\pi}_{h,\texttt{LLM}}^t \right) \widehat{Q}_h^{\pi_z^*}(s_h^t, \tau_h^t, \omega^t) \right] \cdot (1-\epsilon) \\ &\quad + \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{exp}} \right) \widehat{Q}_h^{\pi_z^*}(s_h^t, \tau_h^t, \omega^t) \right] \cdot \epsilon \\ &\leq H \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\pi_{z,h}^*(\cdot \,|\, \tau_h^t, \omega^t), \texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{pt}_h^t) \right) \right] + HT\epsilon \end{aligned} \tag{42}$$

where we write $\pi_h Q_h(s_h, \tau_h, \omega) = \langle \pi_h(\cdot | \tau_h, \omega), Q_h(s_h, \tau_h, \cdot, \omega) \rangle_{\mathcal{G}}$ for all $h \in [H]$, and $\widehat{Q}_h^{\pi}$ denotes the action value function under the practical setting. Furthermore, we have

$$\begin{aligned}&\sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\pi_{z,h}^*(\cdot \,|\, \tau_h^t, \omega^t), \texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{pt}_h^t) \right) \right] \\ &\quad \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{pt}_h^t), \texttt{LLM}(\cdot \,|\, \texttt{pt}_h^t) \right) \right] \\ &\qquad + \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\pi_{z,h}^*(\cdot \,|\, \tau_h^t, \omega^t), \texttt{LLM}(\cdot \,|\, \texttt{pt}_h^t) \right) \right] \\ &\quad \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{pt}_h^t), \texttt{LLM}(\cdot \,|\, \texttt{pt}_h^t) \right) \right] \\ &\qquad + \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[\sum_{z' \neq z} \mathbb{P}_{\mathcal{D}}(z' \,|\, \texttt{pt}_h^t) \right], \end{aligned} \tag{43}$$

where the first inequality arises from the triangle inequality, and the second inequality results from Thoerem 4.2. Furthermore, the first term can be bounded by the pretraining error, following

$$\begin{aligned}&\sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \widehat{\mathbb{P}}_z^{\widehat{\pi}_i}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{pt}_h^t), \texttt{LLM}(\cdot \,|\, \texttt{pt}_h^t) \right) \right] \\ &\quad \leq \lambda_S \cdot \sum_{t=1}^{T} \sum_{h=1}^{H} \bar{\mathbb{E}}_{\texttt{pt}_h^t \sim \mathcal{D}_{\texttt{LLM}}} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{pt}_h^t), \texttt{LLM}(\cdot \,|\, \texttt{pt}_h^t) \right) \right], \\ &\quad = \lambda_S HT \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta), \end{aligned} \tag{44}$$

where the last inequality follows Theorem 5.3 and Assumption 5.6. Under practical setting, $\mathtt{pt}_h^t$ is generated from practical transition $\widehat{\mathbb{P}}_z$, mismatching $\mathbb{P}_{\mathcal{D}}(z\,|\,\mathtt{pt}_h^t)$ in pretraining. Let $\mathcal{X}_{\mathtt{exp}}^t = \{i \in [t] : \widehat{\pi}^i = \pi_{\mathtt{exp}}\}$ and write $\breve{\tau}_{h/t}^i = \tau_H^i$ for all $i < t$ and $\breve{\tau}_{h/t}^t = \tau_h^t$. Define the information gains as

$$L_{h,t}^{\mathtt{exp}}(z') = \sum_{i \in \mathcal{X}_{\mathtt{exp}}^t} \log\left(\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right), \quad L_{h,t}^{\mathtt{LLM}}(z') = \sum_{i \in [t]\setminus\mathcal{X}_{\mathtt{exp}}^t} \log\left(\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right), \tag{45}$$

where $\mathbb{P}_z(\tau_h)$ is defined in (11). Based on the law of total probability, we have

$$\mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{pt}_h^t) = \frac{\mathbb{P}_{z'}(\mathtt{pt}_h^t)\cdot\mathcal{P}_{\mathcal{Z}}(z')}{\sum_{\tilde{z}\in\mathcal{Z}}\mathbb{P}_{\tilde{z}}(\mathtt{pt}_h^t)\cdot\mathcal{P}_{\mathcal{Z}}(\tilde{z})} \le \frac{\mathbb{P}_{z'}(\mathtt{pt}_h^t)}{\mathbb{P}_z(\mathtt{pt}_h^t)}\cdot\frac{\mathcal{P}_{\mathcal{Z}}(z')}{\mathcal{P}_{\mathcal{Z}}(z)}. \tag{46}$$

Let $\mathcal{E}_2$ be the event that Lemma D.1 holds. Based on (46), (45) and conditioned on event $\mathcal{E}_2$, it holds that

$$\begin{aligned}
\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{pt}_h^t) &\le \min\left\{\sum_{z'\neq z}\frac{\mathbb{P}_{z'}(\mathtt{pt}_h^t)}{\mathbb{P}_z(\mathtt{pt}_h^t)}\cdot\frac{\mathcal{P}_{\mathcal{Z}}(z')}{\mathcal{P}_{\mathcal{Z}}(z)}, 1\right\}\\
&\le \min\left\{c_{\mathcal{Z}}\sum_{z'\neq z}\exp\left(L_{h,t}^{\mathtt{exp}}(z') + L_{h,t}^{\mathtt{LLM}}(z')\right), 1\right\}\\
&\le \min\left\{c_{\mathcal{Z}}\sum_{z'\neq z}\exp\Big(t\cdot H\lambda_R^{-1}\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2 - 2\eta|\mathcal{X}_{\mathtt{exp}}^t| + 8\log(|\mathcal{Z}|/\delta) + 2\eta\Big), 1\right\}\\
&\le \min\left\{c_{\mathcal{Z}}\sum_{z'\neq z}\exp\Big(-\big(\eta\epsilon - H\lambda_R^{-1}\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\big)\,t + 8\log(|\mathcal{Z}|/\delta) + 2\eta\Big), 1\right\}\\
&\le \min\left\{c_{\mathcal{Z}}\cdot\exp\Big(-\big(\eta\epsilon - H\lambda_R^{-1}\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\big)\,t + 9\log(|\mathcal{Z}|/\delta) + 2\eta\Big), 1\right\}
\end{aligned} \tag{47}$$

for all $(h,t)\in[H]\times[T]$, where the second inequality follows Assumption 4.5. Here, we suppose that $|\mathcal{X}_{\mathtt{exp}}^t|/t = \epsilon$ for simplicity, which is attainable if we explore at a fixed fraction during episodes. Assume that $\eta\epsilon \ge H\lambda_R^{-1}\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2$ holds temporarily. Following (47) and condition on event $\mathcal{E}_2$, there exists a large constant $c_0 > 0$ such that

$$\sum_{t=1}^T\sum_{h=1}^H\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{pt}_h^t) \le c_0\cdot H\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\big(\eta\epsilon - H\lambda_R^{-1}\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\big)^{-1}, \tag{48}$$

where we use the fact that there exists constant $c_0 > 0$ such that $\sum_{t=1}^T\min\{c_3\exp(-c_1t + c_2), 1\} \le c_0\cdot c_1^{-1}(c_2 + \log c_3)$ for $c_1 \le 1$. Furthermore, based on (48), we can show that

$$\begin{aligned}
&\sum_{t=1}^T\sum_{h=1}^H\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\widehat{\mathbb{P}}_z^{\widehat{\pi}_i}}\mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}}\left[\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{pt}_h^t)\right]\\
&\quad\le \sum_{t=1}^T\sum_{h=1}^H\sum_{z'\neq z}\mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\widehat{\mathbb{P}}_z^{\widehat{\pi}_i}}\mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}}\left[\mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{pt}_h^t)\mathbb{1}\left(\mathcal{E}_2\text{ holds}\right)\right] + 2HT\delta\\
&\quad\le c_0\cdot H\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\big(\eta\epsilon - H\lambda_R^{-1}\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\big)^{-1} + 2HT\delta.
\end{aligned} \tag{49}$$

Combine (42), (47), (44) and (49), it holds that

$$\begin{aligned}
\textbf{(iv)} \le\ & \underbrace{c_0\cdot H^2\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\big(\eta\epsilon - H\lambda_R^{-1}\cdot\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\big)^{-1}}_{\textbf{(v)}}\\
&+ \underbrace{HT\eta^{-1}\left(\eta\epsilon - H\lambda_R^{-1}\cdot\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)}_{\textbf{(vi)}} + \lambda_S H^2T\cdot\Delta_{\mathtt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)\\
&+ H^2T(\eta\lambda_R)^{-1}\cdot\Delta_{\mathtt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2 + 2HT\delta,
\end{aligned} \tag{50}$$

If we explore with probability $\epsilon = H(\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 + (H \log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)/T\eta)^{1/2}$, which satisfies the condition that $\eta\epsilon \geq H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2$ assumed in (47), then we have

$$\textbf{(v)} + \textbf{(vi)} \leq \mathcal{O}\left(H^{\frac{3}{2}}\sqrt{\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta) \cdot T/\eta}\right). \tag{51}$$

Step 3. Conclude the Proof based on Step 1 and Step 2.

Combine (38), (40), (50) and (51), the regret under the practical setting follows

$$\begin{aligned}\mathrm{Reg}_z(T) &\leq \textbf{(i)} + \textbf{(iii)} + \textbf{(iv)} + HT \cdot \mathbb{P}(\mathcal{E}_1 \text{ fails}) \\ &= \mathcal{O}\Big(\underbrace{H^{\frac{3}{2}}\sqrt{\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta) \cdot T/\eta}}_{\text{Planning error}} + \underbrace{H^2 T \cdot \Delta_{\mathrm{p}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi)}_{\text{Pretraining error}}\Big) + 4HT\delta,\end{aligned} \tag{52}$$

where the cumulative pretraining error of the imperfectly pretrained PAR system follows

$$\begin{aligned}\Delta_{\mathrm{p}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi) &= (\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \\ &\quad + 2\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) + \lambda_S \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta).\end{aligned}$$

Here, $\xi = (\eta, \lambda_S, \lambda_R)$ denotes the set of distinguishability and coverage coefficients in Definition 4.4 and Assumption 5.6, and $\Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)$ and $\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)$ are pretraining errors defined in Theorem 5.3 and Theorem 5.5. By taking $\delta = 1/\sqrt{T}$, we complete the proof of Theorem 5.7. □

D.3. Proof of Lemma D.1

In this subsection, we provide a detailed examination of the posterior concentration when there exists a mismatch between the ground-truth environment and the practical environment with pretrained modules. The argument is formalized below.

Lemma D.1. *Suppose that Assumption 4.5 and Theorem 5.5 hold. For all $(z', h, t) \in \mathcal{Z} \times [H] \times [T]$, with probability at least $1 - 2\delta$, it holds that*

(i). $L_{h,t}^{\texttt{LLM}}(z') \leq \big(t - |\mathcal{X}_{\texttt{exp}}^t|\big) H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 + 4\log(|\mathcal{Z}|/\delta)$,

(ii). $L_{h,t}^{\texttt{exp}}(z') \leq |\mathcal{X}_{\texttt{exp}}^t| H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 + 4\log(|\mathcal{Z}|/\delta) - 2\eta \cdot |\mathcal{X}_{\texttt{exp}}^t| + 2\eta$,

where $L_{h,t}^{\texttt{LLM}}(z')$ and $L_{h,t}^{\texttt{exp}}(z')$ are the information gain defined in (45).

Proof of Lemma D.1. Let $\mathfrak{F}_t$ be the filtration induced by $\{\omega^i, \tau_H^i\}_{i<t} \cup \{\mathbb{1}(\pi^i = \pi_{\mathrm{exp}})\}_{i\in[t]}$. Consider a fixed tuple $(z', h, t) \in \mathcal{Z} \times [H] \times [T]$, it holds that

$$\begin{aligned}\widehat{\mathbb{P}}_z(L_{h,t}^{\texttt{LLM}}(z') \geq \beta_{h,t}^{\texttt{LLM}}) &\leq \inf_{\lambda\geq 0} \mathbb{E}_{\mathfrak{F}_{1:t}}\left[\exp(\lambda \cdot (L_{h,t}^{\texttt{LLM}}(z') - \beta_{h,t}^{\texttt{LLM}}))\right] \\ &= \inf_{\lambda\geq 0} \mathbb{E}_{\bigotimes_{i\in[t]\setminus\mathcal{X}_{\texttt{exp}}^t} \widehat{\mathbb{P}}_z^{\widehat{\pi}^i}}\left[\exp\left(\sum_{i\in[t]\setminus\mathcal{X}_{\texttt{exp}}^t} \lambda \cdot \log\left(\frac{\mathbb{P}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right) - \lambda \cdot \beta_{h,t}^{\texttt{LLM}}\right)\right] \\ &= \inf_{\lambda\geq 0} \prod_{i\in[t]\setminus\mathcal{X}_{\texttt{exp}}^t} \mathbb{E}_{\mathbb{P}_z^{\widehat{\pi}^i}}\left[\left(\frac{\mathbb{P}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right)^{\lambda} \cdot \frac{\widehat{\mathbb{P}}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right] \cdot \exp\left(-\lambda \cdot \beta_{h,t}^{\texttt{LLM}}\right) \\ &\leq \inf_{\lambda\geq 0} \prod_{i\in[t]\setminus\mathcal{X}_{\texttt{exp}}^t} \mathbb{E}_{\mathbb{P}_z^{\widehat{\pi}^i}}\left[\left(\frac{\mathbb{P}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right)^{2\lambda}\right]^{1/2} \mathbb{E}_{\mathbb{P}_z^{\widehat{\pi}^i}}\left[\left(\frac{\widehat{\mathbb{P}}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right)^{2}\right]^{1/2} \cdot \exp\left(-\lambda \cdot \beta_{h,t}^{\texttt{LLM}}\right),\end{aligned}$$

where the first inequality is a natural corollary to Lemma F.1, and the last inequality follows the Cauchy-Swartz inequality. By taking $\lambda = \frac{1}{4}$, for all $(h, t) \in [H] \times [T]$, we have

$$\mathbb{E}_{\mathbb{P}_z^{\widehat{\pi}^i}}\left[\left(\frac{\mathbb{P}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right)^{1/2}\right]^{1/2} \mathbb{E}_{\mathbb{P}_z^{\widehat{\pi}^i}}\left[\left(\frac{\widehat{\mathbb{P}}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}{\mathbb{P}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)}\right)^{2}\right]^{1/2} \leq \sqrt{1 + \chi^2\big(\mathbb{P}_{z'}^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i) \,\|\, \widehat{\mathbb{P}}_z^{\widehat{\pi}^i}(\check{\tau}_{h/t}^i)\big)}. \tag{53}$$

Based on Theorem 5.5 and Assumption 4.5, for any policy $\pi \in \Pi$, it holds that

$$
\begin{aligned}
&1+\chi^2\big(\mathbb{P}^{\pi}_{z'}(\tau_h)\,\|\,\widehat{\mathbb{P}}^{\pi}_{z}(\tau_h)\big) \le 1+\chi^2\big(\mathbb{P}^{\pi}_{z'}(\tau_h, s_{1:h})\,\|\,\widehat{\mathbb{P}}^{\pi}_{z}(\tau_h, s_{1:h})\big)\\
&\le 1+\chi^2\left(\prod_{h'=1}^{h}\mathbb{P}^{\pi}_{z'}(g_h, s_{h+1}\,|\,\tau_h, s_h)\cdot\mathbb{O}(o_h\,|\,s_h)\,\|\prod_{h'=1}^{h}\mathbb{P}^{\pi}_{z'}(g_h, s_{h+1}\,|\,\tau_h, s_h)\cdot\mathbb{O}_{\widehat{\gamma}}(o_h\,|\,s_h)\right)\\
&\le \big(1+\max_{s\in\mathcal{S}}\big\{\chi^2\big(\mathbb{O}(\cdot\,|\,s)\,\|\,\mathbb{O}_{\widehat{\gamma}}(\cdot\,|\,s)\big)\big\}\big)^{H} \le \big(1+\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2\big)^{H},
\end{aligned}
\tag{54}
$$

where the first inequality follows data processing inequality and the second inequality arises from the tensorization (Theorem 7.32 and §7.12, Polyanskiy & Wu, 2022). To ensure that $L^{\texttt{LLM}}_{h,t}(z') \le \beta^{\texttt{LLM}}_{h,t}$ holds for all $(z', h, t) \in \mathcal{Z}\times[H]\times[T]$ with probability at least $1-\delta$, we let

$$
\prod_{i\in[t]\setminus\mathcal{X}^t_{\texttt{exp}}}\sqrt{1+\chi^2\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})\,\|\,\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)}\cdot\exp\left(-\frac{\beta^{\texttt{LLM}}_{h,t}}{4}\right)=\frac{\delta}{|\mathcal{Z}|},
$$

with a union bound taken over $\mathcal{Z}$, since Lemma F.1 has ensured the inequality holds for all $(h,t)\in[H]\times[T]$. Thus, the constant $\beta^{\texttt{LLM}}_{h,t}$ is then chosen as

$$
\begin{aligned}
\beta^{\texttt{LLM}}_{h,t} &= 2\sum_{i\in[t]\setminus\mathcal{X}^t_{\texttt{exp}}}\log\Big(1+\chi^2\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})\,\|\,\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)\Big)+4\log(|\mathcal{Z}|/\delta)\\
&\le \big(t-|\mathcal{X}^t_{\texttt{exp}}|\big)\cdot H\log\big(1+\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2\big)+4\log(|\mathcal{Z}|/\delta)\\
&\le \big(t-|\mathcal{X}^t_{\texttt{exp}}|\big)\cdot H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2+4\log(|\mathcal{Z}|/\delta),
\end{aligned}
$$

which is based on (53), (54) by taking a union bound over $\mathcal{Z}$, and the last inequality results from $\log(1+x)\le x$ for all $x\ge 0$. Similarly, for the exploration episodes, we let

$$
\begin{aligned}
&\widehat{\mathbb{P}}_z\big(L^{\texttt{exp}}_{h,t}(z')\ge\beta^{\texttt{exp}}_{h,t}\big)\le\inf_{\lambda\ge 0}\mathbb{E}\Big[\exp\big(\lambda\cdot(L^{\texttt{exp}}_{h,t}-\beta^{\texttt{exp}}_{h,t})\big)\Big]\\
&\le\prod_{i\in\mathcal{X}^t_{\texttt{exp}}}\sqrt{1-D^2_{\mathrm{H}}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}),\mathbb{P}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)}\cdot\sqrt{1+\chi^2\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})\,\|\,\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)}\cdot\exp\left(-\frac{1}{4}\beta^{\texttt{exp}}_{h,t}\right).
\end{aligned}
$$

Furthermore, based on Definition 4.4, the expolration episodes satisfies that

$$
\sum_{i\in\mathcal{X}^t_{\texttt{exp}}}D^2_{\mathrm{H}}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}),\mathbb{P}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)\ge\sum_{i\in\mathcal{X}^{t-1}_{\texttt{exp}}}D^2_{\mathrm{H}}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\tau_H),\mathbb{P}^{\widehat{\pi}^i}_{z}(\tau_H)\big)\ge\eta\cdot|\mathcal{X}^{t-1}_{\texttt{exp}}|.
\tag{55}
$$

To ensure that $L^{\texttt{exp}}_{h,t}(z')\le\beta^{\texttt{exp}}_{h,t}$ holds for all $(z',h,t)\in\mathcal{Z}\times[H]\times[T]$ with high probability, we take

$$
\prod_{i\in[t]\setminus\mathcal{X}^t_{\texttt{exp}}}\sqrt{1-D^2_{\mathrm{H}}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}),\mathbb{P}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)}\cdot\sqrt{1+\chi^2\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})\,\|\,\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)}\cdot\exp\left(-\frac{\beta^{\texttt{exp}}_{h,t}}{4}\right)=\frac{\delta}{|\mathcal{Z}|},
$$

with a union bound taken over $\mathcal{Z}$, and thus the constant $\beta^{\texttt{exp}}_{h,t}$ is chosen as

$$
\begin{aligned}
\beta^{\texttt{exp}}_{h,t} &= 2\sum_{i\in\mathcal{X}^t_{\texttt{exp}}}\log\Big(1-D^2_{\mathrm{H}}\Big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}),\mathbb{P}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\Big)\Big)\\
&\quad+2\sum_{i\in\mathcal{X}^t_{\texttt{exp}}}\log\Big(1+\chi^2\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})\,\|\,\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z}(\breve{\tau}^i_{h/t})\big)\Big)+4\log(|\mathcal{Z}|/\delta)\\
&\le|\mathcal{X}^t_{\texttt{exp}}|\cdot H\log\big(1+\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2\big)+4\log(|\mathcal{Z}|/\delta)-2\eta\cdot|\mathcal{X}^{t-1}_{\texttt{exp}}|\\
&\le|\mathcal{X}^t_{\texttt{exp}}|\cdot H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2+4\log(|\mathcal{Z}|/\delta)-2\eta\cdot(|\mathcal{X}^t_{\texttt{exp}}|-1),
\end{aligned}
$$

where the first inequality results from (54), (55) and facts that $\log(1-x)\le -x$ for all $x\le 1$ and $\log(1+x)\le x$ for all $x\ge 0$, and then we complete the proof of Lemma D.1. □

D.4. Proof of Lemma D.2

Lemma D.2 (Learning Target of Contrastive Loss)**.** *For any observation-state pair $(o, s) \in \mathcal{O} \times \mathcal{S}$ sampled from the contrastive collection process, the learning target is $f^*(o, s) = \mathbb{O}(o \,|\, s)/\mathcal{P}^-(o)$.*

Proof of Lemma D.2. For any $(o, s) \in \mathcal{O} \times \mathcal{S}$, the posterior probability of label y follows that

$$\mathbb{D}(y \,|\, o, s) := \mathbb{P}_{\mathcal{C}}(y \,|\, o, s) = \frac{\mathbb{P}_{\mathcal{C}}(o \,|\, s, y) \cdot \mathbb{P}_{\mathcal{C}}(s \,|\, y)}{\sum_{y \in \{0,1\}} \mathbb{P}_{\mathcal{C}}(o \,|\, s, y) \cdot \mathbb{P}_{\mathcal{C}}(s \,|\, y)},$$

where the equation follows Baye's Theorem and $\mathbb{P}_{\mathcal{C}}(y = 0) = \mathbb{P}_{\mathcal{C}}(y = 1) = 1/2$. Moreover, the contrastive data collection process in §3.2 indicates that

$$\mathbb{P}_{\mathcal{C}}(\cdot \,|\, s, y = 0) = \mathbb{O}(\cdot \,|\, s), \quad \mathbb{P}_{\mathcal{C}}(\cdot \,|\, s, y = 1) = \mathcal{P}^-(\cdot), \tag{56}$$

and data are labeled independent of data itself, such that $\mathbb{P}_{\mathcal{C}}(s \,|\, y) = \mathbb{P}_{\mathcal{C}}(s)$. Thus, $\mathbb{P}_{\mathcal{C}}(y \,|\, o, s) = \mathbb{P}_{\mathcal{C}}(o \,|\, s, y)/(\mathcal{P}^-(o) + \mathbb{O}(o \,|\, s))$. Recall that the population risk is

$$\mathcal{R}_{\mathrm{CT}}(\gamma; \mathcal{D}_{\texttt{Rep}}) = \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbb{D}_\gamma(\cdot|o, s) \,\|\, \mathbb{D}(\cdot|o, s)\right) + \mathrm{Ent}(\mathbb{D}(\cdot|o, s))\right].$$

As the minimum is attained at $\mathbb{D}_\gamma(\cdot \,|\, o, s) = \mathbb{D}(\cdot \,|\, o, s)$. Following (7), the learning target follows

$$\frac{\mathbb{P}_{\mathcal{C}}(o \,|\, s, y)}{\mathcal{P}^-(o) + \mathbb{O}(o \,|\, s)} = \left(\frac{f^*(o, s)}{1 + f^*(o, s)}\right)^y \left(\frac{1}{1 + f^*(o, s)}\right)^{1-y}. \tag{57}$$

By solving the equation in (57), the learning target follows that $f^*(o, s) = \mathbb{O}(o \,|\, s)/\mathcal{P}^-(o)$ for the contrastive loss in (6), and then we conclude the proof of Lemma D.2. □

E. Proof for Section B: Extentions

E.1. Proof of Proposition B.1

Proof of Proposition B.1. Based on the law of total probability, it holds that

$$\mathbb{P}_{\mathcal{D}}\left(o_h \,|\, (o, g)_{1:h-1}, \mathcal{H}_t\right) = \sum_{z \in \mathcal{Z}} \mathbb{P}_z\left(o_h \,|\, (o, g)_{1:h-1}\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, (o, g)_{1:h-1}, \mathcal{H}_t\right) \tag{58}$$

Furthermore, based on Baye's theorem, we have

$$\mathbb{P}_{\mathcal{D}}\left(z \,|\, (o, g)_{1:h-1}, \mathcal{H}_t\right) = \frac{\prod_{h'=1}^{h-2} \mathbb{P}_z\left(o_{h'+1} \,|\, (o, g)_{1:h'}\right)}{\prod_{h'=1}^{h-2} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1} \,|\, (o, g)_{1:h'}, \mathcal{H}_t\right)} \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t), \tag{59}$$

Hence, (58) and (59) jointly indicates that

$$\begin{aligned}
\prod_{h'=1}^{h-1} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1} \,|\, (o, g)_{1:h'}, \mathcal{H}_t\right) &= \mathbb{P}_{\mathcal{D}}\left(o_h \,|\, (o, g)_{1:h-1}, \mathcal{H}_t\right) \cdot \prod_{h'=1}^{h-2} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1} \,|\, (o, g)_{1:h'}, \mathcal{H}_t\right) \\
&= \sum_{z \in \mathcal{Z}} \mathbb{P}_z\left(o_h \,|\, (o, g)_{1:h-1}\right) \cdot \prod_{h'=1}^{h-2} \mathbb{P}_z\left(o_{h'+1} \,|\, (o, g)_{1:h'}\right) \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t) \\
&= \sum_{z \in \mathcal{Z}} \left(\prod_{h'=1}^{h-1} \mathbb{P}_z\left(o_{h'+1} \,|\, (o, g)_{1:h'}\right)\right) \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t).
\end{aligned} \tag{60}$$

Thus, following the definition of marginal distributions, it holds that

$$\begin{aligned}
\mathbb{P}^t_{\texttt{LLM}}\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) &= \int_{o_{2:h-1}} \prod_{h'=1}^{h-1} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1} \,|\, (o, g)_{1:h'}, \mathcal{H}_t\right) \mathrm{d}o_{2:h-1} \\
&= \sum_{z \in \mathcal{Z}} \left(\int_{o_{2:h-1}} \prod_{h'=1}^{h-1} \mathbb{P}_z\left(o_{h'+1} \,|\, (o, g)_{1:h'}\right) \mathrm{d}o_{2:h-1}\right) \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t) \\
&= \sum_{z \in \mathcal{Z}} \mathbb{P}_z\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t),
\end{aligned}$$

where the second equation follows (60) and then we complete the proof of Proposition B.1. □

E.2. Proof of Corollary B.3

Notations. Denote $(\mathcal{J}, \widehat{\mathcal{J}})$ and $(\pi_z^*, \widehat{\pi}_z^*)$, and $(\mathbb{P}_{z,h}, \widehat{\mathbb{P}}_{z,h})$ as the value functions, optimal policies, and probability under the environment concerning the ground-truth $\mathbb{O}$ and the pretrained $\mathbb{O}_{\widehat{\gamma}}$. Let $(\widehat{\mathcal{J}}_{t,\texttt{LLM}}, \widehat{\pi}_{\texttt{LLM}}^{t,*})$ denote the value function of the environment simulated by pretrained $\texttt{LLM}_{\widehat{\theta}}$ and its optimal policy; $\mathcal{J}_{t,\texttt{LLM}}$ denote the value function of the environment simulated by perfect LLM; $(\mathbb{P}_{\texttt{LLM}}^t, \widehat{\mathbb{P}}_{\texttt{LLM}}^t)$ are the probability under environment simulated by perfect LLM or pretrained $\texttt{LLM}_{\widehat{\theta}}$.

Proof of Corollary B.3. Condition on the event $\mathcal{E}_1$ that both Theorem 5.3 and 5.5 hold, the regret can be decomposed as

$$\begin{aligned}
\mathrm{Reg}_z(T) \le & \underbrace{\sum_{t=1}^T \widehat{\mathcal{J}}_z(\widehat{\pi}_z^*, \omega^t) - \mathcal{J}_z(\widehat{\pi}_z^*, \omega^t)}_{\textbf{(i)}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\mathcal{J}_z(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_z^*, \omega^t)\right]}_{\textbf{(ii)}} \\
& + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t)\right]}_{\textbf{(iii)}} \\
& + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t) - \mathcal{J}_z(\widehat{\pi}^t, \omega^t)\right]}_{\textbf{(iv)}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\mathcal{J}_z(\widehat{\pi}^t, \omega^t) - \widehat{\mathcal{J}}_z(\widehat{\pi}^t, \omega^t)\right]}_{\textbf{(v)}}. \qquad (61)
\end{aligned}$$

Step 1. Bound (i) and (v) with Translator's Pretraining Error.

Similar to (39) in the proof of Theorem 5.7, it holds that

$$\textbf{(i)} + \textbf{(vi)} \le 2H^2 T \lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta), \qquad (62)$$

following the pretraining error in Theorem 5.5.

Step 2. Bound (iii) via Optimality in Planner's Algorithm.

Recall that the Planner follows policy $\pi_h^t(\cdot \,|\, \tau_h^t, \omega^t) \sim (1-\epsilon) \cdot \widehat{\pi}_{h,\texttt{LLM}}^{t,*}(\cdot \,|\, \tau_h^t, \omega^t) + \epsilon \cdot \pi_{h,\texttt{exp}}(\cdot | \tau_h^t)$. Then, it holds that

$$\begin{aligned}
\textbf{(iii)} &= \sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t)\right] + \sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t)\right] \\
&\le \sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t) - (1-\epsilon) \cdot \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t) - \epsilon \cdot \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi_{\texttt{exp}}, \omega^t)\right] \le 2HT\epsilon, \qquad (63)
\end{aligned}$$

where the the first inequality results from the optimality of $\widehat{\pi}_{\texttt{LLM}}^{t,*}$ under simulated environment.

Step 3. Bound (ii) and (iv) with LLM's Pretraining Error.

For any policy $\pi \in \Pi$, given history $\mathcal{H}_t$, the performance difference follows

$$\begin{aligned}
\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_z(\pi, \omega^t) &= \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_{t,\texttt{LLM}}(\pi, \omega^t) + \mathcal{J}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_z(\pi, \omega^t) \\
&\le \underbrace{\mathbb{E}\left[\sum_{h=1}^H \int_{o_h} \left(\widehat{\mathbb{P}}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) - \mathbb{P}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right)\right) \mathrm{d}o_h\right]}_{\textbf{(vi)}} \\
&\quad + \underbrace{\sup_{g_{1:H-1}} \sum_{h=1}^H \int_{o_h} \left(\mathbb{P}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) - \mathbb{P}_z\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right)\right) \mathrm{d}o_h}_{\textbf{(vii)}}, \qquad (64)
\end{aligned}$$

where the inequality arises from $\|r_h\|_\infty \leq 1$ depending solely on o_h. Furthermore, we have

$$\begin{aligned}
&\int_{o_h} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) - \mathbb{P}^t_{\texttt{LLM}}\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) \mathrm{d}o_h \\
&\quad = \int_{o_{2:h}} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{2:h} \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) - \mathbb{P}^t_{\texttt{LLM}}\left(o_{2:h} \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) \mathrm{d}o_{2:h} \\
&\quad = \int_{o_{2:h}} \left(\prod_{h'=1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) - \prod_{h'=1}^{h-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) \right) \mathrm{d}o_{2:h}. \qquad (65)
\end{aligned}$$

Following the arguments above, the difference can be decomposed as

$$\begin{aligned}
&\prod_{h'=1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) - \prod_{h'=1}^{h-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) \\
&\quad = \sum_{h'=1}^{h-1} \left(\widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) - \mathbb{P}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}\right) \right) \\
&\qquad \cdot \prod_{k=h'+1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (o,g)_{1:k}\right) \cdot \prod_{k=1}^{h'-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (o,g)_{1:k}\right) \\
&\quad = \sum_{h'=1}^{h-1} \left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) - \texttt{LLM}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) \right) \\
&\qquad \cdot \prod_{k=h'+1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (o,g)_{1:k}\right) \cdot \prod_{k=1}^{h'-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (o,g)_{1:k}\right). \qquad (66)
\end{aligned}$$

Combine (65) and (66), it holds that

$$\begin{aligned}
\textbf{(vi)} &\leq \sum_{h=1}^{H} \int_{o_{2:h}} \sum_{h'=1}^{h-1} \left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) - \texttt{LLM}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) \right) \\
&\qquad \cdot \prod_{k=h'+1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (o,g)_{1:k}\right) \cdot \prod_{k=1}^{h'-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (o,g)_{1:k}\right) \mathrm{d}o_{2:h} \\
&\leq \sum_{h=1}^{H} \sum_{h'=1}^{h-1} \mathbb{E}_{(o,g)_{1:h'} | \mathcal{H}_t} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right), \texttt{LLM}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) \right) \right]. \qquad (67)
\end{aligned}$$

Following (67), for any policy $\pi \in \Pi$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t} \left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_{t,\texttt{LLM}}(\pi, \omega^t) \right] \\
&\quad \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{h'=1}^{h-1} \mathbb{E}_{\mathcal{H}_t} \mathbb{E}_{(o,g)_{1:h'} | \mathcal{H}_t} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right), \texttt{LLM}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) \right) \right] \\
&\quad \leq \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{h'=1}^{h-1} \lambda_{S,1} \lambda_{S,2}^{-1} \cdot \bar{\mathbb{E}}_{\mathcal{D}_{\texttt{LLM}}} \left[D_{\mathrm{TV}} \left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right), \texttt{LLM}\left(o_{h'+1} \,|\, (o,g)_{1:h'}, \mathcal{H}_t\right) \right) \right] \\
&\quad \leq H^2 T \lambda_{S,1} \lambda_{S,2}^{-1} \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) \qquad (68)
\end{aligned}$$

where the first inequality follows Theorem 5.3 and Assumption B.2. Based on Proposition B.1, the term **(vii)** can be upper bounded using the Bayesian aggregated arguments such that

$$\textbf{(vii)} = \sup_{g_{1:H-1}} \sum_{z' \neq z} \sum_{h=1}^{H} \int_{o_h} \left(\mathbb{P}_{z'}\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) - \mathbb{P}_{z}\left(o_h \,|\, o_1, \mathbf{do}\, g_{1:h-1}\right) \right) \cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathcal{H}_t) \mathrm{d}o_h \leq H \sum_{z' \neq z} \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathcal{H}_t).$$

Following the arguments above, for any policy $\pi \in \Pi$, it holds that

$$\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi,\omega^t) - \mathcal{J}_z(\pi,\omega^t)\right] \le H \sum_{t=1}^{T}\sum_{z'\neq z} \mathbb{E}_{\mathcal{H}_t}\left[\mathbb{P}_{\mathcal{D}}(z' \,|\, \mathcal{H}_t)\right], \tag{69}$$

Combine (68), (69) and the similar concentration arguments of posterior probability in (48), denoted by event $\mathcal{E}_2$ (see proof of Theorem 5.7 in §D.2), it holds that

$$\begin{aligned}
\textbf{(ii)} + \textbf{(iv)} &\le \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left[\left(\mathcal{J}_z(\widehat{\pi}_z^*,\omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_z^*,\omega^t)\right)\cdot \mathbb{1}\left(\mathcal{E}_2 \text{ holds}\right)\right] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left[\left(\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t,\omega^t) - \mathcal{J}_z(\widehat{\pi}^t,\omega^t)\right)\cdot \mathbb{1}\left(\mathcal{E}_2 \text{ holds}\right)\right] + 2HT\delta \\
&\le 2H^2T\lambda_{S,1}\lambda_{S,2}^{-1}\cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta) + 2HT\delta \\
&\quad + c_0\cdot 2H\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\left(\eta\epsilon - H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)^{-1}
\end{aligned} \tag{70}$$

Step 4. Conclude the Proof based on Step 1, Step 2, and Step 3.

Combine (62), (63) and (70), we have

$$\begin{aligned}
\mathrm{Reg}_z(T) &\le \underbrace{c_0\cdot 2H\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\left(\eta\epsilon - H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)^{-1}}_{\textbf{(viii)}} + 4HT\delta \\
&\quad + \underbrace{2HT\eta^{-1}\left(\eta\epsilon - H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)}_{\textbf{(ix)}} + 2H^2T\lambda_{S,1}\lambda_{S,2}^{-1}\cdot\Delta_{\texttt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta) \\
&\quad + 2H^2T(\eta\lambda_R)^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2 + 2H^2T\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta) \\
&\le \mathcal{O}\Big(H\sqrt{\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot T/\eta} + H^2T\cdot\Delta_{\mathrm{p,wm}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta,\xi)\Big) + 4HT\delta,
\end{aligned} \tag{71}$$

if we choose $\epsilon = (\log(c_{\mathcal{Z}}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2} + H(\eta\lambda_{\min})^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2$ to strike an exploration-exploitation balance between **(viii)** and **(ix)**. Thus, the cumulative pretraining error follows

$$\begin{aligned}
\Delta_{\mathrm{p,wm}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta,\xi) &= 2(\eta\lambda_R)^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2 \\
&\quad + 2\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta) + 2\lambda_{S,1}\lambda_{S,2}^{-1}\cdot\Delta_{\texttt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta).
\end{aligned}$$

Here, $\xi = (\eta,\lambda_{S,1},\lambda_{S,2},\lambda_R)$ denotes the set of distinguishability and coverage coefficients in Definition 4.4 and Assumption 5.6, and $\Delta_{\texttt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)$ and $\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)$ are pretraining errors defined in Theorem 5.3 and Theorem 5.5. By taking $\delta = 1/\sqrt{T}$, we complete the entire proof. □

E.3. Proof of Corollary B.4

The proof is similar to that in §C.2.

Proof Sketch of Corollary B.4. We first verify the claim in (14), which is akin to Proposition 4.2. Note that for all $(h,t)\in[H]\times[T]$, based on the law of total probability, it holds that

$$\begin{aligned}
\pi_{h,\texttt{LLM}}^t(\mathbf{g}_h^t \,|\, \tau_h^t,\omega^t) &= \prod_{k\in\mathcal{K}} \texttt{LLM}\big(g_{h,k}^t \,|\, \texttt{pt}_{h,k}^t\big) \\
&= \prod_{k\in\mathcal{K}}\left(\sum_{z\in\mathcal{Z}} \mathbb{P}\left(g_{h,k}^t \,|\, \texttt{pt}_{h,k}^t, z\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{pt}_{h,k}^t\right)\right) \\
&= \prod_{k\in\mathcal{K}}\left(\sum_{z\in\mathcal{Z}} \pi_{z,h,k}^*\left(g_{h,k}^t \,|\, \tau_h^t,\omega^t\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{pt}_h^t\right)\right),
\end{aligned} \tag{72}$$

where the first equation arises from the autoregressive manner of LLM, and the last equation follows the generating distribution. The Planner takes a mixture policy of $\pi_{\texttt{exp}}$ and $\pi_{\texttt{LLM}}$ such that

$$\pi_h^t(\mathbf{g}_h^t \,|\, \tau_h^t, \omega^t) \sim (1-\epsilon) \cdot \pi_{h,\texttt{LLM}}^t(\mathbf{g}_h^t \,|\, \tau_h^t, \omega^t) + \epsilon \cdot \pi_{h,\texttt{exp}}(\mathbf{g}_h^t \,|\, \tau_h^t), \tag{73}$$

for any $(h,t) \in [H] \times [T]$ given an η-distinguishable policy $\pi_{\texttt{exp}}$ (see Definition 4.4). Given a sequence of high-level tasks $\{\omega^t\}_{t\in[T]}$, the regret can be decomposed as

$$\mathrm{Reg}(T) \le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \mathbb{P}_z^{\pi_i}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t \right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t) \right] + HT\epsilon, \tag{74}$$

Recall that (17) indicates that for all $(h,t) \in [H] \times [T]$, we have

$$\begin{aligned}
&\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t \right)(\mathbf{g}_h \,|\, \tau_h, \omega) \\
&= \prod_{k\in\mathcal{K}} \left(\sum_{z'\in\mathcal{Z}} \pi_{z',h,k}^*(g_{h,k} \,|\, \tau_h, \omega) \cdot \mathbb{P}_{\mathcal{D}}\left(z' \,|\, \texttt{pt}_h^t\right) \right) - \prod_{k\in\mathcal{K}} \pi_{z,h,k}^*(g_{h,k} \,|\, \tau_h, \omega) \\
&\le H \sum_{k\in\mathcal{K}} \left(\sum_{z'\neq z} \left(\pi_{z',h,k}^* - \pi_{z,h,k}^* \right)(g_{h,k} \,|\, \tau_h, \omega) \cdot \mathbb{P}_{\mathcal{D}}\left(z' \,|\, \texttt{pt}_h^t\right) \right) \\
&\quad \cdot \prod_{k'=1}^{k-1} \left(\sum_{z''\in\mathcal{Z}} \pi_{z'',h,k'}^*(g_{h,k,k'} \,|\, \tau_h, \omega) \cdot \mathbb{P}_{\mathcal{D}}\left(z' \,|\, \texttt{pt}_h^t\right) \right) \cdot \prod_{k'=k+1}^{K} \pi_{z,h}^*(g_{h,k} \,|\, \tau_h, \omega).
\end{aligned}$$

Following this, we have

$$\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t \right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t) \le HK \cdot \sum_{z'\neq z} \mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{pt}_h^t\right), \tag{75}$$

for all $(h,t) \in [H] \times [T]$. Based on Lemma C.1 and the similar arguments in the proof Theorem 4.6 in §C.2, with probability at least $1-\delta$, the following event $\mathcal{E}_1$ holds: for all $(h,t) \in [H] \times [T]$,

$$\sum_{z'\neq z} \mathbb{P}_{\mathcal{D}}(z' \,|\, \texttt{pt}_h^t) \le \mathcal{O}\left(\min\left\{ \log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right) \eta^{-1} / |\mathcal{X}_{\texttt{exp}}^{t-1}|, 1 \right\} \right), \tag{76}$$

where $\mathcal{X}_{\texttt{exp}}^t = \{i \in [t] : \pi^i = \pi_{\texttt{exp}}\}$ denotes the set of exploration episodes. Based on (72), (75) and conditioned on $\mathcal{E}_1$, it holds that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \mathbb{P}_z^{\pi_i}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t \right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t) \right] \\
&\quad \le HK \cdot \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{z'\neq z} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \mathbb{P}_z^{\pi_i}} \mathbb{E}_{\tau_h^t \sim \mathbb{P}_z^{\pi^t}} \left[\mathbb{P}_{\mathcal{D}}(z' \,|\, \texttt{pt}_h^t) \right] \\
&\quad \le 2\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta) HK\eta^{-1} \cdot \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \bigotimes_{i=1}^{t-1} \mathbb{P}_z^{\pi_i}} \mathbb{E}_{\tau_h^t \sim \mathbb{P}_z^{\pi^t}} \left[\min\left\{ 1/|\mathcal{X}_{\texttt{exp}}^{t-1}|, 1 \right\} \right],
\end{aligned} \tag{77}$$

Note that $\mathbb{1}(\pi^t = \pi_{\texttt{exp}}) \overset{\text{iid}}{\sim} \text{Bernuolli}(\epsilon)$ for all $t \in [T]$. Besides, with probability at least $1-\delta$, the following event $\mathcal{E}_2$ holds:

$$\sum_{t=1}^{T} \min\left\{ 1/|\mathcal{X}_{\texttt{exp}}^{t-1}|, 1 \right\} \le \mathcal{O}(\epsilon^{-1} \log(T \log T/\delta)). \tag{78}$$

based on Lemma F.5. Combine (74), (77) and (78), it follows that

$$\begin{aligned}
\mathrm{Reg}_z(T) &\le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}} \mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\mathtt{LLM}}^t\right) Q_{z,h}^*(s_h,\tau_h,\omega^t)\mathbb{1}\left(\mathcal{E}_1\cap\mathcal{E}_2 \text{ holds}\right)\right] \\
&\quad + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\bigotimes_{i=1}^{t-1}\mathbb{P}_z^{\pi_i}} \mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\mathtt{LLM}}^t\right) Q_{z,h}^*(s_h,\tau_h,\omega^t)\mathbb{1}\left(\mathcal{E}_1\cap\mathcal{E}_2 \text{ fails}\right)\right] + HT\epsilon \\
&\le \mathcal{O}\Big(\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)H^2K\log(T\log T/\delta)\cdot(\eta\epsilon)^{-1} + HT\epsilon + H\sqrt{T}\log(1/\delta) + 2HT\delta\Big) \\
&\le \tilde{\mathcal{O}}\left(H^{\frac{3}{2}}\sqrt{TK/\eta\cdot\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)}\right),
\end{aligned}$$

where we choose to expolre with probability $\epsilon = \left(HK\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)/T\eta\right)^{1/2}$ in the last inequality. If we take $\delta = 1/\sqrt{T}$ in the arguments above, then we conclude the proof of Corollary B.4. □

F. Technical Lemmas

Lemma F.1 (Martingale Concentration Inequality). *Let $X_1,\dots,X_T$ be a sequence of real-valued random variables adapted to a filter $(\mathscr{F}_t)_{t\le T}$. For any $\delta\in(0,1)$ and $\lambda>0$, it holds that*

$$\mathbb{P}\left(\exists T'\in[T] : -\sum_{t=1}^{T'} X_t \ge \sum_{t=1}^{T'}\frac{1}{\lambda}\log\mathbb{E}\left[\exp(-\lambda X_t)|\mathscr{F}_{t-1}\right] + \frac{1}{\lambda}\log\left(1/\delta\right)\right) \le \delta.$$

Proof of Lemma F.1. See Lemma A.4 in Foster et al. (2021) and Theorem 13.2 in Zhang (2023) for detailed proof. Lemma A.4 in Foster et al. (2021) is a special case by taking $\lambda = 1$.

Lemma F.2 (Donsker-Varadhan). *Let P and Q be the probability measures over $\mathcal{X}$, then*

$$D_{\mathrm{KL}}(P\,\|\,Q) = \sup_{f\in\mathcal{F}}\left\{\mathbb{E}_{x\sim P}\left[f(x)\right] - \log\mathbb{E}_{x\sim Q}\left[\exp(f(x))\right]\right\},$$

where $\mathcal{F} = \{f : \mathcal{X}\mapsto\mathbb{R}\,|\,\mathbb{E}_{x\sim Q}\left[\exp(f(x))\right]\le\infty\}$.

Proof of Lemma F.2. See Donsker & Varadhan (1976) for detailed proof.

Lemma F.3 (MLE guarantee). *Let $\mathcal{F}$ be finite function class and there exists $f^*\in\mathcal{F}$ such that $f^*(x,y) = \mathbb{P}(y\,|\,x)$, where $\mathbb{P}(y\,|\,x)$ is the conditional distribution for estimation. Given a dataset $\mathcal{D} = \{x_i,y_i\}_{i\in[N]}$ where $x_i\sim\mathbb{P}_{\mathcal{D}}(\cdot\,|\,x_{1:i-1},y_{1:i-1})$ and $y_i\sim\mathbb{P}_{\mathcal{D}}(\cdot\,|\,x_i)$ for all $i\in[N]$, we have*

$$\bar{\mathbb{E}}_{\mathcal{D}}\left[D_{\mathrm{TV}}^2\left(\widehat{f}(x,\cdot), f^*(x,\cdot)\right)\right] \le 2\log(N|\mathcal{F}|/\delta)/N$$

with propbability at least $1-\delta$, where $\widehat{f}$ is the maximum likelihood estimator such that

$$\widehat{f} := \operatorname*{argmax}_{f\in\mathcal{F}}\widehat{\mathbb{E}}_{\mathcal{D}}\left[\log f(x,y)\right].$$

Proof of Lemma F.3. See Theorem 21 in Agarwal et al. (2020) for detailed proof.

Lemma F.4 (Performance Difference Lemma for POMDP). *Consider policies $\pi,\pi'\in\Pi$, it holds*

$$\mathcal{J}(\pi) - \mathcal{J}(\pi') = \sum_{h=1}^{H}\mathbb{E}_{\pi}\left[Q_h^{\pi'}(s_h,\tau_h,g_h) - V_h^{\pi'}(s_h,\tau_h)\right].$$

For fixed policy $\pi\in\Pi$ under different POMDPs, denoted by $\mathcal{M}$ and $\mathcal{M}'$, then it holds that

$$\mathcal{J}_{\mathcal{M}}(\pi) - \mathcal{J}_{\mathcal{M}'}(\pi) = \sum_{h=1}^{H}\mathbb{E}_{\mathcal{M}}^{\pi}\left[(\mathbb{P}_{h,\mathcal{M}}V_{h+1,\mathcal{M}'}^{\pi} - \mathbb{P}_{h,\mathcal{M}'}V_{h+1,\mathcal{M}'}^{\pi})(s_h,\tau_h,g_h)\right],$$

where $\mathbb{P}_{h,\mathcal{M}}V_{h+1,\mathcal{M}'}^{\pi}(s_h,\tau_h,g_h) = \langle V_{h+1,\mathcal{M}'}^{\pi}(\cdot,\cdot), \mathbb{P}_{h,\mathcal{M}}(\cdot,\cdot\,|\,s_h,\tau_h,g_h)\rangle_{\mathcal{S}\times\mathcal{T}^*}$.

Lemma F.5. *Let $X_t \overset{\text{iid}}{\sim}$ Bernuolli(ρ) and $Y_t = \sum_{\tau=1}^{t} X_\tau$. For any $\delta \in (0,1)$ and $\rho > 0$, with probability greater than $1-\delta$, it holds that $\sum_{t=1}^{T} \min\{1/Y_t, 1\} \leq \mathcal{O}(\rho^{-1}\log(T\log T/\delta))$.*

Proof of Lemma F.5. Note that $\{Y_t\}_{t\in[T]}$ is non-decreasing and it holds that

$$\sum_{t=1}^{T} \min\left\{\frac{1}{Y_t}, 1\right\} = \#\{t\in[T] : Y_t = 0\} + \sum_{t\in[T]:Y_t>0} \frac{1}{Y_t}, \tag{79}$$

and with probability at least $1-\delta$, the following event $\mathcal{E}_0$ holds:

$$t_0 := \#\{t\in[T] : Y_t = 0\} \leq \frac{\log(\delta)}{\log(1-\rho)} \leq \rho^{-1}\log(1/\delta),$$

where the first inequality results from the property of Bernuolli random variable, and the second inequality uses fact that $\log(1-x) \leq -x$ for all $x \leq 1$. For notational simplicy, we write $\{t\in[T] : Y_t > 0\} = \{t_0, \ldots, t_0 + 2^{N_T} - 1\}$. With probability at least $1-\delta$, the following event $\mathcal{E}_n$ holds:

$$Y_{t_0+2^n} = \sum_{\tau=1}^{t_0+2^n} X_t = \sum_{\tau=t_0+1}^{t_0+2^n} X_t \geq 2^n\rho - \sqrt{2^{n-1}\log(1/\delta)}. \tag{80}$$

based on the Hoeffding inequality. Suppose that $\{\mathcal{E}_n\}_{n\in[N_T]}$ holds, then we have

$$\sum_{t\in[T]:Y_t>0} \frac{1}{Y_t} = \sum_{n=0}^{N_T} \sum_{t=t_0+2^n}^{2^{n+1}-1} \frac{1}{Y_t} \leq \sum_{n=0}^{N_T} \frac{2^n}{Y_{t_0+2^n}} \leq \sum_{n=0}^{N_T} \frac{2^n}{\max\{2^n\rho - \sqrt{2^{n-1}\log(1/\delta)}, 1\}}. \tag{81}$$

Let $n_0 = 1 + \lceil \log_2(\rho^{-2}\log(1/\delta))\rceil$ such that $\rho - \sqrt{\log(1/\delta)/2^{n+1}} \geq \rho/2$. Following (81), it holds

$$\sum_{t\in[T]:Y_t>0} \frac{1}{Y_t} \leq \sum_{n=0}^{n_0} 2^n + \sum_{n=n_0+1}^{N_T} 2\rho^{-1} \leq 2^{n_0+1} + 2\rho^{-1}N_T \leq 8\rho^{-2}\log(1/\delta) + 4\rho^{-1}\log T. \tag{82}$$

Combine (80) and (82), by taking a union bound over $\mathcal{E}_0, \ldots, \mathcal{E}_{N_T}$, then we can get

$$\begin{aligned}\sum_{t=1}^{T} \min\left\{\frac{1}{Y_t}, 1\right\} &\leq 8\rho^{-2}\log(2N_T/\delta) + 4\rho^{-1}\log(2TN_T/\delta) \\ &\leq 8\rho^{-2}\log(4\log T/\delta) + 4\rho^{-1}\log(4T\log T/\delta) \leq \mathcal{O}(\rho^{-1}\log(T\log T/\delta)),\end{aligned}$$

where we use the fact that $\log_2 T \leq 2\log T$, and then we finish the proof of Lemma F.5. □