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

Jianliang He* Siyu Chen† Fengzhuo Zhang‡ Zhuoran Yang§

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.

Contents

*Department of Statistics and Data Science, Fudan University. Email: hejl20@fudan.edu.cn.
†Department of Statistics and Data Science, Yale University. Email: siyu.chen.sc3226@yale.edu.
‡Department of Electrical and Computer Engineering, National University of Singapore. fzzhang@u.nus.edu.
§Department of Statistics and Data Science, Yale University. Email: zhuoran.yang@yale.edu.

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 towards achieving artificial general 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 understand natural languages. 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 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 the 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, such as reinforcement learning (RL), where the RL policy is specifically trained for a singular task. 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 through prompting with the LLM fixed. The Actor, as language-conditioned policies, can be trained 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 and Mahadevan, 2003; Pateria et al., 2021), positioning the LLM Planner

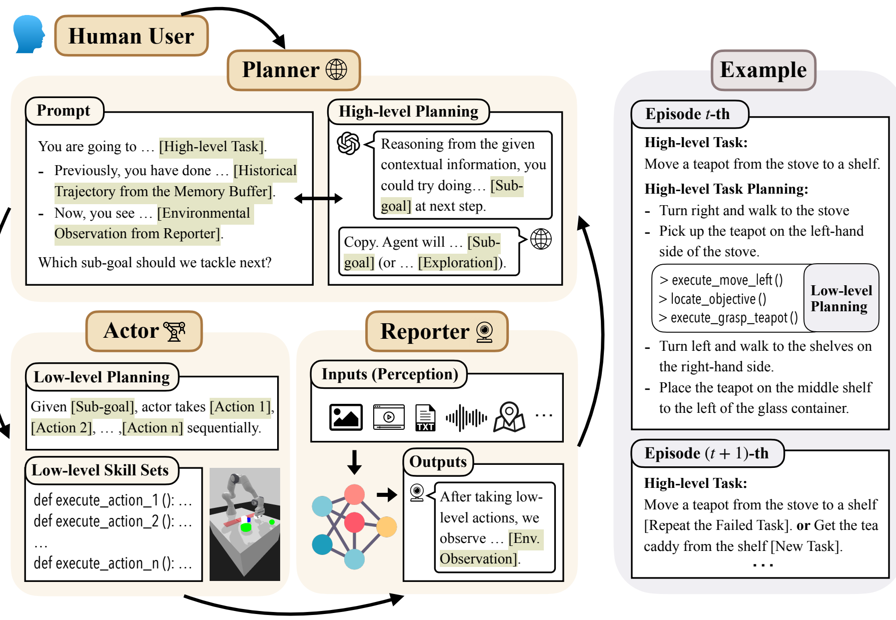

Figure 1: Overview of the Planner-Actor-Reporter (PAR) system as LLM Agents. Acting as a central controller, `Planner` conducts the high-level planning by storing the history and reasoning through the iterative use of the ICL ability of LLMs, coupled with explorations. `Actor` handles low-level planning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

and the Actor as policies operating within high-level POMDPs and low-level MDPs, respectively. Both levels share the same state spacenamely, 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 the 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.

Furthermore, to answer Question (b), we prove that when the 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. 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 through 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 inadequate exploration, resulting in a linear regret. To mitigate this (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. Specifically, to address Question (d) we establish that the regret is bounded by a sum of two terms: 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 pretaining the LLM Planner and Reporter via maximum likelihood estimation (MLE) and contrastive learning, respectively.

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

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 $\ell_{t+1} \sim \mathtt{LLM}_\theta(\cdot \,|\, S_t)$, where $S_t = (\ell_1, \dots, \ell_t)$ and $t \in \mathbb{N}$. Each token $\ell_t \in \mathfrak{L}$ specifies a word or 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 the 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 $\mathtt{prompt}_t = (\ell_1, \dots, \ell_t) \in \mathfrak{L}^*$ is generated based on a latent variable $z \in \mathcal{Z}$ autoregressively. The token follows a generating distribution such that $\ell_t \sim \mathbb{P}(\cdot \,|\, \mathtt{prompt}_{t-1}, z)$ and $\mathtt{prompt}_t = (\mathtt{prompt}_{t-1}, \ell_t)$, where $\mathcal{Z}$ represents the space of hidden information or 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 this Bayesian statement.

LLM Agents. LLMs, as highlighted in (OpenAI, 2023), are powerful tools for task planning (Wei et al., 2022a; Hu and 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., 2023; Hao et al., 2023; Liu et al., 2023). By breaking down complex tasks into subgoals, LLM

Agent facilitates effective zero-shot resource allocation across environments. For instance, 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 General Framework for LLM Agents

To formalize the architecture of LLM Agents, we propose a general framework—Planner-Actor-Reporter (PAR) system. Futhermore, the problem is modeled as a hierarchical RL problem (Pateria et al., 2021). 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 perceive 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 `Planner` and translators for `Reporter` in §3.2.

3.1 Planner-Actor-Reporter System

In this section, we delve into details of PAR system under Hierarchical Markov Decision Process (HMDP). At the high level, the `Planner` (e.g., LLM) handles task planning by decomposing tasks into subgoals using a language-conditioned Partially Observable Markov Decision Process (POMDP) with a finite horizon H. At the low level, the `Actor` translates these subgoals into actionable steps in the physical world within the framework of a language-conditioned Markov Decision Process (MDP) with a finite horizon H_a[1]. See Figure 1 for an overview.

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 rewards depending on a subgoal $g \in \mathcal{G}$. Thus, the `Actor` is modeled as a language-conditioned policy $\mu = \{\mu_g\}_{g\in\mathcal{G}}$, where $\mu_g = \{\mu_{\bar{h}}\}_{\bar{h}\in[H_a]}$ and $\mu_{\bar{h}} : \mathcal{S} \times \mathcal{G} \mapsto \Delta(\mathcal{A})$. Assume that the `Actor` terminates at step $H_a + 1$, regardless of subgoal achievement. Subsequently, the `Planner` receives the observation of current state $\bar{s}_{h,H_a+1}$ from the `Reporter`, and sends a new subgoal to the `Actor` based on feedback.

High-level POMDP. Suppose that each low-level episode corresponds to a single high-level action of the `Planner`. Thus, the high-level POMDP reuses the physical state space $\mathcal{S}$ as the state

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

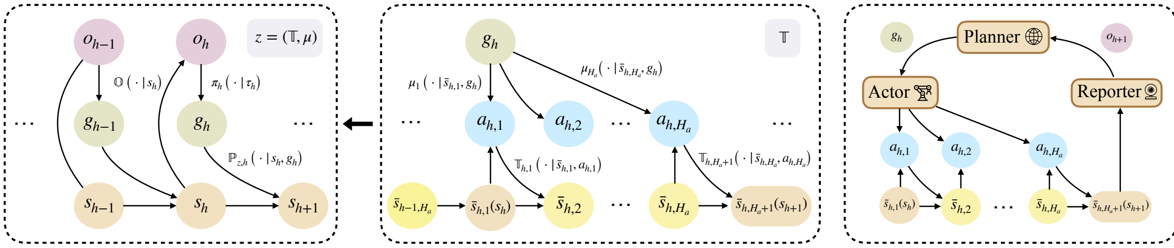

Figure 2: Illustration of structure of HMDP. The low-level MDP is featured by transition kernel $\mathbb{T}$, which characterizes the dynamics of the physical environment. The high-level transition is a result of a sequence of low-level actions in the physical environment, guided by policies $\mu = \{\mu_g\}_{g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

space, but takes the subgoal space $\mathcal{G}$ as the action space. As a result, the high-level transition kernel is jointly determined by the low-level policy μ and the physical transition kernel $\mathbb{T}$ as

$$\mathbb{P}_{z,h}(s_{h+1} \,|\, s_h, g_h) = \mathbb{P}\big(\bar{s}_{h,H_a+1} = s_{h+1} \,|\, \bar{s}_{h,1} = s_h,$$
$$a_{h,\bar{h}} \sim \mu_{\bar{h}}(\cdot \,|\, \bar{s}_{h,\bar{h}}, g_h), \bar{s}_{h,\bar{h}+1} \sim \mathbb{T}_{h,\bar{h}}(\cdot \,|\, \bar{s}_{h,\bar{h}}, \bar{a}_{h,\bar{h}})\big),$$

where we write $z = (\mathbb{T}, \mu)$. Since the LLM-empowered `Planner` cannot directly process 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 tasks in language.

Interactive Protocal. 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 the latent information to the high-level `Planner` and we denote by $\mathcal{Z}$ the space of all potential latent variables. The interactive protocol is as below: at the beginning of each episode t, `Planner` receives a task ω_t. At step h, we let

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 sequences. Specifically, LLM's recommendations are obtained by invoking the ICL ability of LLMs with history-dependent prompt:

$$\texttt{prompt}_h^t = \mathcal{H}_t \cup \left\{\omega^t, \tau_h^t\right\}, \quad \mathcal{H}_t = \textstyle\bigcup_{i=1}^{t-1} \left\{\omega^i, \tau_H^i\right\}, \tag{3.1}$$

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 $\pi_{h,\mathtt{LLM}}^t(\cdot\,|\,\tau_h^t, \omega^t) := \mathtt{LLM}_\theta(\cdot\,|\,\mathtt{prompt}_h^t)$. The `Planner` selects g_h^t and $\pi_h^t(\cdot\,|\,\tau_h^t, \omega^t)$, and sends it to `Actor`.

Module 2: Actor. Upon receiving g_h^t from `Planner`, the `Actor` plans to implement g_h^t in physical world with pretrained 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 $a_{h,\bar{h}} \sim \mu_{\bar{h}}(\cdot\,|\,\bar{s}_{h,\bar{h}}, g_h^t)$ and $\bar{s}_{h,\bar{h}+1} \sim \mathbb{T}_{h,\bar{h}}(\cdot\,|\,\bar{s}_{h,\bar{h}}, a_{h,\bar{h}})$ starting from $\bar{s}_{h,1} = s_h^t$. The low-level episode concludes at an ending state $s_{h+1}^t = \bar{s}_{h,H_a+1}$.

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

The strength of the PAR system lies in its resemblance to the reinforcement learning (RL, Sutton and 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 effectively be pretrained using the 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. Throughout this paper, we focus on the `Planner`'s performance, and regard `Actor` as an autonomous agent with pretrained skill sets. 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

$$\mathcal{J}_z(\pi, \omega) := \mathbb{E}_\pi\left[\sum_{h=1}^H r_h\left(o_h, \omega\right)\right], \tag{3.2}$$

where 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 each $(z, \omega) \in \mathcal{Z}\times\Omega$, the optimal policy is defined as

$$\pi_z^*(\omega) := \underset{\pi\in\Pi}{\operatorname{argmax}}\ \mathcal{J}_z(\pi, \omega), \tag{3.3}$$

where $\Pi = \{\pi = (\pi_1, \ldots, \pi_H), \pi_h : (\mathcal{O}\times\mathcal{G})^{h-1}\times\mathcal{O}\times\Omega \mapsto \Delta(\mathcal{G}), \forall h \in [H]\}$. To characterize the performance under the practical setting, we denote $\widehat{\mathcal{J}}_z(\pi, \omega)$ as the value function concerning the pretrained $\mathbb{O}_{\widehat{\phi}}$, and let $\widehat{\pi}_z^*(\omega) = \operatorname{argmax}_{\pi\in\Pi} \widehat{\mathcal{J}}_z(\pi, \omega)$ for all $\omega \in \Omega$ be the optimal policy in practice. The regret under practical setting is defined as

$$\mathrm{Reg}_z(T) := \sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_z(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_z(\widehat{\pi}^t, \omega^t)\right], \tag{3.4}$$

where $\widehat{\pi}^t$ represents the `Planner`'s policy empowered by a pretrained $\mathtt{LLM}_{\widehat{\theta}}$ and the expectation is taken with respect to the context $\mathcal{H}_t$ in (3.1) by sequentially taking $\{\widehat{\pi}^i\}_{i<t}$. Here, we focus

on the performance when the `Planner` collaborates with a pretrained PAR system, considering a specific environment characterized by z and tasks $\{\omega^t\}_{t\in[T]}$. Our goal is to design a sample-efficient algorithm that attains a sublinear regret, i.e., $\mathrm{Reg}_z(T) = o(T)$.

Pretraining Dataset Collection. The pretraining dataset consists of N_p independent samples with T_p episodes such that $\mathcal{D} = \{D_n\}_{n\in[N_\mathrm{p}]}$, where $D_n = \{\omega^t, \tau_H^t, \mathbf{s}_{1:H}^t\}_{t\in[T_\mathrm{p}]}$. For each sample, $z \sim \mathcal{P}_\mathcal{Z}$ specifies a low-level MDP, and $\omega^t \overset{\text{iid}}{\sim} \mathcal{P}_\Omega$ specifies the sequential tasks, where $\mathcal{P}_\mathcal{Z}$ and $\mathcal{P}_\Omega$ are the prior distributions. Assume that the joint distribution of the data, denoted by $\mathbb{P}_\mathcal{D}$, follows:

$$\mathbb{P}_\mathcal{D}(D) = \mathcal{P}_\mathcal{Z}(z) \prod_{t=1}^{T_\mathrm{p}} \mathcal{P}_\Omega(\omega^t) \prod_{h=1}^{H} \pi_{z,h}^*(g_h^{t,*} \,|\, \tau_h^t, \omega^t) \cdot \mathbb{O}(o_h^t \,|\, s_h^t) \cdot \widetilde{\pi}_{z,h}(g_h^t \,|\, \tau_h^t, \omega^t) \cdot \mathbb{P}_{z,h}(s_{h+1}^t \,|\, s_h^t, g_h^t), \tag{3.5}$$

where $\widetilde{\pi} = \{\widetilde{\pi}_h\}_{h\in[H]}$ is the behavior policy that characterizes how the contextual information is collected, and the label is then sampled from the optimal policy. Subsequently, the latent variable 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, write

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

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 in each. Following this, the pretraining dataset for LLM is autoregressively constructed with the expert knowledge, denoted by $\mathcal{D}_{\mathtt{LLM}} = \{(\widetilde{\ell}_t^n, S_t^n)\}_{(n,t)\in[N_\mathrm{p}]\times[\bar{T}_\mathrm{p}]}$, where $S_{t+1}^n = (S_t^n, \ell_t^n)$ with $S_t^n \in \mathfrak{L}^*$, and set

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

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

$$\mathcal{L}_{\mathrm{CE}}(\theta; \mathcal{D}_{\mathtt{LLM}}) := \widehat{\mathbb{E}}_{\mathcal{D}_{\mathtt{LLM}}}\left[-\log \mathtt{LLM}_\theta(\ell \,|\, S)\right], \tag{3.6}$$

and $\mathtt{LLM}_{\widehat{\theta}}$ is the pretrained LLM by algorithm in (3.6). More implementation details are deferred to §5.2.

Translator Pretraining. To pretrain translators, we employ a self-supervised contrastive learning approach, which aligns with the 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 $\mathcal{D}$. For each observation-state pair (o, s), a positive or a negative data point is generated with equal probability, following that

- **(Positive Data):** Collect $(o, s, 1)$ with a label $y = 1$;
- **(Negative Data):** Collect $(o, s^-, 0)$ with a label $y = 0$, where s^- is sampled from the designated negative sampling distribution, denoted by $\mathcal{P}^- : \mathcal{S} \mapsto \mathcal{O}$.

Denote $\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}$ as the joint distribution of the data point induced by the contrastive data collection process above. The learning algorithm follows $\widehat{\phi} = \mathrm{argmin}_{\phi \in \Phi} \mathcal{L}_{\mathrm{CT}}(\phi; \mathcal{D}_{\texttt{Rep}})$, where the contrastive loss $\mathcal{L}_{\mathrm{CT}}(\phi; \mathcal{D}_{\texttt{Rep}})$ is defined as

$$\mathcal{L}_{\mathrm{CT}}(\phi; \mathcal{D}_{\texttt{Rep}}) := \widehat{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}[y \log\left(1 + f_\phi(o, s)^{-1}\right) + (1 - y) \log\left(1 + f_\phi(o, s)\right)]. \tag{3.7}$$

Let $\mathcal{F}$ be a finite function class, where $\mathcal{F} \subseteq (\mathcal{S} \times \mathcal{O} \mapsto \mathbb{R})$ is 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 details). Following this, the pretrained translator for the `Reporter` by algorithm in (3.7) is defined as $\mathbb{O}_{\widehat{\phi}}(\cdot \,|\, \cdot) = f_{\widehat{\phi}}(\cdot, \cdot) \cdot \mathcal{P}^-(\cdot)$. More details are deferred to §5.2.

4 LLM Planning via Bayesian Aggregated Imitation Learning

In this section, we first demonstrate that the LLMs can conduct high-level planning through Bayesian aggregated imitation learning (BAIL) in §4.1, utilizing the ICL ability of LLMs with 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 and 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 (3.5) 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). In this section, we temporarily consider the perfect setting.

Definition 4.1 (Perfect Setting)**.** We say the PAR system is perfectly pretrained if (i) $\mathbb{O}_{\widehat{\phi}}(\cdot \,|\, s) = \mathbb{O}(\cdot \,|\, s)$ 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 $t \leq \bar{T}_{\mathrm{p}}$.

Algorithm 1 Planning with PAR System - `Planner`

Input: Policy $\pi_{\texttt{exp}}$ with $\eta \in (0,1)$, $c_{\mathcal{Z}} > 0$, and $|\mathcal{Z}| \in \mathbb{N}$.
Initialize: $\mathcal{H}_0 \leftarrow \{\}$, and $\epsilon \leftarrow (H \log(c_{\mathcal{Z}}^{-1}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.

```
for episode t from 1 to T do
    Receive the high-level task ω^t from the human user.
    Sample I_t ~ Bernuolli(ε).
    for step h from 1 to H do
        Collect the observation o_h^t from the Reporter.
        Set prompt_h^t ← H_t ∪ {ω^t, o_1^t, g_1^t, ..., o_h^t}.
        Sample g_{h,LLM}^t ~ LLM(·|prompt_h^t) via prompting LLM.
        If I_t = 1 then g_h^t ← g_{h,LLM}^t, else sample g_h^t ~ π_{h,exp}(·|τ_h^t).
        Send the subgoal g_h^t to the Actor.
    end for
    Update H_{t+1} ← H_t ∪ {ω^t, τ_H^t}.
end for
```

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

Proposition 4.2 (LLM Performs BAIL)**.** Assume that the pretraining data distribution is given by (3.5). 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

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

where $\pi_{h,\texttt{LLM}}^t$ is LLM's policy and prompt is defined in (3.1).

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 $\texttt{prompt}_h^t$. 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` with a perfect PAR system. From a high level, the process of task planning is an implementation of policies from imitation learning (Ross and 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 the 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 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 if $|Z| = T$. Thus, an explorative strategy 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). Similar to ϵ-greedy algorithms (Tokic and Palm, 2011; Dann et al., 2022), we provide a simple but efficient algorithm for LLM-empowered task planning. The pseudocode is presented in Algorithm 1. In each episode, the Planner performs two main steps:

- **Policy Decision** (Line 5): Randomly decide whether to execute the exploration policy $\pi_{\mathtt{exp}}$ 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, the subgoal is obtained by prompting LLMs with $\mathtt{prompt}_h^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t\}$, equivalently sampling from $\mathtt{LLM}(\cdot \,|\, \mathtt{prompt}_h^t)$. Otherwise, Planner takes sub-goal from $\pi_{h,\mathtt{exp}}(\cdot \,|\, \tau_h^t)$.

In standard ϵ-greedy algorithms, explorations are taken uniformly over the action space $\mathcal{G}$, i.e., $\pi_{\mathtt{exp}} = \mathrm{Unif}_{\mathcal{G}}$. Recent works extend it to a collection of distributions (e.g., softmax, Gaussian noise) under function approximation (Dann et al., 2022). Hence, we consider a broader class of exploration strategies satisfying the η-distinguishability below.

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

The η-distinguishability property implies the existence of an exploration policy $\pi_{\mathtt{exp}}$ that could well-distinguish the latent structure with an η-gap in Hellinger divergence concerning the distribution of the whole trajectory. Similar assumptions are adopted in ICL analysis (Xie et al., 2021) to guarantee that latent models are well-separated. Next, we introduce the assumption on the coverage of latent models.

Assumption 4.5 (Priori Coverage)**.** There exists constant $c_{\mathcal{Z}} > 0$ such that $\mathcal{P}_{\mathcal{Z}}(z)/\mathcal{P}_{\mathcal{Z}}(z') \leq c_{\mathcal{Z}}$ for all $z, z' \in \mathcal{Z}$.

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 the scenario almost surely doesn't occur, rendering planning in such scenarios unnecessary. Now, we are ready to present the main theorem of the Planner under the perfect setting.

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

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

for all fixed $z \in \mathcal{Z}$ 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 $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for task planning facilitated by LLMs. The multiplicative factor of this 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 latent models and the priori coverage in Assumption 4.5, which jointly characterizes the complexity of the external 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 $\widetilde{\mathcal{O}}(H^{\frac{3}{2}}\sqrt{T})$ Bayesian regret, i.e., $\mathbb{E}_{z\sim\mathcal{P}_{\mathcal{Z}}}[\text{Reg}(T)]$ via ICL. In comparison, we focus on a more challenging setting that aims to control the frequentist regret for specific z, which is closer to applications, and attain a comparable result with exploration.

5 Performance under Practical Setting

5.1 Pretraining Large Language Model

In this subsection, we elaborate on the pretraining of LLMs using the transformer architecture. Here, we employ a supervised learning algorithm, minimizing the cross-entropy loss such that $\widehat{\theta} = \text{argmin}_{\theta\in\Theta} \mathcal{L}_{\text{CE}}(\theta; \mathcal{D}_{\texttt{LLM}})$, as detailed in (3.7). Following this, the population risk follows that

$$\mathcal{R}_{\text{CE}}(\theta; \mathcal{D}_{\texttt{LLM}}) = \mathbb{E}_t[\mathbb{E}_{S_t}[D_{\text{KL}}(\texttt{LLM}(\cdot|S_t) \,\|\, \texttt{LLM}_\theta(\cdot|S_t)) + H_{\text{s}}(\texttt{LLM}(\cdot|S_t))]],$$

where $t \sim \text{Unif}_{[\bar{T}_{\text{p}}]}$ and $H_{\text{s}}(\mathbb{P}) = \mathbb{E}_{x\sim\mathbb{P}}[\log \mathbb{P}(x)]$ denotes the Shannon entropy. Note that the minimum is achieved if $\texttt{LLM}_\theta(\cdot|S) = \texttt{LLM}(\cdot|S)$. Thus, $\texttt{LLM}_\theta$ and $\texttt{LLM}$ are expected to converge under the learning algorithm in (3.6) with a sufficiently large dataset. 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 below 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 latent variable $z \in \mathcal{Z}$ and $t \le \bar{T}_{\text{p}}$, there exists an absolute constant $R > 0$ such that all induced sequence $S_t = (\ell_1, \dots, \ell_t) \sim \mathbb{P}_{\mathcal{D}}(\cdot\,|\,z)$ with $S_t \in \mathfrak{L}^*$ satisfies that $\|S_t^\top\|_{2,\infty} \le R$ 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 an absolute constant $c_0 > 0$ such that for all $\ell_{t+1} \in \mathfrak{L}$ and $S_t = (\ell_1, \dots, \ell_t) \in \mathfrak{L}^*$ with length $t < \bar{T}_{\text{p}}$, it holds that $\mathbb{P}_{\mathcal{D}}(\ell_{t+1}\,|\,S_t, z) \ge c_0$.

The ambiguity assumption states that the generating distribution is lower bounded, and the assumption is grounded in reasoning as there are 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 (3.6) satisfies that

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

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 $\mathcal{D}_{\mathtt{LLM}}$, converges at $\mathcal{O}\left((N_{\mathrm{p}}\bar{T}_{\mathrm{p}})^{-1/2}\right)$, where we write $\bar{\mathbb{E}}_{\mathcal{D}_t}[f(x)] = \sum_{i\in[t]}\mathbb{E}_{x_i}[f]/t$ with $x_i \sim \mathbb{P}(\cdot\,|\,x_{1:i-1})$. 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 the increasing volume of network (Proposition C.4, Zhang et al., 2023). For notational simplicity, let the right-hand side of the theorem be $\Delta_{\mathtt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)$.

5.2 Pretraining Observation-to-Language Translator

In this subsection, we focus on the pretraining approach of observation-to-language translators under a self-supervised architecture with contrastive loss. Consider function class

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

with finite functions, the contrastive loss $\mathcal{L}_{\mathrm{CT}}(\phi;\mathcal{D}_{\mathtt{Rep}})$ in (3.7) is then defined over $\mathcal{F}$. Note that the contrastive loss can be equivalently written as the negative log-likelihood loss of a binary discriminator, following that

$$
\mathbb{D}_\phi(y\,|\,o,s) := \left(\frac{f_\phi(o,s)}{1+f_\phi(o,s)}\right)^y \left(\frac{1}{1+f_\phi(o,s)}\right)^{1-y} \text{ and } \mathcal{L}_{\mathrm{CT}}(\phi;\mathcal{D}_{\mathtt{Rep}}) = \widehat{\mathbb{E}}_{\mathcal{D}_{\mathtt{Rep}}}\left[-\mathbb{D}_\phi(y\,|\,o,s)\right]. \quad (5.1)
$$

Based on the algorithm $\widehat{\phi} = \mathrm{argmin}_{\phi\in\Phi}\,\mathcal{L}_{\mathrm{CT}}(\phi;\mathcal{D}_{\mathtt{Rep}})$ and the implied form in (5.1), the population loss follows that

$$
\mathcal{R}_{\mathrm{CT}}(\phi;\mathcal{D}_{\mathtt{Rep}}) = \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbb{D}_\phi(\cdot|o,s)\,\|\,\mathbb{D}(\cdot|o,s)\right) + H_{\mathrm{s}}(\mathbb{D}(\cdot|o,s))\right]. \quad (5.2)
$$

[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 $i\leq t$.

Thus, the minimum is attained at $\mathbb{D}_\phi(\cdot\,|\,o,s) = \mathbb{D}(\cdot\,|\,o,s)$, where $\mathbb{D}(\cdot\,|\,o,s) = \mathbb{P}_{\mathcal{D}_\mathrm{r}}(\cdot\,|\,o,s)$ is the distribution of the label conditioned on the (o,s) pair in contrastive data collection. Following this, the learning target is the ground-truth likelihood ratio $f^*(o,s) = \mathbb{O}(o\,|\,s)/\mathcal{P}^-(o)$ (see Lemma D.2). Below, we assume the learning target is realizable in $\mathcal{F}$, which is standard in the literature (Qiu et al., 2022).

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

We present the performance of the pretrained translator.

Theorem 5.5 (Pretrained Translator)**.** Suppose that Assumption 5.4 holds. With probability at least $1-\delta$, the pretrained model $\mathbb{O}_{\widehat{\phi}}$ by the algorithm in (5.1) satisfies that

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

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

Theorem 5.5 posits that the average expectation of the total variation of the translation distribution regarding $\mathcal{D}_{\texttt{Rep}}$ 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 $\Delta_{\texttt{Rep}}(N_\mathrm{p}, T_\mathrm{p}, H, \delta)$. Furthermore, the algorithm also ensures a more stringent convergence guarantee concerning the χ^2-divergence:

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

5.3 Performance with Pretrained PAR System

In this subsection, we delve into the regret 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 $\boldsymbol{\pi}_t = \{\pi^i\}_{i \leq \lceil t/2H \rceil}$ from the `Planner`, it holds (i) $\widehat{\mathbb{P}}_z^{\boldsymbol{\pi}_t}(S_t) \leq \lambda_S \cdot \bar{\mathbb{P}}_{\mathcal{D}_{\mathrm{LLM}}}(S_t)$ with $S_t \in \mathfrak{L}^*$, and (ii) $\bar{\mathbb{P}}_{\mathcal{D}_{\texttt{Rep}}}(s) \geq \lambda_R$ for all state $s \in \mathcal{S}$.

Here, $\widehat{\mathbb{P}}$ represents the distribution of the dynamic system with the pretrained translator. The assumption asserts that (i) the distribution of ICL prompts induced by policy sequences $\boldsymbol{\pi}_t = \{\pi^i\}_{i \leq \lceil t/2H \rceil}$ from the `Planner` under practical scenarios is covered by the average distribution in LLM's pretraining dataset. Here, $\lceil t/2H \rceil$ is the number of the complete 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 widely adopted in the ICL analysis (Zhang et al., 2023) and the 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 attainable if an explorative behavior strategy $\widetilde{\pi}$ is deployed with sufficiently large N_p during data collection. We now present the main theorem regarding 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 $T \leq T_{\mathrm{p}}$, under the practical setting, Algorithm 1 ensures that

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

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

$$\begin{aligned}\Delta_{\mathrm{p}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi) &= (\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \\ &+ 4H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) + 2\lambda_S \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta),\end{aligned}$$

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 the perfect scenario, the `Planner` can achieve an approximate $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, but incorporating an additional pretraining error term that diminishes with an increase in the volume of pretraining data. Besides, it further 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. In response, we propose a novel regret decomposition and provide the convergence rate of posterior probability with bounded errors, distinguishing ours from Lee et al. (2023); Liu et al. (2023).

Extentions. We also present two extensions. First, we discuss the design of the `Planner` by taking LLMs as World Model (WM) in §B.1, where the `Planner` prompts LLM to predict the next observation instead of subgoals, thus eliminating the dependence on expert knowledge. Besides, we extend the result to accommodate the multi-agent collaboration, i.e., multiple `Actors`, in §B.2.

6 Conclusion

In this work, we embedded the LLM agent into a hierarchical RL framework named PAR system. At the high level, the LLM planner decomposed the user-specified task into subgoals, and at the low level, actor(s) translated the linguistic subgoals into physical realizations. The actor also provided feedbacks for augmenting the planning process through a trained reporter. In the context of a perfectly pretrained scenario, we elucidated the Bayesian Adaptive Imitation Learning (BAIL) nature of the LLM-aided planning pipeline. Our findings demonstrated that LLMs comprehend the environment through Bayesian inference on historical data and make decisions following expert policy learned during pretraining. In practical scenarios, we examined how training errors from both the LLM and the reporter contribute to the ICL error. To navigate the exploration-exploitation tradeoff, we introduced an ϵ-greedy exploration strategy, achieving $\sqrt{T}$-regret for both ideal and practical settings. 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.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., and Tang, J. (2022b). 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)*, pages 61–68.

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

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

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

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

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

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

OpenAI, R. (2023). Gpt-4 technical report. *arXiv*, pages 2303–08774.

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

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

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

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

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

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

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

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

Ross, S., Gordon, G., and Bagnell, D. (2011). 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*, pages 627–635. JMLR Workshop and Conference Proceedings.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

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

A Additional Background Knowledge and Related Works

In this appendix, we present the additional background knowledge and related works that are omitted due to the space limit. We first lay out the notations used this paper.

Notations. For some $n \in \mathbb{N}^+$, we take the convention to use $[n] = \{1, \ldots, n\}$. Denote $\Delta(\mathcal{X})$ be the probability simplex over set $\mathcal{X}$. Consider two non-negative sequence $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$, if $\limsup a_n/b_n < \infty$, then we write it as $a_n = \mathcal{O}(b_n)$ and use $\widetilde{\mathcal{O}}$ to omit logarithmic terms. Else if $\liminf a_n/b_n < \infty$, then we write $a_n = \Omega(b_n)$. For continuum $\mathcal{S}$, let $|\mathcal{S}|$ be the cardinality. Given two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, we define the total variance, Helliger distance, KL divergence and χ^2-divergence as $D_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2}\mathbb{E}_{x\sim\mathbb{P}_1}\big[\big|\mathrm{d}\mathbb{P}_2(x)/\mathrm{d}\mathbb{P}_1(x) - 1\big|\big]$, $D_{\mathrm{H}}^2(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2}\mathbb{E}_{x\sim\mathbb{P}_1}\big[\big(\sqrt{\mathrm{d}\mathbb{P}_2(x)/\mathrm{d}\mathbb{P}_1(x)} - 1\big)^2\big]$, $D_{\mathrm{KL}}(\mathbb{P}_1\|\mathbb{P}_2) = \mathbb{E}_{x\sim\mathbb{P}_1}\big[\log \mathrm{d}\mathbb{P}_1(x)/\mathrm{d}\mathbb{P}_2(x)\big]$, and $D_{\chi}^2(\mathbb{P}_1\|\mathbb{P}_2) = \mathbb{E}_{x\sim\mathbb{P}_1}\big[(\mathrm{d}\mathbb{P}_2(x)/\mathrm{d}\mathbb{P}_1(x) - 1)^2\big]$. Let $\bar{\mathbb{E}}_{\mathcal{D}}$ and $\bar{\mathbb{P}}_{\mathcal{D}}$ be average expectation and average distribution such that $\bar{\mathbb{E}}_{\mathcal{D}_t}[f(x)] = \sum_{i\in[t]}\mathbb{E}_{x_i}[f]/t$ for any function $f : \mathcal{X} \mapsto \mathbb{R}$, where $x_i \sim \mathbb{P}(\cdot\,|\,x_{1:i-1})$ and $\mathcal{D}_t = \{x_i\}_{i\in[t]}$, and $\bar{\mathbb{P}}_{\mathcal{D}_t}(x) = \bar{\mathbb{E}}_{\mathcal{D}_t}[\mathbb{1}_x]$ such that $\bar{\mathbb{P}}_{\mathcal{D}_t}[f(x)] = \sum_{i\in[t]}\mathbb{P}(\cdot\,|\,x_{1:i-1})/t$. Let $\widehat{\mathbb{E}}_{\mathcal{D}}$ be the emprical average expectation, i.e., $\widehat{\mathbb{E}}_{\mathcal{D}}[f(x)] = \sum_{x\in\mathcal{D}} f(x)/|\mathcal{D}|$.

A.1 Additional 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; Zhou et al., 2022). 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 and 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 and 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 the 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., 2023; Hu et al., 2023; Liu et al., 2023; Nottingham et al., 2023). Moreover, another recent line of research has been dedicated to devising diverse prompting schemes to enhance the reasoning capabilities 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 on 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 Additional Background on Hierarchical Markov Decision Process

In this subsection, we present a formalized definition of the HMDP model introduced in 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 such that

$$\mathbb{P}_{z,h}(s_{h+1} \,|\, s_h, g_h) = \mathbb{P}\big(\bar{s}_{h,H_a+1} = s_{h+1} \,|\, \bar{s}_{h,1} = s_h, a_{h,\bar{h}} \sim \mu_{\bar{h}}(\cdot \,|\, \bar{s}_{h,\bar{h}}, g_h), \bar{s}_{h,\bar{h}+1} \sim \mathbb{T}_{h,\bar{h}}(\cdot \,|\, \bar{s}_{h,\bar{h}}, \bar{a}_{h,\bar{h}})\big),$$

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

Please refer to Figure 2 for the interactive protocol of HMDP. Furthermore, for the high-level POMDP, the state value function of policy π, the state value function is defined as

$$V^\pi_{z,h}(s, \tau, \omega) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(o_{h'}, \omega)\Big| s_h = s, \tau_h = \tau\right], \tag{A.1}$$

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

$$Q_{z,h}^{\pi}(s, \tau, g, \omega) = \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r_{h'}(o_{h'}, \omega)\Big| s_h = s, \tau_h = \tau, g_h = g\right], \tag{A.2}$$

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

$$\mathbb{P}_z^{\pi}(\tau_h) = \pi(\tau_h) \cdot \mathbb{P}_z(\tau_h), \quad \mathbb{P}_z(\tau_h) = \prod_{h'=1}^{h-1} \mathbb{P}(o_{h'-1} \,|\, \tau_{h'}, g_{h'}), \quad \pi(\tau_h) = \prod_{h'=1}^{h-1} \pi_h(g_{h'} \,|\, \tau_{h'}), \tag{A.3}$$

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 Additional Background on LLM Pretraining

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 $\mathtt{Attn}(\mathbf{H}, \mathbf{K}, \mathbf{V}) = \mathtt{Softmax}(\mathbf{H}\mathbf{K}^\top)\mathbf{V} \in \mathbb{R}^{n\times d_s}$, where softmax function is applied row-wisely and normalize each vector via the exponential function such that $[\mathtt{Softmax}(\mathbf{h})]_i = \exp(\mathbf{h}_i)/\sum_{j=1}^d \exp(\mathbf{h}_j)$ 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 $\mathtt{Mha}(\mathbf{H}, \mathbf{W}) = \sum_{i=1}^h \mathtt{Softmax}(\mathbf{H}\mathbf{W}_i^H, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V)$, 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 $\mathtt{Proj}[\cdot]$ and a residual link. After receiving intermediate $\overline{\mathbf{H}}^{(t)} \in \mathbb{R}^{n\times d}$, FF module maps each row through a same single-hidden layer neural network with d_F neurons such that $\mathtt{ReLU}(\overline{\mathbf{H}}^{(t)}\mathbf{A}_1^{(t)})\mathbf{A}_2^{(t)}$, where $\mathbf{A}_1^{(t)} \in \mathbb{R}^{d\times d_F}$, $\mathbf{A}_2^{(t)} \in \mathbb{R}^{d_F\times d}$, and $[\mathtt{ReLU}(\mathbf{X})]_{i,j} = \max\{\mathbf{X}_{i,j}, 0\}$. Specifically, the output of the t-th layer with $t \in [D]$ can be summarized as below:

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

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

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

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 $\boldsymbol{\theta}^{(t)} = \big(\mathbf{W}^{(t)}, \mathbf{A}^{(t)}, \boldsymbol{\gamma}^{(t)}\big)$ 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

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

B Extentions

B.1 Extention for LLM as World Model

Recall that the pretraining algorithm in §3.2 also equips LLM with the capability to predict observation generation, i.e., $\mathbb{P}_h(o_h \,|\, (\mathbf{o}, \mathbf{g})_{1:h-1})$. 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 specific expert knowledge (Hu and Shu, 2023). Specifically, when regarding LLM as a simulated World Model (WM), `Planner` can interactively prompt LLM to internally simulate the entire trajectory. 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), `Planner` can utilize the LLM-simulated environment to optimize its task planning based on historical feedbacks. 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 $\mathbf{g}_{1:H}$ sequentially to predict observations $\mathbf{o}_{1:H}$. Subsequently, a simulation dataset $\mathcal{D}_t^{\sharp}$ is collected, allowing `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).

Proposition B.1 (LLM as BAWM)**.** Assume that the pretraining data distribution is given by (3.5). Under the perfect setting in Definition 4.1, for each $h \in [H]$ and $t \in [T]$, the LLM serves as a Bayesian aggregated world model, following that

$$\mathbb{P}^t_{h,\texttt{LLM}}(\cdot\,|\,o_1,\mathrm{do}(\mathbf{g}_{1:h-1}))=\sum_{z\in\mathcal{Z}}\mathbb{P}_{z,h}\left(\cdot\,|\,o_1,\mathrm{do}(\mathbf{g}_{1:h-1})\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z\,|\,\mathcal{H}_t\right), \tag{B.1}$$

where the marginal distribution of observation at h-th step is defined as

$$\mathbb{P}^t_{h,\texttt{LLM}}(\cdot\,|\,o_1,\mathrm{do}(\mathbf{g}_{1:h-1}))=\int_{\mathbf{o}_{2:h-1}}\prod_{h'=1}^{h-1}\mathbb{P}_{\mathcal{D}}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'},\mathcal{H}_t\right)\,\mathrm{d}\mathbf{o}_{2:h-1},$$
$$\mathbb{P}_{z,h}(\cdot\,|\,o_1,\mathrm{do}(\mathbf{g}_{1:h-1}))=\int_{\mathbf{o}_{2:h-1}}\prod_{h'=1}^{h-1}\mathbb{P}_{z,h}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'}\right)\,\mathrm{d}\mathbf{o}_{2:h-1}.$$

We remark that the generation distribution of observation $\mathbb{P}^t_{\texttt{LLM}}(\cdot\,|\,(\mathbf{o},\mathbf{g})_{1:h})=\texttt{LLM}(\cdot\,|\,(\mathbf{o},\mathbf{g})_{1:h},\mathcal{H}_t)$ is non-stationary due to the autoregressive nature of LLMs, resulting in inaccuracies in the posterior probability $\mathbb{P}_{\mathcal{D}}(z\,|\,(\mathbf{o},\mathbf{g})_{1:h},\mathcal{H}_t)$, which fluctuates due to simulated trajectory $(\mathbf{o},\mathbf{g})_{1:h}$. Instead, Proposition B.1 establishes that the marginal distribution is stationary with aggregation over posterior probability $\mathbb{P}_{\mathcal{D}}\left(z\,|\,\mathcal{H}_t\right)$. 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 $\boldsymbol{\pi}_t=\{\pi^i\}_{i\le\lfloor t/2H\rfloor}$ from `Planner`, it holds (i) $\widehat{\mathbb{P}}^{\boldsymbol{\pi}_t}_z(\mathcal{H}_{\lfloor t/2H\rfloor})\le\lambda_{S,1}\cdot\bar{\mathbb{P}}_{\mathcal{D}_{\texttt{LLM}}}(\mathcal{H}_{\lfloor t/2H\rfloor})$ and $\bar{\mathbb{P}}_{\mathcal{D}_{\texttt{LLM}}}(\widetilde{S}\,|\,\mathcal{H}_{\lfloor t/2H\rfloor})\ge\lambda_{S,2}$, where $S_t=(\mathcal{H}_{\lfloor t/2H\rfloor},\widetilde{S})\in\mathfrak{L}^*$ with $\mathcal{H}_{\lfloor t/2H\rfloor}\in\mathfrak{L}^*$, and (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 compared with Assumption 5.6, particularly on in-episode trajectories. Here, $\lceil t/2H\rceil$ is the number of the complete episodes described in S_t. This heightened demand arises from LLM now serving as a comprehensive WM, necessitating more extensive information for effective simulation and introducing additional prerequisites for coverage in the pretraining dataset. Suppose that the `Planner` can learn optimal policies $\widehat{\pi}^{t,*}_{\texttt{LLM}}=\mathrm{argmax}_{\pi\in\Pi}\,\widehat{\mathcal{J}}^t_{\texttt{LLM}}(\pi,\omega)$ from simulation with sufficiently large simulation steps $|\mathcal{D}^\sharp_t|$, where $\widehat{\mathcal{J}}^t_{\texttt{LLM}}$ denotes the value function concerning $\texttt{LLM}_{\widehat{\theta}}$ and history $\mathcal{H}_t$. Akin to Algorithm 1, we propose an LLM-empowered planning algorithm, which incorporates an ϵ-greedy exploration with an η-distinguishable $\pi_{\texttt{exp}}$. The pseudocode is in Algorithm 2. The corollary below presents the performance of planning by taking LLMs as WM under practical settings.

Corollary B.3 (Regret under Practical Setting with LLM as WM)**.** 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

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

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

$$\begin{aligned}\Delta_{\mathrm{p,wm}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta,\xi)&=2(\eta\lambda_R)^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\\&\quad+4H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)+4\lambda_{S,1}\lambda_{S,2}^{-1}\cdot\Delta_{\texttt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta).\end{aligned}$$

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: $\mathcal{H}_0 \leftarrow \{\}$, $\mathcal{D}_t^\sharp \leftarrow \{\}, \forall t \in [T]$, and $\epsilon \leftarrow (H \log(c_{\mathcal{Z}}^{-1}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.

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 $\mathbf{g}_n^{t,\sharp} \sim \text{Unif}(\mathcal{G}^H)$ and set $\texttt{prompt}_{1,n}^t \leftarrow \mathcal{H}_t \cup \{o_1^t, g_{1,n}^{t,\sharp}\}$.
    for step h from 1 to H **do**
      Set $\texttt{prompt}_{h,n}^t \leftarrow \mathcal{H}_t \cup \{o_{1,n}^t, g_{1,n}^{t,\sharp}, \dots, o_{h,n}^{t,\sharp}, g_{h,n}^{t,\sharp}\}$.
      Predict $\widetilde{o}_{h+1,n}^t \sim \texttt{LLM}(\cdot\,|\,\texttt{prompt}_{h,n}^t)$ via prompting LLM.
    end for
    Update $\mathcal{D}_t^\sharp \leftarrow \mathcal{D}_t^\sharp \cup \{o_{1,n}^t, g_{1,n}^{t,\sharp}, \dots, o_{H-1,n}^{t,\sharp}, g_{H-1,n}^{t,\sharp}, o_{H,n}^{t,\sharp}\}$.
  end for
  for step h from 1 to H **do**
    Collect the observation o_h^t from the `Reporter`.
    Calculate $\pi_{\texttt{LLM}}^t \leftarrow \textsc{Optimal-planning}(\omega^t, \mathcal{D}_t^\sharp, \{r_h\})$
    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

where $\xi = (\eta, \lambda_{S,1}, \lambda_{S,2}, \lambda_R)$ are in Definition 4.4 and Assumption B.2, 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

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

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

B.2 Extention for Multi-Agent Collaboration

To characterize the multi-agent interactive process, i.e., several `Actors`, of task planning, we consider a *cooperative* hierarchical Markov Game (HMG), corresponding to the 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

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: $\mathcal{H}_0 \leftarrow \varnothing$, and $\epsilon \leftarrow (HK \log(c_{\mathcal{Z}}^{-1}|\mathcal{Z}|\sqrt{T})/T\eta)^{1/2}$.

for episode t from 1 to T **do**
  Receive the high-level task ω^t from the human user.
  Sample $\mathcal{I}_t \sim \text{Bernoulli}(\epsilon)$.
  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 $\texttt{prompt}_{h,k}^t \leftarrow \mathcal{H}_t \cup \{\omega^t, o_1^t, \mathbf{g}_1^t, \dots, o_h^t, k\}$.
      Sample $g_{h,k,\texttt{LLM}}^t \sim \texttt{LLM}(\cdot \,|\, \texttt{prompt}_{h,k}^t)$ 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

$\{\mu_{h,k}\}_{k\in\mathcal{K}}$ can be correlated, e.g., within zero-sum game, Stackelberg game (Başar and Olsder, 1998). (ii) high-level POMG: under cooperation, we assume that the high-level policies can be factorized as $\pi_h(\mathbf{g}_h \,|\, \tau_{h-1}, \omega) = \prod_{k=1}^{K} \pi_{h,k}(g_{h,k} \,|\, \tau_{h-1}, \omega)$. The remaining concepts remain consistent with the HMDP. In particular, `Planner` empowered by LLMs 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., physical transition kernel $\mathbb{T}$, from the historical context, and then assign subgoal for each `Actor`. Specifically, LLM's recommendations are obtained by invoking the ICL ability of LLMs with the history-dependent prompt akin to (3.1) sequentially for each `Actor`. For the k-th `Actor`, prompt LLM with $\texttt{prompt}_{h,k}^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t, k\}$, where denote $\mathcal{H}_t = \bigcup_{i=1}^{t-1} \{\omega^i, \tau_H^i\}$ and $\tau_h^t = \{o_h^1, \mathbf{g}_h^1, \dots, o_h^t\}$. Under the perfect setting in Definition 4.1, LLM's joint policy for recommendations follows:

$$\pi_{h,\texttt{LLM}}^t(\mathbf{g}_h^t \,|\, \tau_h^t, \omega^t) = \prod_{k\in\mathcal{K}} \left(\sum_{z\in\mathcal{Z}} \pi_{z,h,k}^* \left(g_{h,k}^t \,|\, \tau_h^t, \omega^t\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{prompt}_h^t\right) \right), \tag{B.2}$$

which is akin to Proposition 4.2 and the proof of the statement is in §E.3. The pseudocode is presented in Algorithm 3. Then, we give the following corollary for `Planner`'s algorithm 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 $\pi_{\texttt{exp}}$ and $T \leq T_{\text{p}}$, the `Planner`'s algorithm in Algorithm 3 guarantees that

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

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

Corollary B.4 is akin to Theorem 4.6 with an additional $\sqrt{K}$ in regret and the proof is in §E.3. Besides, the multi-agent space of latent variable $|\mathcal{Z}| = |\mathcal{Z}_{\mathbb{T}}| \times |\mathcal{Z}_{\mathrm{m}}|$, where $\mathcal{Z}_{\mathrm{m}}$ is the space of joint policy, is generally larger than the single-agent space. Specifically, if the responses are uncorrelated, then we have $\log|\mathcal{Z}_{\mathrm{m}}| = K\log|\mathcal{Z}_{\mathrm{s}}|$, resulting in a $\sqrt{K}$ times larger regret. The proof of extension to practical setting aligns closely with Corollary B.4 based on Theorem 5.7, and is omitted.

C Proof of Performance under Perfect Setting in Section 4

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

$$\begin{aligned}
\pi_{h,\texttt{LLM}}^{t}\left(g_h^t \,|\, \tau_h^t, \omega^t\right) &= \sum_{z\in\mathcal{Z}} \mathbb{P}_{\mathcal{D}}\left(g_h^t \,|\, \texttt{prompt}_h^t, z\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{prompt}_h^t\right) \\
&= \sum_{z\in\mathcal{Z}} \mathbb{P}_{\mathcal{D}}\left(g_h^t \,|\, \mathcal{H}_t, \tau_h^t, \omega^t, z\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{prompt}_h^t\right) \\
&= \sum_{z\in\mathcal{Z}} \pi_{z,h}^{*}\left(\cdot \,|\, \tau_h^t, \omega^t\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{prompt}_h^t\right),
\end{aligned} \tag{C.1}$$

where the second equation results from Baye's theorem, the third equation follows the definition of prompts in (3.1), 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

$$\pi_h^t(\cdot \,|\, \tau_h^t, \omega^t) \sim (1-\epsilon)\cdot \pi_{h,\texttt{LLM}}^t(\cdot \,|\, \tau_h^t, \omega^t) + \epsilon \cdot \pi_{h,\texttt{exp}}(\cdot \,|\, \tau_h^t), \tag{C.2}$$

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

$$\begin{aligned}
&\pi_{h,\texttt{LLM}}^{t}\left(\cdot \,|\, \tau_h^t, \omega^t\right) = \sum_{z\in\mathcal{Z}} \pi_{z,h}^{*}\left(\cdot \,|\, \tau_h^t, \omega^t\right) \cdot \mathbb{P}_{\mathcal{D}}\left(z \,|\, \texttt{prompt}_h^t\right), \\
&\text{where } \texttt{prompt}_h^t = \mathcal{H}_t \cup \tau_h^t \text{ with } \mathcal{H}_t = \left\{\omega^i, \tau_H^i\right\}_{i\in[t-1]},
\end{aligned} \tag{C.3}$$

for all $(h,t) \in [H] \times [T]$. Following (C.2), given $z \in \mathcal{Z}$ and $\{\omega^t\}_{t\in[T]}$, the regret is decomposed as

$$
\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left[\mathcal{J}_z(\pi^*, \omega^t) - \mathcal{J}_z(\pi^t, \omega^t)\right] \\
&= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \mathbb{P}_z^{\pi_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t)\sim \mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\mathtt{exp}}\right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t)\right] \cdot \epsilon}_{\textbf{(i)}} \\
&\quad + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \mathbb{P}_z^{\pi_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t)\sim \mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\mathtt{LLM}}^t\right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t)\right] \cdot (1-\epsilon)}_{\textbf{(ii)}} \\
&\le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \mathbb{P}_z^{\pi_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t)\sim \mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\mathtt{LLM}}^t\right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t)\right] + HT\epsilon,
\end{aligned} \tag{C.4}
$$

where the second equation results from the performance difference lemma (PDL, see Lemma F.4), and we write $\pi_h Q_h(s_h, \tau_h, \omega) = \langle \pi_h(\cdot|\tau_h, \omega), Q_h(s_h, \tau_h, \cdot, \omega)\rangle_{\mathcal{G}}$, and $\mathbb{P}_z^{\pi}(\tau_h)$ is defined in (A.3). Term **(i)** denotes the exploration error bounded by $HT\epsilon$ and term **(ii)** is the planning error. Based on Lemma C.1, with probability at least $1-\delta$, the following event $\mathcal{E}_1$ holds: for all $(h,t) \in [H] \times [T]$,

$$
\sum_{z'\in\mathcal{Z}} \sum_{i\in[t]} D_{\mathrm{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big) \cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}_h^t) \le \log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right), \tag{C.5}
$$

where the randomness is incurred by the $\mathtt{prompt}_h^t$ and we define $\breve{\tau}_{h/t}^i = \tau_H^i$ for all $i \in [t-1]$ and $\breve{\tau}_{h/t}^t = \tau_h^t$. Suppose that the event $\mathcal{E}_1$ in (C.5) holds true, and denote $\mathcal{X}_{\mathtt{exp}}^t = \{i \in [t] : \pi^i = \pi_{\mathtt{exp}}\}$ as the set of exploration episodes. Note that for all $(h,t) \in [H] \times [T]$, it holds that

$$
\sum_{z'\in\mathcal{Z}} \sum_{i\in[t]} D_{\mathrm{H}}^2\big(\mathbb{P}_z^{\pi^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_{z'}^{\pi^i}(\breve{\tau}_{h/t}^i)\big) \ge \sum_{z'\in\mathcal{Z}} \sum_{i\in\mathcal{X}_{\mathtt{exp}}^{t-1}} D_{\mathrm{H}}^2\left(\mathbb{P}_z^{\pi_{\mathtt{exp}}}(\tau_H), \mathbb{P}_{z'}^{\pi_{\mathtt{exp}}}(\tau_H)\right) \ge \eta \cdot |\mathcal{X}_{\mathtt{exp}}^{t-1}|, \tag{C.6}
$$

where the last inequality results from $\pi_{\mathtt{exp}}$ is η-distinguishable (see Definition 4.4) and the fact that $D_{\mathrm{H}}^2(\mathbb{P}_1, \mathbb{P}_2) \le 1$ for all $\mathbb{P}_1, \mathbb{P}_2 \in \Delta(\mathcal{X})$. Combine (C.5) and (C.6), we can get

$$
\sum_{z'\neq z} \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}_h^t) \le \min\left\{\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)\eta^{-1}/|\mathcal{X}_{\mathtt{exp}}^{t-1}|, 1\right\}, \tag{C.7}
$$

for all $(h,t) \in [H] \times [T]$. Based on Proposition 4.2 and conditioned on $\mathcal{E}_1$, it holds that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\mathcal{H}_t\sim\mathbb{P}_z^{\pi_{1:t-1}}}\mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}}\left[\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t\right) Q_{z,h}^*(s_h^t,\tau_h^t,\omega^t)\right] \\
&\quad = \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\mathcal{H}_t\sim\mathbb{P}_z^{\pi_{1:t-1}}}\mathbb{E}_{(s_h^t,\tau_h^t)\sim\mathbb{P}_z^{\pi^t}}\left[\sum_{z'\in\mathcal{Z}}\left(\pi_{z,h}^* - \pi_{z',h}^*\right) Q_{z,h}^*(s_h^t,\tau_h^t,\omega^t)\cdot\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}_h^t)\right], \\
&\quad \le H\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{z'\neq z}\mathbb{E}_{\mathcal{H}_t\sim\mathbb{P}_z^{\pi_{1:t-1}}}\mathbb{E}_{\tau_h^t\sim\mathbb{P}_z^{\pi^t}}\left[\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}_h^t)\right] \\
&\quad \le \log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)H\eta^{-1}\cdot\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\mathcal{H}_t\sim\mathbb{P}_z^{\pi_{1:t-1}}}\mathbb{E}_{\tau_h^t\sim\mathbb{P}_z^{\pi^t}}\left[\min\left\{1/|\mathcal{X}_{\texttt{exp}}^{t-1}|,1\right\}\right], \qquad \text{(C.8)}
\end{aligned}$$

Note that $\mathbb{1}(\pi^t = \pi_{\texttt{exp}}) \overset{\text{iid}}{\sim} \text{Bernuolli}(\epsilon)$ for all episode $t \in [T]$. Thus, Bernstein's inequality indicates that with probability at least $1-\delta$, the following event $\mathcal{E}_2$ holds:

$$\sum_{t=1}^{T}\min\left\{1/|\mathcal{X}_{\texttt{exp}}^{t-1}|,1\right\} \le \mathcal{O}\left(\log T\cdot\epsilon^{-1} + \sqrt{T}\log(1/\delta)\right). \qquad \text{(C.9)}$$

Combine (C.4), (C.8) and (C.9), it follows that

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

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

C.3 Proof of Lemma C.1

Lemma C.1. (Online Guarantee) 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

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

where denote $\breve{\tau}_{h/t}^i = \tau_H^i$ for all $i \in [t-1]$ and $\breve{\tau}_{h/t}^t = \tau_h^t$, and $\mathbb{P}_z^{\pi}(\tau_h)$ is defined in (A.3).

Proof of Lemma C.1. For all $(h,t,z') \in [H]\times[T]\times\mathcal{Z}$, the information gain follows that

$$
\begin{aligned}
L_{h,t}(z') &:= \sum_{i\in[t]} \log\left(\frac{\mathbb{P}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}_z(\breve{\tau}^i_{h/t})}\right) = \sum_{i\in[t]} \log\left(\frac{\mathbb{P}^{\pi^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\pi^i}_z(\breve{\tau}^i_{h/t})}\right) \\
&\le 2\cdot \sum_{i\in[t]} \mathbb{E}_{\breve{\tau}^i_{h/t}\sim\mathbb{P}^{\pi^i}_z}\left[\sqrt{\frac{\mathbb{P}^{\pi^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\pi^i}_z(\breve{\tau}^i_{h/t})}} - 1\right] + \log(|\mathcal{Z}|/\delta) \\
&= -2\cdot \sum_{i\in[t]} D^2_{\mathrm{H}}\big(\mathbb{P}^{\pi^i}_z(\breve{\tau}^i_{h/t}), \mathbb{P}^{\pi^i}_{z'}(\breve{\tau}^i_{h/t})\big) + \log(|\mathcal{Z}|/\delta),
\end{aligned}
\tag{C.10}
$$

with probability greater than $1-\delta$, and $\mathbb{P}_z(\tau_h)$ and $\mathbb{P}^\pi_z(\tau_h)$ are defined in (A.3). Specifically, the inequality in (C.10) follows that $\log(1-x) \le -x$ for all $x \le 1$, and Lemma F.1 by defining the filtration $(\mathscr{F}_{t'})_{t'\le t}$ as $\mathscr{F}_{t'} = \sigma(\{\omega^i, \tau^i_H\}_{i\le t'-1})$ and taking a union bound over $\mathcal{Z}$. Recall that $\mathtt{prompt}^t_h = \{\omega^i, \tau^i_H\}_{i\in[t-1]} \cup \{\omega^t, \tau^t_h\}$ where $\{\omega^i\}_{i\le t}$ is assigned independent of z, then the posterior probability satisfies that

$$
\mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}^t_h) \propto \exp\big(L_{h,t}(z') + \log \mathcal{P}_{\mathcal{Z}}(z')\big). \tag{C.11}
$$

Based on (C.11) and the Donsker-Varadhan representation in Lemma F.2, it holds

$$
\begin{aligned}
&\sum_{z'\in\mathcal{Z}} L_{h,t}(z')\cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}^t_h) - D_{\mathrm{KL}}\big(\mathbb{P}_{\mathcal{D}}(\cdot \,|\, \mathtt{prompt}^t_h) \,\|\, \mathcal{P}_{\mathcal{Z}}(\cdot)\big) \\
&\qquad \ge \sum_{z'\in\mathcal{Z}} L_{h,t}(z')\cdot \mathcal{Q}(z') - D_{\mathrm{KL}}\big(\mathcal{Q}(\cdot) \,\|\, \mathcal{P}_{\mathcal{Z}}(\cdot)\big), \quad \forall \mathcal{Q} \in \Delta(\mathcal{Z}).
\end{aligned}
\tag{C.12}
$$

Denote $\delta_z(\cdot)$ as the Dirac distribution over singleton z. By taking $\mathcal{Q}(\cdot) = \delta_z(\cdot)$ in (C.12), we have

$$
\sum_{z'\in\mathcal{Z}} L_{h,t}(z')\cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}^t_h) \ge D_{\mathrm{KL}}\big(\mathbb{P}_{\mathcal{D}}(\cdot \,|\, \mathtt{prompt}^t_h) \,\|\, \mathcal{P}_{\mathcal{Z}}(\cdot)\big) + \log \mathcal{P}_{\mathcal{Z}}(z) \ge \log \mathcal{P}_{\mathcal{Z}}(z). \tag{C.13}
$$

Thus, for all $(h,t) \in [H]\times[T]$, with probability at least $1-\delta$, it holds that

$$
\begin{aligned}
&\sum_{z'\in\mathcal{Z}} \sum_{i\in[t]} D^2_{\mathrm{H}}\big(\mathbb{P}^{\pi^i}_z(\breve{\tau}^i_{h/t}), \mathbb{P}^{\pi^i}_{z'}(\breve{\tau}^i_{h/t})\big)\cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}^t_h) \\
&\qquad \le -\frac{1}{2}\sum_{z'\in\mathcal{Z}} L_{h,t}(z')\cdot \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathtt{prompt}^t_h) + \frac{1}{2}\log\left(|\mathcal{Z}|/\delta\right) \le \log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right),
\end{aligned}
\tag{C.14}
$$

where the first inequality results from (C.10), and the last inequality follows (C.13) and Assumption 4.5, which suggests that $\mathcal{P}_{\mathcal{Z}}(z) \ge (c_{\mathcal{Z}}|\mathcal{Z}|)^{-1}$. Therefore, we conclude the proof of Lemma C.1. $\square$

C.4 Proof of Proposition 4.3

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). Suppose that the high-level POMDP is fully observable, i.e., $\mathbb{O}(s) = s$, with $H = 2$, and $|\Omega|$=1. Let $\mathcal{S} =$

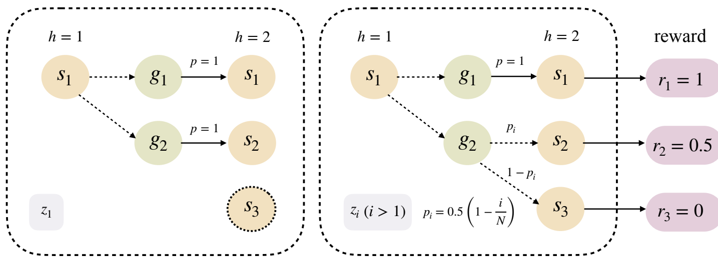

Figure 3: Illustration of hard-to-distinguish example in Proposition 4.3. The left figure demonstrates the state transition probability with respect to latent variable z_1, and the right figure shows the transition concerning z_i $(i>1)$. As $|\Omega|=1$, rewards are fixed for each state.

$\{s_1,s_2,s_3\}$ with reward $r(s_1)=0.5$, $r(s_2)=1$, and $r(s_3)=0$, $\mathcal{G}=\{g_1,g_2\}$, and $\mathcal{Z}=\{z_1,\ldots,z_N\}$. Starting from fixed initial state s_1 with latent variable z_1, the transition kernel is

$$\begin{cases} \mathbb{P}_{z_i}(s_1\,|\,s_1,g_1)=1, & \mathbb{P}_{z_i}(s_2\,|\,s_1,g_1)=0, \quad \mathbb{P}_{z_i}(s_3\,|\,s_1,g_1)=0, \qquad \forall i\in[N], \\ \mathbb{P}_{z_1}(s_1\,|\,s_1,g_2)=0, & \mathbb{P}_{z_1}(s_1\,|\,s_1,g_2)=2, \quad \mathbb{P}_{z_1}(s_3\,|\,s_1,g_2)=0, \qquad \text{if } i=1, \\ \mathbb{P}_{z_i}(s_1\,|\,s_1,g_2)=0, & \mathbb{P}_{z_i}(s_2\,|\,s_1,g_2)=p_i, \quad \mathbb{P}_{z_i}(s_3\,|\,s_1,g_2)=1-p_i, \quad \text{if } i\neq 1, \end{cases}$$

where $p_i=0.5(1-\frac{i}{N})$ for all $i\in[N]$. Thus, 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 a uniform. At episode $t=1$, without any information, the posterior distribution $\mathbb{P}(\cdot\,|\,\texttt{prompt}_1)$ degenerates to the prior $\mathcal{P}_{\mathcal{Z}}(\cdot)=\mathrm{Unif}_{\mathcal{Z}}(\cdot)$. Following this, the LLM's recommended policy at 1-st step follows that

$$\pi_{\texttt{LLM}}(\cdot\,|\,s_1)=\sum_{z\in\mathcal{Z}}1/N\cdot\pi^*_z(\cdot\,|\,s_1)=(1-1/N)\cdot\delta_{g_1}(\cdot\,|\,s_1)+1/N\cdot\delta_{g_2}(\cdot\,|\,s_1),$$

where δ_g denotes the Dirac distribution over singleton g. 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 $i\in[N]$, recommending g_1 by LLM does not provide information to differentiate the environment $z_i\in\mathcal{Z}$, and the posterior distribution remains uniform. Furthermore, this situation, denoted as $\mathbb{P}(\cdot\,|\,\texttt{prompt}_t)=\mathrm{Unif}_{\mathcal{Z}}(\cdot)$ and $\pi_t=\pi_{\texttt{LLM}}$, stops only if LLM suggests g_2 at epsiode $t\in[T]$. Consider specific trajectory $\tau_{\text{hard}}=(s_1,g_1,s_1)_{t\in[T]}$ which is achievable only if LLM consistently adheres to the initial policy $\pi_{\texttt{LLM}}$ and recommends taking subgoal g_1. Thus, the probability of trajectory τ_{hard} is $\mathbb{P}_{z_1}(\tau_{\text{hard}})=(1-1/N)^T$, indicating that $\mathrm{Reg}_{z_1}(T)\geq 0.5T\cdot(1-1/N)^T$. If $T=N$, we have the frequentist regret concerning z_1 at the rate of $0.5T\cdot(1-1/T)^T=\Omega(T)$, and complete the proof of Proposition 4.3. □

D Proof of Performance under Practical Setting in Section 5

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

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

and the contrastive learning algorithm in (3.7) follows $\widehat{\phi} = \mathrm{argmax}_{f\in\mathcal{F}}\ \widehat{\mathbb{E}}_{\mathcal{D}_{\texttt{Rep}}}\big[\log \mathbb{D}_\phi(y\,|\,o,s)\big]$, and thus $f_{\widehat{\phi}}$ is the maximum likelihood estimator (MLE) concerning the dataset $\mathcal{D}_{\texttt{Rep}}$. Based on Lemma F.3, the MLE-type algorithm ensures that, with probability at least $1-\delta$, it holds that

$$\bar{\mathbb{E}}_{(o,s)\sim\mathcal{D}_{\texttt{Rep}}}\left[D^2_{\mathrm{TV}}\left(\mathbb{D}_{\widehat{\phi}}(\cdot\,|\,o,s),\mathbb{D}(\cdot\,|\,o,s)\right)\right] \le 2\log(N_{\mathrm{p}}T_{\mathrm{p}}H|\mathcal{F}|/\delta)/N_{\mathrm{p}}T_{\mathrm{p}}H, \tag{D.1}$$

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

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

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

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s) &= \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s,y=0)\cdot\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y=0\,|\,s) + \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s,y=1)\cdot\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y=1\,|\,s)\\
&= \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s,y=0)\cdot\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y=0) + \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s,y=1)\cdot\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y=1) := \bar{\mathbb{O}}(o\,|\,s),
\end{aligned} \tag{D.3}$$

where the second equation results from the fact that contrastive data are labeled independent of data itself such that $\mathbb{P}(s\,|\,y) = \mathbb{P}(s)$ for all $y\in\{0,1\}$. Based on (D.3), we can get

$$\bar{\mathbb{E}}_{(o,s)\sim\mathcal{D}_{\texttt{Rep}}}\left[\left(\frac{\bar{\mathbb{O}}_{\widehat{\phi}}(o\,|\,s) - \bar{\mathbb{O}}(o\,|\,s)}{\bar{\mathbb{O}}(o\,|\,s)}\right)^2\right] = \bar{\mathbb{E}}_{s\sim\mathcal{D}_{\texttt{Rep}}}\left[\mathbb{E}_{o\sim\bar{\mathbb{O}}(\cdot\,|\,s)}\left[\left(\frac{\bar{\mathbb{O}}_{\widehat{\phi}}(\cdot\,|\,s) - \bar{\mathbb{O}}(\cdot\,|\,s)}{\bar{\mathbb{O}}(\cdot\,|\,s)}\right)^2\right]\right], \tag{D.4}$$

where equations results from the fact that $\mathbb{P}_{\mathcal{D}}(o,s) = \bar{\mathbb{O}}(o\,|\,s)\cdot\mathbb{P}_{\mathcal{D}}(s)$ and definition of χ^2-divergence. Therefore, combine (D.2) and (D.4), it holds that

$$\bar{\mathbb{E}}_{(o,s)\sim\mathcal{D}_{\texttt{Rep}}}\left[D^2_{\mathrm{TV}}\left(\mathbb{D}_{\widehat{\phi}}(\cdot\,|\,o,s),\mathbb{D}(\cdot\,|\,o,s)\right)\right] \le \frac{1}{(1+R_{\mathcal{F}})^2}\cdot\bar{\mathbb{E}}_{s\sim\mathcal{D}_{\texttt{Rep}}}\left[D_{\chi^2}\left(\bar{\mathbb{O}}_{\widehat{\phi}}(\cdot\,|\,s)\,\|\,\bar{\mathbb{O}}(\cdot\,|\,s)\right)\right]. \tag{D.5}$$

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

$$\begin{aligned}
D_{\chi^2}\left(\bar{\mathbb{O}}_{\widehat{\phi}}(\cdot\,|\,s)\,\|\,\bar{\mathbb{O}}(\cdot\,|\,s)\right) &= \sup_{g:\mathcal{O}\mapsto\mathbb{R}}\left\{\frac{\left(\mathbb{E}_{\bar{\mathbb{O}}_{\widehat{\phi}}}[g(o)\,|\,s] - \mathbb{E}_{\bar{\mathbb{O}}}[g(o)\,|\,s]\right)^2}{\mathrm{Var}_{\bar{\mathbb{O}}}[g(o)\,|\,s]}\right\}\\
&= \sup_{g:\mathcal{O}\mapsto\mathbb{R}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\phi}}}[g(o)\,|\,s] - \mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{4\cdot\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\cdot\frac{\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}{\mathrm{Var}_{\bar{\mathbb{O}}}[g(o)\,|\,s]}\right\}\\
&\ge \sup_{\substack{g:\mathcal{O}\mapsto\mathbb{R},\\ \mathbb{E}_{\mathbb{O}}[g(o)|s]=0}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\phi}}}[g(o)|s] - \mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{4\cdot\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\cdot\frac{\mathbb{E}_{\mathbb{O}}[g(o)^2\,|\,s]}{\mathbb{E}_{\bar{\mathbb{O}}}[g(o)^2\,|\,s]}\right\}
\end{aligned} \tag{D.6}$$

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

$$\frac{\mathbb{E}_{\mathbb{O}}[g(o)^2\,|\,s]}{\mathbb{E}_{\bar{\mathbb{O}}}[g(o)^2\,|\,s]}=2\left(1+\frac{\mathbb{E}_{\mathcal{P}^-}[g(o)^2\,|\,s]}{\mathbb{E}_{\mathbb{O}}[g(o)^2\,|\,s]}\right)^{-1}\leq 2\left(1+\left\|\frac{\mathcal{P}^-(\cdot)}{\mathbb{O}(\cdot\,|\,s)}\right\|_\infty\right)^{-1}\leq 2(1+R_{\mathcal{F}}^-)^{-1}, \tag{D.7}$$

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

$$\begin{aligned}\sup_{\substack{g:\mathcal{O}\mapsto\mathbb{R},\\ \mathbb{E}_{\mathbb{O}}[g(o)|s]=0}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\phi}}}[g(o)\,|\,s]-\mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\right\}&=\sup_{g:\mathcal{O}\mapsto\mathbb{R}}\left\{\frac{\left(\mathbb{E}_{\mathbb{O}_{\widehat{\phi}}}[g(o)\,|\,s]-\mathbb{E}_{\mathbb{O}}[g(o)\,|\,s]\right)^2}{\mathrm{Var}_{\mathbb{O}}[g(o)\,|\,s]}\right\}\\ &=D_{\chi^2}\left(\mathbb{O}_{\widehat{\phi}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right),\end{aligned} \tag{D.8}$$

Based on (D.1), (D.5), (D.6), (D.7) and (D.8), then we have

$$\bar{\mathbb{E}}_{\mathcal{D}_{\mathrm{Rep}}}\left[D_{\chi^2}\left(\mathbb{O}_{\widehat{\phi}}(\cdot\,|\,s)\,\|\,\mathbb{O}(\cdot\,|\,s)\right)\right]\leq\mathcal{O}\left(\frac{(1+R_{\mathcal{F}}^-)(1+R_{\mathcal{F}})^2}{N_{\mathrm{p}}T_{\mathrm{p}}H}\cdot\log(N_{\mathrm{p}}T_{\mathrm{p}}H|\mathcal{F}|/\delta)\right). \tag{D.9}$$

Combine (D.9) and the divergence inequalities (Polyanskiy and Wu, 2022), we have

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

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

D.2 Proof of Theorem 5.7

Proof of Theorem 5.7. To clarify the notations, $\mathcal{J}$ and $\widehat{\mathcal{J}}$ denote the value function concerning the ground-truth $\mathbb{O}$ and the pretrained $\mathbb{O}_{\widehat{\phi}}$; π_z^* and $\widehat{\pi}_z^*$ represent the optimal policy under the ground-truth and pretrained envrionment; π^t and $\widehat{\pi}^t$ are the `Planner`'s policy in Algorithm 1 empowered by perfect `LLM` or pretrained $\texttt{LLM}_{\widehat{\theta}}$. Condition 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

$$\begin{aligned}\mathrm{Reg}_z(T)&=\sum_{t=1}^T\mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_z(\widehat{\pi}_z^*,\omega^t)-\widehat{\mathcal{J}}_z(\widehat{\pi}^t,\omega^t)\right]\\ &\leq\underbrace{\sum_{t=1}^T\widehat{\mathcal{J}}_z(\widehat{\pi}_z^*,\omega^t)-\mathcal{J}_z(\widehat{\pi}_z^*,\omega^t)}_{\textbf{(i)}}+\underbrace{\sum_{t=1}^T\mathcal{J}_z(\widehat{\pi}_z^*,\omega^t)-\mathcal{J}_z(\pi_z^*,\omega^t)}_{\textbf{(ii)}}\\ &\quad+\underbrace{\sum_{t=1}^T\mathcal{J}_z(\pi_z^*,\omega^t)-\widehat{\mathcal{J}}_z(\pi_z^*,\omega^t)}_{\textbf{(iii)}}+\underbrace{\sum_{t=1}^T\mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_z(\pi_z^*,\omega^t)-\widehat{\mathcal{J}}_z(\widehat{\pi}^t,\omega^t)\right]}_{\textbf{(iv)}},\end{aligned} \tag{D.10}$$

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

$$\begin{aligned}
\sum_{t=1}^{T} \widehat{\mathcal{J}}_z(\pi_t, \omega^t) - \mathcal{J}_z(\pi_t, \omega^t) &= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \mathbb{P}_z^{\pi_t}} \left[(\mathbb{P}_{z,h} \widehat{V}_h^{\pi_t} - \widehat{\mathbb{P}}_{z,h} \widehat{V}_h^{\pi_t})(s_h, \tau_h, \omega^t) \right] \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} 2H \cdot \mathbb{E}_{(s_h^t, \tau_h^t) \sim \mathbb{P}_z^{\pi_t}} \left[D_{\mathrm{TV}} \left(\mathbb{P}_{z,h}(\cdot \,|\, s_h, \tau_h), \widehat{\mathbb{P}}_{z,h}(\cdot \,|\, s_h, \tau_h) \right) \right] \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} 2H^2 \cdot \max_{s \in \mathcal{S}} \left\{ D_{\mathrm{TV}} \left(\mathbb{O}(\cdot \,|\, s), \mathbb{O}_{\widehat{\phi}}(\cdot \,|\, s) \right) \right\}, \qquad \text{(D.11)}
\end{aligned}$$

where $\mathbb{P}_{z,h}$ and $\widehat{\mathbb{P}}_{z,h}$ respectively denote the probability distribution concerning ground-truth $\mathbb{O}$ and pretrained $\mathbb{O}_{\widehat{\phi}}$. Based on (D.11), by taking policies $\widehat{\pi}_z^*$ and π_z^*, then we have

$$\begin{aligned}
\textbf{(i)} + \textbf{(iii)} &= \sum_{t=1}^{T} \widehat{\mathcal{J}}_z(\widehat{\pi}_z^*, \omega^t) - \mathcal{J}_z(\widehat{\pi}_z^*, \omega^t) + \sum_{t=1}^{T} \mathcal{J}_z(\pi_z^*, \omega^t) - \widehat{\mathcal{J}}_z(\pi_z^*, \omega^t) \\
&\leq 4H^3 T \lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta). \qquad \text{(D.12)}
\end{aligned}$$

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

$$\pi_h^t(\cdot \,|\, \tau_h^t, \omega^t) \sim (1-\epsilon) \cdot \widehat{\pi}_{h,\texttt{LLM}}^t(\cdot \,|\, \tau_h^t, \omega^t) + \epsilon \cdot \pi_{h,\texttt{exp}}(\cdot | \tau_h^t), \qquad \text{(D.13)}$$

where $\widehat{\pi}_{h,\texttt{LLM}}^t(\cdot \,|\, \tau_h^t, \omega^t) = \texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{prompt}_h^t)$ and $\texttt{prompt}_h^t = \mathcal{H}_t \cup \{\omega^t, \tau_h^t\}$. Based on PDL in Lemma F.4, the performance difference in term **(iv)** can be decomposed as

$$\begin{aligned}
\textbf{(iv)} &= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t} \left[\widehat{\mathcal{J}}_z(\pi_z^*, \omega^t) - \widehat{\mathcal{J}}_z(\widehat{\pi}^t, \omega^t) \right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[(\pi_{z,h}^* - \widehat{\pi}_h^t) \widehat{Q}_h^{\pi_z^*}(s_h^t, \tau_h^t, \omega^t) \right] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[\left(\pi_{z,h}^* - \widehat{\pi}_{h,\texttt{LLM}}^t \right) \widehat{Q}_h^{\pi_z^*}(s_h^t, \tau_h^t, \omega^t) \right] \cdot (1-\epsilon) \\
&\quad + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{exp}} \right) \widehat{Q}_h^{\pi_z^*}(s_h^t, \tau_h^t, \omega^t) \right] \cdot \epsilon \\
&\leq 2H \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}} \left(\pi_{z,h}^*(\cdot \,|\, \tau_h^t, \omega^t), \texttt{LLM}_{\widehat{\theta}}(\cdot \,|\, \texttt{prompt}_h^t) \right) \right] + HT\epsilon \qquad \text{(D.14)}
\end{aligned}$$

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

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}}\left(\pi_{z,h}^*(\cdot\,|\,\tau_h^t, \omega^t), \mathtt{LLM}_{\widehat{\theta}}(\cdot\,|\,\mathtt{prompt}_h^t)\right)\right] \\
&\quad \le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}}\left(\mathtt{LLM}_{\widehat{\theta}}(\cdot\,|\,\mathtt{prompt}_h^t), \mathtt{LLM}(\cdot\,|\,\mathtt{prompt}_h^t)\right)\right] \\
&\qquad + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}}\left(\pi_{z,h}^*(\cdot\,|\,\tau_h^t, \omega^t), \mathtt{LLM}(\cdot\,|\,\mathtt{prompt}_h^t)\right)\right] \\
&\quad \le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}}\left(\mathtt{LLM}_{\widehat{\theta}}(\cdot\,|\,\mathtt{prompt}_h^t), \mathtt{LLM}(\cdot\,|\,\mathtt{prompt}_h^t)\right)\right] \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[\sum_{z'\neq z} \mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{prompt}_h^t)\right],
\end{aligned}
\tag{D.15}
$$

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

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}} \left[D_{\mathrm{TV}}\left(\mathtt{LLM}_{\widehat{\theta}}(\cdot\,|\,\mathtt{prompt}_h^t), \mathtt{LLM}(\cdot\,|\,\mathtt{prompt}_h^t)\right)\right] \\
&\quad \le \lambda_S \cdot \sum_{t=1}^{T}\sum_{h=1}^{H} \bar{\mathbb{E}}_{\mathtt{prompt}_h^t\sim\mathcal{D}_{\mathtt{LLM}}} \left[D_{\mathrm{TV}}\left(\mathtt{LLM}_{\widehat{\theta}}(\cdot\,|\,\mathtt{prompt}_h^t), \mathtt{LLM}(\cdot\,|\,\mathtt{prompt}_h^t)\right)\right], \\
&\quad = \lambda_S HT \cdot \Delta_{\mathtt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta),
\end{aligned}
\tag{D.16}
$$

where the last inequality follows Theorem 5.3 and Assumption 5.6. Under practical setting, $\mathtt{prompt}_h^t$ is generated from practical transition $\widehat{\mathbb{P}}_z$, mismatching the conditional distribution $\mathbb{P}_{\mathcal{D}}(z\,|\,\mathtt{prompt}_h^t)$ under the perfect scenarios 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 $i \in [t-1]$ and $\breve{\tau}_{h/t}^t = \tau_h^t$. Define information gains as

$$
L_{h,t}^{\mathtt{exp}}(z') = \sum_{i\in\mathcal{X}_{\mathtt{exp}}^t} \log\left(\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right), \quad L_{h,t}^{\mathtt{LLM}}(z') = \sum_{i\in[t]\setminus\mathcal{X}_{\mathtt{exp}}^t} \log\left(\frac{\mathbb{P}_{z'}(\breve{\tau}_{h/t}^i)}{\mathbb{P}_z(\breve{\tau}_{h/t}^i)}\right), \tag{D.17}
$$

where $\mathbb{P}_z(\tau_h)$ is defined in (A.3). Based on the Baye's Theorem, we have

$$
\mathbb{P}_{\mathcal{D}}(z'\,|\,\mathtt{prompt}_h^t) \le \frac{\mathbb{P}_{z'}(\mathtt{prompt}_h^t)}{\mathbb{P}_z(\mathtt{prompt}_h^t)} \cdot \frac{\mathcal{P}_{\mathcal{Z}}(z')}{\mathcal{P}_{\mathcal{Z}}(z)}. \tag{D.18}
$$

Let $\mathcal{E}_2$ be the event that Lemma D.1 holds. Based on (D.18), (D.17) and condition on event $\mathcal{E}_2$, it holds that

$$
\begin{aligned}
\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}_h^t) &\leq \min\left\{\sum_{z'\neq z}\frac{\mathbb{P}_{z'}(\texttt{prompt}_h^t)}{\mathbb{P}_{z}(\texttt{prompt}_h^t)}\cdot\frac{\mathcal{P}_{\mathcal{Z}}(z')}{\mathcal{P}_{\mathcal{Z}}(z)},1\right\}\\
&\leq \min\left\{c_{\mathcal{Z}}\sum_{z'\neq z}\exp\left(L_{h,t}^{\texttt{exp}}(z')+L_{h,t}^{\texttt{LLM}}(z')\right),1\right\}\\
&\leq \min\left\{c_{\mathcal{Z}}\sum_{z'\neq z}\exp\left(t\cdot H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2-2\eta|\mathcal{X}_{\texttt{exp}}^t|+8\log(|\mathcal{Z}|/\delta)+2\eta\right),1\right\}\\
&\leq \min\left\{c_{\mathcal{Z}}\sum_{z'\neq z}\exp\left(-\left(\eta\epsilon-H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)t+8\log(|\mathcal{Z}|/\delta)+2\eta\right),1\right\}\\
&\leq \min\left\{c_{\mathcal{Z}}\cdot\exp\left(-\left(\eta\epsilon-H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)t+9\log(|\mathcal{Z}|/\delta)+2\eta\right),1\right\} \qquad (\text{D.19})
\end{aligned}
$$

for all $(h,t)\in[H]\times[T]$, where the second inequality follows Assumption 4.5, and we suppose that $|\mathcal{X}_{\texttt{exp}}^t|/t=\epsilon$ is fixed for simplicity, which is attainable if we explore at a fixed fraction in episodes. Assume that $\eta\epsilon\geq H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2$ holds temporarily. Following (D.19) and condition on event $\mathcal{E}_2$, there exists a large constant $c_0>0$ such that

$$
\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}_h^t)\leq c_0\cdot H\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\left(\eta\epsilon-H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)^{-1}, \qquad (\text{D.20})
$$

where we use the fact that there exists constant $c_0>0$ such that $\sum_{t=1}^{T}\min\{c_3\exp(-c_1t+c_2),1\}\leq c_0\cdot c_1^{-1}(c_2+\log c_3)$. Furthermore, based on (D.20), we can show that

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}}\mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}}\left[\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}_h^t)\right]\\
&\quad\leq\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{z'\neq z}\mathbb{E}_{\mathcal{H}_t\sim\widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}}\mathbb{E}_{\tau_h^t\sim\widehat{\mathbb{P}}_z^{\widehat{\pi}^t}}\left[\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}_h^t)\,\mathbb{1}\left(\mathcal{E}_2\text{ holds}\right)\right]+2HT\delta\\
&\quad\leq c_0\cdot H\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\left(\eta\epsilon-H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)^{-1}+2HT\delta. \qquad (\text{D.21})
\end{aligned}
$$

Combine (D.14), (D.19), (D.16) and (D.21), it holds that

$$
\begin{aligned}
\textbf{(iv)}\leq&\ \underbrace{c_0\cdot H^2\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)\cdot\left(\eta\epsilon-H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)^{-1}}_{\textbf{(v)}}\\
&+\underbrace{HT\eta^{-1}\left(\eta\epsilon-H\lambda_R^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2\right)}_{\textbf{(vi)}}+2\lambda_S H^2T\cdot\Delta_{\texttt{LLM}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)\\
&+H^2T(\eta\lambda_R)^{-1}\cdot\Delta_{\texttt{Rep}}(N_{\mathrm{p}},T_{\mathrm{p}},H,\delta)^2+2HT\delta, \qquad (\text{D.22})
\end{aligned}
$$

If we explore with probability $\epsilon = H(\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\rm p}, T_{\rm p}, H, \delta)^2 + (H \log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta)/T\eta)^{1/2}$, which satisfies the condition $\eta\epsilon \geq H\lambda_R^{-1}\Delta_{\texttt{Rep}}(N_{\rm p}, T_{\rm p}, H, \delta)^2$ assumed in (D.19), then we have

$$\textbf{(v)} + \textbf{(vi)} \leq \mathcal{O}\left(H^{\frac{3}{2}}\sqrt{\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta) \cdot T/\eta}\right). \tag{D.23}$$

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

Combine (D.10), (D.12), (D.22) and (D.23), the regret under the practical setting follows

$$\begin{aligned}\mathrm{Reg}_z(T) &\leq \textbf{(i)} + \textbf{(iii)} + \textbf{(iv)} + 2HT \cdot \mathbb{P}(\mathcal{E}_1 \text{ fails}) \\ &= \mathcal{O}\Big(\underbrace{H^{\frac{3}{2}}\sqrt{\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta) \cdot T/\eta}}_{\text{Planning error}} + \underbrace{H^2 T \cdot \Delta_{\rm p}(N_{\rm p}, T_{\rm p}, H, \delta, \xi)}_{\text{Pretraining error}}\Big) + 4HT\delta,\end{aligned} \tag{D.24}$$

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

$$\begin{aligned}\Delta_{\rm p}(N_{\rm p}, T_{\rm p}, H, \delta, \xi) &= (\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\rm p}, T_{\rm p}, H, \delta)^2 \\ &\quad + 4H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\rm p}, T_{\rm p}, H, \delta) + 2\lambda_S \cdot \Delta_{\texttt{LLM}}(N_{\rm p}, T_{\rm p}, H, \delta).\end{aligned}$$

Here, $\xi = (\eta, \lambda_S, \lambda_R)$ denotes the set of distinguishability and coverage coefficients in Definition 4.4 and Assumption 5.6, and $\Delta_{\texttt{LLM}}(N_{\rm p}, T_{\rm p}, H, \delta)$ and $\Delta_{\texttt{Rep}}(N_{\rm p}, T_{\rm p}, H, \delta)$ 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. $\square$

D.3 Proof of Lemma D.1

In this subsection, we provide a detailed examination of the concentration arguments with respect to the posterior probability when there exists mismatch between the ground-truth envrionment and the practical environment with pretrained modules. The argument is formalized as 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). $L_{h,t}^{\texttt{LLM}}(z') \leq \left(t - |\mathcal{X}_{\texttt{exp}}^t|\right) H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\rm p}, T_{\rm p}, H, \delta)^2 + 4\log(|\mathcal{Z}|/\delta)$,

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

where $L_{h,t}^{\texttt{LLM}}(z')$ and $L_{h,t}^{\texttt{exp}}(z')$ are the information gain defined in (D.17).

Proof of Lemma D.1. Similar to the proof of Theorem 5.7, we denote $\mathbb{P}_{z,h}$ and $\widehat{\mathbb{P}}_{z,h}$ respectively as the probability concerning $\mathbb{O}$ and $\mathbb{O}_{\widehat{\phi}}$. Consider a fixed tuple $(z', h, t) \in \mathcal{Z} \times [H] \times [T]$, it holds that

$$\begin{aligned}
&\widehat{\mathbb{P}}_z(L^{\texttt{LLM}}_{h,t}(z') \geq \beta^{\texttt{LLM}}_{h,t}) \leq \inf_{\lambda \geq 0} \mathbb{E}_{\tau^i_H \sim \widehat{\mathbb{P}}^{\widehat{\pi}^i}_z} \left[\exp(\lambda \cdot (L^{\texttt{LLM}}_{h,t}(z') - \beta^{\texttt{LLM}}_{h,t}))\right] \\
&\leq \inf_{\lambda \geq 0} \mathbb{E}_{\tau^i_H \sim \widehat{\mathbb{P}}^{\widehat{\pi}}_z} \left[\exp\left(\sum_{i \in [t] \setminus \mathcal{X}^t_{\exp}} \lambda \cdot \log\left(\frac{\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})}\right) - \lambda \cdot \beta^{\texttt{LLM}}_{h,t}\right)\right] \\
&= \inf_{\lambda \geq 0} \prod_{i \in [t] \setminus \mathcal{X}^t_{\exp}} \mathbb{E}_{\tau^i_H \sim \mathbb{P}^{\widehat{\pi}^i}_z} \left[\left(\frac{\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})}\right)^{\lambda} \cdot \frac{\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})}\right] \cdot \exp\left(-\lambda \cdot \beta^{\texttt{LLM}}_{h,t}\right) \\
&\leq \inf_{\lambda \geq 0} \prod_{i \in [t] \setminus \mathcal{X}^t_{\exp}} \mathbb{E}_{\tau^i_H \sim \mathbb{P}^{\widehat{\pi}^i}_z} \left[\left(\frac{\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})}\right)^{2\lambda}\right]^{1/2} \mathbb{E}_{\mathbb{P}^{\widehat{\pi}^i}_z} \left[\left(\frac{\widehat{\mathbb{P}}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t})}{\mathbb{P}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})}\right)^{2}\right]^{1/2} \cdot \exp\left(-\lambda \cdot \beta^{\texttt{LLM}}_{h,t}\right),
\end{aligned}$$

where the first inequality follows Lemma F.1 by defining the filtration $(\mathscr{F}_{t'})_{t' \leq t}$ as $\mathscr{F}_{t'} = \sigma(\{\omega^i, \tau^i_H\}_{i \leq t'-1})$, and the last inequality results from the Cauchy-Swartz inequality. By taking $\lambda = \frac{1}{4}$, we have

$$\begin{aligned}
\widehat{\mathbb{P}}_z\left(L^{\texttt{LLM}}_{h,t}(z') \geq \beta^{\texttt{LLM}}_{h,t}\right) &\leq \prod_{i \in [t] \setminus \mathcal{X}^t_{\exp}} \left(1 - D^2_{\mathrm{H}}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}), \mathbb{P}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})\big)\right)^{1/2} \\
&\quad \cdot \left(1 + D^2_{\chi}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}) \,\|\, \widehat{\mathbb{P}}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})\big)\right)^{1/2} \cdot \exp\left(-\frac{1}{4}\beta^{\texttt{LLM}}_{h,t}\right) \\
&\leq \exp\left(\sum_{i \in [t] \setminus \mathcal{X}^t_{\exp}} \frac{1}{2} \log\left(1 + D^2_{\chi}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}) \,\|\, \widehat{\mathbb{P}}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})\big)\right) - \frac{1}{4}\beta^{\texttt{LLM}}_{h,t}\right). \qquad \text{(D.25)}
\end{aligned}$$

Based on (D.25), Theorem 5.5 and Assumption 4.5, for all $(z', h, t) \in \mathcal{Z} \times [H] \times [T]$, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&1 + D^2_{\chi}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}) \,\|\, \widehat{\mathbb{P}}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t})\big) \leq 1 + D^2_{\chi}\big(\mathbb{P}^{\widehat{\pi}^i}_{z'}(\breve{\tau}^i_{h/t}, \breve{\mathbf{s}}^i_{h/t}) \,\|\, \widehat{\mathbb{P}}^{\widehat{\pi}^i}_z(\breve{\tau}^i_{h/t}, \breve{\mathbf{s}}^i_{h/t})\big) \\
&\quad \leq \left(1 + \max_{s \in \mathcal{S}} D_{\chi^2}\big(\mathbb{O}(\cdot \,|\, s) \,\|\, \mathbb{O}_{\widehat{\phi}}(\cdot \,|\, s)\big)\right)^H \leq \left(1 + \lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2\right)^H, \qquad \text{(D.26)}
\end{aligned}$$

where the first inequality results from the data processing inequality with $\breve{\mathbf{s}}^i_{h/t} = \breve{\mathbf{s}}^i_{1:H}$ for $i \in [t-1]$ and $\breve{\mathbf{s}}^i_{h/t} = \breve{\mathbf{s}}^t_{1:h}$, and the second inequality arises from tensorization. To ensure that $L^{\texttt{LLM}}_{h,t}(z') \geq \beta^{\texttt{LLM}}_{h,t}$ holds for all $(z', h, t) \in \mathcal{Z} \times [H] \times [T]$ with probability at least $1 - \delta$, we take constant $\beta^{\texttt{LLM}}_{h,t}$ as

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

which is based on (D.25), (D.26) 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 exploration episodes, we have

$$
\begin{aligned}
\widehat{\mathbb{P}}_z(L_{h,t}^{\mathtt{exp}}(z') \ge \beta_{h,t}^{\mathtt{exp}}) &\le \inf_{\lambda \ge 0} \mathbb{E}_{\tau_H^i \sim \widehat{\mathbb{P}}_z}\Big[\exp(\lambda \cdot (L_{h,t}^{\mathtt{exp}} - \beta_{h,t}^{\mathtt{exp}}))\Big] \\
&\le \prod_{i \in \mathcal{X}_{\mathtt{exp}}^t} \Big(1 - D_{\mathrm{H}}^2\big(\mathbb{P}_{z'}^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_z^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i)\big)\Big)^{1/2} \\
&\qquad \cdot \Big(1 + D_{\chi}^2\big(\mathbb{P}_{z'}^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i) \,\|\, \widehat{\mathbb{P}}_z^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i)\big)\Big)^{1/2} \cdot \exp\Big(-\frac{1}{4}\beta_{h,t}^{\mathtt{exp}}\Big) \\
&\le \exp\Bigg(\frac{1}{2}\sum_{i \in \mathcal{X}_{\mathtt{exp}}^t} -D_{\mathrm{H}}^2\Big(\mathbb{P}_{z'}^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_z^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i)\Big) \\
&\qquad + \frac{1}{2}\sum_{i \in \mathcal{X}_{\mathtt{exp}}^t} \log\Big(1 + D_{\chi}^2\big(\mathbb{P}_{z'}^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i) \,\|\, \widehat{\mathbb{P}}_z^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i)\big)\Big) - \frac{1}{4}\beta_{h,t}^{\mathtt{exp}}\Bigg),
\end{aligned} \tag{D.27}
$$

where the last inequality follows $\log(1-x) \le -x$ for all $x \le 1$. Based on Definition 4.4, we have

$$
\sum_{i \in \mathcal{X}_{\mathtt{exp}}^t} D_{\mathrm{H}}^2\big(\mathbb{P}_{z'}^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i), \mathbb{P}_z^{\widehat{\pi}^i}(\breve{\tau}_{h/t}^i)\big) \ge \eta \cdot |\mathcal{X}_{\mathtt{exp}}^{t-1}|. \tag{D.28}
$$

Combine the arguments in (D.26), (D.27) and (D.28), to ensure that $L_{h,t}^{\mathtt{exp}}(z') \ge \beta_{h,t}^{\mathtt{exp}}$ holds for all $(z', h, t) \in \mathcal{Z} \times [H] \times [T]$ with probability greater than $1 - \delta$, we set constant $\beta_{h,t}^{\mathtt{exp}}$ as

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

by taking a union bound over $\mathcal{Z} \times [H] \times [T]$, and then we finish 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

$$
f^*(o, s) = \frac{\mathbb{O}(o \,|\, s)}{\mathcal{P}^-(o)}.
$$

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

$$
\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y \,|\, o, s) = \frac{\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o, s \,|\, y) \cdot \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y)}{\sum_{y \in \{0,1\}} \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o, s \,|\, y) \cdot \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y)} = \frac{\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o \,|\, s, y) \cdot \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(s \,|\, y)}{\sum_{y \in \{0,1\}} \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o \,|\, s, y) \cdot \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(s \,|\, y)}, \tag{D.29}
$$

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

$$\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(\cdot\,|\,s, y=0) = \mathbb{O}(\cdot\,|\,s), \quad \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(\cdot\,|\,s, y=1) = \mathcal{P}^{-}(\cdot), \tag{D.30}$$

and data are labeled independent of data itself, such that $\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(s\,|\,y) = \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(s)$. By dividing both the numerator and denominator by $\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(s)$, then $\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(y\,|\,o,s) = \mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s,y)/(\mathcal{P}^{-}(o) + \mathbb{O}(o\,|\,s))$. Recall that the population loss is

$$\mathcal{R}_{\mathrm{CT}}(\phi; \mathcal{D}_{\mathtt{Rep}}) = \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbb{P}_{\phi}(\cdot|o,s)\,\|\,\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(\cdot|o,s)\right) + H_{\mathrm{s}}(\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(\cdot|o,s))\right],$$

and the minimum is achieved at $\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(\cdot\,|\,o,s) = \mathbb{P}_{\phi^*}(\cdot|o,s)$. Following the form of (5.1), the learning target satisfies that

$$\frac{\mathbb{P}_{\mathcal{D}_{\mathrm{r}}}(o\,|\,s,y)}{\mathcal{P}^{-}(o) + \mathbb{O}(o\,|\,s)} = \left(\frac{f^*(o,s)}{1+f^*(o,s)}\right)^{y}\left(\frac{1}{1+f^*(o,s)}\right)^{1-y}. \tag{D.31}$$

By solving the equation in (D.31), the learning target follows that $f^*(o,s) = \mathbb{O}(o\,|\,s)/\mathcal{P}^{-}(o)$ for the contrastive loss in (3.7), and then we conclude the proof of Lemma D.2. □

E Proof of Results for Extentions

E.1 Proof of Proposition B.1

Proof of Proposition B.1. For all $h \in [H]$, the marginal distribution of observation follows that

$$\mathbb{P}^{t}_{h,\mathtt{LLM}}(o_h\,|\,o_1, \mathrm{do}(\mathbf{g}_{1:h-1})) = \int_{\mathbf{o}_{2:h-1}} \mathbb{P}_{\mathcal{D}}\left(\mathbf{o}_{2:h}\,|\,o_1, \mathrm{do}(\mathbf{g}_{1:h-1}), \mathcal{H}_t\right)\, \mathrm{d}\mathbf{o}_{2:h-1},$$

$$\text{where } \mathbb{P}_{\mathcal{D}}\left(\mathbf{o}_{2:h}\,|\,o_1, \mathrm{do}(\mathbf{g}_{1:h-1}), \mathcal{H}_t\right) = \prod_{h'=1}^{h-1} \mathbb{P}_{\mathcal{D}}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right). \tag{E.1}$$

Similarly, we define the marginal distribution concerning latent variable z as

$$\mathbb{P}_{z}\left(\mathbf{o}_{2:h}\,|\,o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) = \prod_{h'=1}^{h-1} \mathbb{P}_{z,h}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'}\right). \tag{E.2}$$

Based on the Baye's Theorem and data generation distribution denoted by $\mathcal{D}_{\mathrm{p}}$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}\left(o_h\,|\,(\mathbf{o},\mathbf{g})_{1:h-1}, \mathcal{H}_t\right) &= \sum_{z\in\mathcal{Z}} \mathbb{P}_{z}\left(o_h\,|\,(\mathbf{o},\mathbf{g})_{1:h-1}\right)\cdot \mathbb{P}_{\mathcal{D}}\left(z\,|\,(\mathbf{o},\mathbf{g})_{1:h-1}, \mathcal{H}_t\right)\\ &= \sum_{z\in\mathcal{Z}} \mathbb{P}_{z}\left(o_h\,|\,(\mathbf{o},\mathbf{g})_{1:h-1}\right)\cdot \frac{\prod_{h'=1}^{h-2}\mathbb{P}_{z}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'}\right)\cdot \mathbb{P}_{\mathcal{D}}(z\,|\,\mathcal{H}_t)}{\prod_{h'=1}^{h-2}\mathbb{P}_{\mathcal{D}}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)}\\ &= \sum_{z\in\mathcal{Z}} \frac{\mathbb{P}_{z}\left(\mathbf{o}_{2:h}\,|\,o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right)\cdot \mathbb{P}_{\mathcal{D}}(z\,|\,\mathcal{H}_t)}{\prod_{h'=1}^{h-2}\mathbb{P}_{\mathcal{D}}\left(o_{h'+1}\,|\,(\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)},\end{aligned} \tag{E.3}$$

Following the arguments in (E.3) and definitions in (E.1) and (E.2), it holds that

$$\mathbb{P}_{\mathcal{D}}\left(\mathbf{o}_{2:h} \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1}), \mathcal{H}_t\right) = \sum_{z \in \mathcal{Z}} \mathbb{P}_z\left(\mathbf{o}_{2:h} \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t), \tag{E.4}$$

Based on (E.4) and integration taken over $\mathbf{o}_{2:h}$ at both side, we can get

$$\mathbb{P}^t_{h,\texttt{LLM}}\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) = \sum_{z \in \mathcal{Z}} \mathbb{P}_{z,h}\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) \cdot \mathbb{P}_{\mathcal{D}}(z \,|\, \mathcal{H}_t), \tag{E.5}$$

and then we complete the proof of Proposition B.1. □

E.2 Proof of Corollary B.3

Proof of Corollary B.3. To clarify notations, $\mathcal{J}$, $\widehat{\mathcal{J}}$ and $\mathcal{J}_{t,\texttt{LLM}}$ respectively denote the value function with ground-truth translation distribution $\mathbb{O}$, pretrained $\mathbb{O}_{\widehat{\phi}}$ and simulated environment concerning pretrained $\texttt{LLM}_{\widehat{\theta}}$; $\widehat{\pi}^*_z$ and $\widehat{\pi}^{t,*}_{\texttt{LLM}}$ are the optimal policy under the practical setting and simulated environment; $\widehat{\pi}^t$ is the policy in Algorithm 2 with pretrained $\texttt{LLM}_{\widehat{\theta}}$. Condition 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

$$\begin{aligned}
\mathrm{Reg}_z(T) \leq\; & \underbrace{\sum_{t=1}^T \widehat{\mathcal{J}}_z(\widehat{\pi}^*_z, \omega^t) - \mathcal{J}_z(\widehat{\pi}^*_z, \omega^t)}_{\textbf{(i)}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\mathcal{J}_z(\widehat{\pi}^*_z, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^*_z, \omega^t)\right]}_{\textbf{(ii)}} \\
& + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^*_z, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t)\right]}_{\textbf{(iii)}} \\
& + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t) - \mathcal{J}_z(\widehat{\pi}^t, \omega^t)\right]}_{\textbf{(iv)}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{\mathcal{H}_t}\left[\mathcal{J}_z(\widehat{\pi}^t, \omega^t) - \widehat{\mathcal{J}}_z(\widehat{\pi}^t, \omega^t)\right]}_{\textbf{(v)}}.
\end{aligned} \tag{E.6}$$

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

Similar to (D.11) in the proof of Theorem 5.7, it holds that

$$\textbf{(i)} + \textbf{(vi)} \leq 4H^3 T \lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta), \tag{E.7}$$

following the pretraining error in Theorem 5.5.

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

Recall that `Planner` conducts task planning via the mixture policy:

$$\pi^t_h(\cdot \,|\, \tau^t_h, \omega^t) \sim (1-\epsilon) \cdot \widehat{\pi}^{t,*}_{h,\texttt{LLM}}(\cdot \,|\, \tau^t_h, \omega^t) + \epsilon \cdot \pi_{h,\texttt{exp}}(\cdot | \tau^t_h, \omega^t), \tag{E.8}$$

Following this, it holds that

$$
\begin{aligned}
\textbf{(iii)} &= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\Big[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t)\Big] + \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\Big[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t)\Big] \\
&\le \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\Big[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t) - (1-\epsilon)\cdot \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_{\texttt{LLM}}^{t,*}, \omega^t) - \epsilon \cdot \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi_{\texttt{exp}}, \omega^t)\Big] \le 2HT\epsilon, \qquad \text{(E.9)}
\end{aligned}
$$

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

$$
\begin{aligned}
&\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_z(\pi, \omega^t) = \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_{t,\texttt{LLM}}(\pi, \omega^t) + \mathcal{J}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_z(\pi, \omega^t) \\
&\quad \le \mathbb{E}_{\mathbf{g}_{1:h-1}\sim\pi}\underbrace{\left[\sum_{h=1}^{H}\int_{o_h}\Big(\widehat{\mathbb{P}}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) - \mathbb{P}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right)\Big)\,\mathrm{d}o_h\right]}_{\textbf{(vi)}} \\
&\qquad + \sup_{\mathbf{g}_{1:H-1}} \underbrace{\sum_{h=1}^{H}\int_{o_h}\Big(\mathbb{P}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) - \mathbb{P}_z\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right)\Big)\,\mathrm{d}o_h}_{\textbf{(vii)}}, \qquad \text{(E.10)}
\end{aligned}
$$

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

$$
\begin{aligned}
&\int_{o_h} \widehat{\mathbb{P}}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) - \mathbb{P}_{\texttt{LLM}}^t\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right)\ \mathrm{d}o_h \\
&\quad = \int_{\mathbf{o}_{2:h}} \widehat{\mathbb{P}}_{\texttt{LLM}}^t\left(\mathbf{o}_{2:h} \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) - \mathbb{P}_{\texttt{LLM}}^t\left(\mathbf{o}_{2:h} \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right)\ \mathrm{d}\mathbf{o}_{2:h} \\
&\quad = \int_{\mathbf{o}_{2:h}} \prod_{h'=1}^{h-1} \widehat{\mathbb{P}}_{\texttt{LLM}}^t\left(o_{h'+1} \,|\, (\mathbf{o}, \mathbf{g})_{1:h'}\right) - \prod_{h'=1}^{h-1} \mathbb{P}_{\texttt{LLM}}^t\left(o_{h'+1} \,|\, (\mathbf{o}, \mathbf{g})_{1:h'}\right)\ \mathrm{d}\mathbf{o}_{2:h}. \qquad \text{(E.11)}
\end{aligned}
$$

Following the arguments above, the difference can be decomposed as

$$
\begin{aligned}
&\prod_{h'=1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}\right) - \prod_{h'=1}^{h-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}\right) \\
&\quad= \sum_{h'=1}^{h-1} \Big(\widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}\right) - \mathbb{P}^t_{\texttt{LLM}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}\right)\Big) \\
&\qquad \cdot \prod_{k=h'+1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (\mathbf{o},\mathbf{g})_{1:k}\right) \cdot \prod_{k=1}^{h'-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (\mathbf{o},\mathbf{g})_{1:k}\right) \\
&\quad\le \sum_{h'=1}^{h-1} \left\|\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right) - \texttt{LLM}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)\right\|_1 \\
&\qquad \cdot \prod_{k=h'+1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (\mathbf{o},\mathbf{g})_{1:k}\right) \cdot \prod_{k=1}^{h'-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (\mathbf{o},\mathbf{g})_{1:k}\right). \qquad \text{(E.12)}
\end{aligned}
$$

Combine (E.11) and (E.12), it holds that

$$
\begin{aligned}
\textbf{(vi)} &\le \sum_{h=1}^{H}\sum_{h'=1}^{h-1} \int_{\mathbf{o}_{2:h}} \left\|\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right) - \texttt{LLM}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)\right\|_1 \\
&\qquad \cdot \prod_{k=h'+1}^{h-1} \widehat{\mathbb{P}}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (\mathbf{o},\mathbf{g})_{1:k}\right) \cdot \prod_{k=1}^{h'-1} \mathbb{P}^t_{\texttt{LLM}}\left(o_{k+1} \,|\, (\mathbf{o},\mathbf{g})_{1:k}\right) \ \mathrm{d}\mathbf{o}_{2:h} \\
&= 2\sum_{h=1}^{H}\sum_{h'=1}^{h-1} \mathbb{E}_{\mathbf{o}_{1:h'}|\mathcal{H}_t}\left[D_{\mathrm{TV}}\left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right), \texttt{LLM}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)\right)\right]. \qquad \text{(E.13)}
\end{aligned}
$$

Following (E.13), for any policy $\pi \in \Pi$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}}\left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi,\omega^t) - \mathcal{J}_{t,\texttt{LLM}}(\pi,\omega^t)\right] \\
&\quad= 2\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{h'=1}^{h-1} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\boldsymbol{\pi}}_{1:t-1}}} \mathbb{E}_{(\mathbf{o},\mathbf{g})_{1:h'}|\mathcal{H}_t}\left[D_{\mathrm{TV}}\left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right), \texttt{LLM}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)\right)\right] \\
&\quad\le 2\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{h'=1}^{h-1} \lambda_{S,1}\lambda_{S,2}^{-1} \cdot \bar{\mathbb{E}}_{(\mathbf{o},\mathbf{g})_{1:h'},\mathcal{H}_t \sim \mathcal{D}_{\texttt{LLM}}}\left[D_{\mathrm{TV}}\left(\texttt{LLM}_{\widehat{\theta}}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right), \texttt{LLM}\left(o_{h'+1} \,|\, (\mathbf{o},\mathbf{g})_{1:h'}, \mathcal{H}_t\right)\right)\right] \\
&\quad\le 2H^2T \cdot \lambda_{S,1}\lambda_{S,2}^{-1} \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) \qquad \text{(E.14)}
\end{aligned}
$$

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

$$
\begin{aligned}
\textbf{(vii)} &\le \sum_{h=1}^{H} \int_{o_h} \left\|\mathbb{P}^t_{\texttt{LLM}}\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) - \mathbb{P}_z\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right)\right\|_1 \mathrm{d}o_h \\
&= \sum_{h=1}^{H}\sum_{z'\in\mathcal{Z}} \int_{o_h} \mathbb{P}_z\left(o_h \,|\, o_1, \mathrm{do}(\mathbf{g}_{1:h-1})\right) \left\|\mathbb{P}_{\mathcal{D}}(z' \,|\, \mathcal{H}_t) - \delta_z(z')\right\|_1 \mathrm{d}o_h \le H \sum_{z'\ne z} \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathcal{H}_t).
\end{aligned}
$$

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

$$\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}_{1:t-1}}} \left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\pi, \omega^t) - \mathcal{J}_z(\pi, \omega^t) \right] \leq H \cdot \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t \sim \widehat{\mathbb{P}}_z^{\widehat{\pi}_{1:t-1}}} \left[\sum_{z' \neq z} \mathbb{P}_{\mathcal{D}}(z' \,|\, \mathcal{H}_t) \right], \tag{E.15}$$

Combine (E.14), (E.15) and the similar concentration arguments of posterior probability of (D.20), denoted by event $\mathcal{E}_2$ (see proof of Theorem 5.7 in §D.2), it holds that

$$\begin{aligned}
\textbf{(ii)} + \textbf{(iv)} &\leq \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t} \left[\mathcal{J}_z(\widehat{\pi}_z^*, \omega^t) - \widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}_z^*, \omega^t) \, \mathbb{1}\left(\mathcal{E}_2 \text{ holds}\right) \right] \qquad \text{(E.16)} \\
&\quad + \sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t} \left[\widehat{\mathcal{J}}_{t,\texttt{LLM}}(\widehat{\pi}^t, \omega^t) - \mathcal{J}_z(\widehat{\pi}^t, \omega^t) \, \mathbb{1}\left(\mathcal{E}_2 \text{ holds}\right) \right] + 2HT\delta \\
&\leq c_0 \cdot 2H^2 \log(c_{\mathcal{Z}} |\mathcal{Z}|/\delta) \cdot \left(\eta\epsilon - H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \right)^{-1} \\
&\quad + 4H^2 T \lambda_{S,1} \lambda_{S,2}^{-1} \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) + 2HT\delta. \qquad \text{(E.17)}
\end{aligned}$$

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

Combine (E.7), (E.9) and (E.17), we have

$$\begin{aligned}
\mathrm{Reg}_z(T) &\leq \textbf{(i)} + \textbf{(ii)} + \textbf{(iii)} + \textbf{(iv)} + \textbf{(v)} + 2HT \cdot \mathbb{P}(\mathcal{E}_1 \text{ fails}) \qquad \text{(E.18)} \\
&\leq c_0 \cdot 2H^2 \log(c_{\mathcal{Z}} |\mathcal{Z}|/\delta) \cdot \left(\eta\epsilon - H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \right)^{-1} + 4HT\delta \\
&\quad + 2HT\eta^{-1} \left(\eta\epsilon - H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \right) + 4\lambda_{S,1}\lambda_{S,2}^{-1} H^2 T \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) \\
&\quad + 2H^2 T (\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 + 4H^3 T \lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) \\
&\leq \mathcal{O}\Big(\underbrace{H^{\frac{3}{2}} \sqrt{\log(c_{\mathcal{Z}} |\mathcal{Z}|/\delta) \cdot T/\eta}}_{\text{Planning error}} + \underbrace{H^2 T \cdot \Delta_{\mathrm{p,wm}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi)}_{\text{Pretraining error}} \Big) + 4HT\delta, \qquad \text{(E.19)}
\end{aligned}$$

if we explore with probability $\epsilon = H(\eta\lambda_{\min})^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 + \sqrt{H \log\left(c_{\mathcal{Z}} |\mathcal{Z}|/\delta\right)/T\eta}$ and the cumulative pretraining error follows that

$$\begin{aligned}
\Delta_{\mathrm{p,wm}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta, \xi) &= 2(\eta\lambda_R)^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta)^2 \\
&\quad + 4H\lambda_R^{-1} \cdot \Delta_{\texttt{Rep}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta) + 4\lambda_{S,1}\lambda_{S,2}^{-1} \cdot \Delta_{\texttt{LLM}}(N_{\mathrm{p}}, T_{\mathrm{p}}, H, \delta).
\end{aligned}$$

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. $\square$

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 (B.2), which is akin to Proposition 4.2.

Note that for all $(h,t)\in[H]\times[T]$, it holds that

$$
\begin{aligned}
\pi^t_{h,\texttt{LLM}}\big(\mathbf{g}^t_h\,|\,\tau^t_h,\omega^t\big) &= \prod_{k\in\mathcal{K}} \texttt{LLM}\big(g^t_{h,k}\,|\,\texttt{prompt}^t_h\big)\\
&= \prod_{k\in\mathcal{K}}\left(\sum_{z\in\mathcal{Z}} \mathbb{P}\left(g^t_{h,k}\,|\,\texttt{prompt}^t_h, z\right)\cdot \mathbb{P}_{\mathcal{D}}\left(z\,|\,\texttt{prompt}^t_h\right)\right)\\
&= \prod_{k\in\mathcal{K}}\left(\sum_{z\in\mathcal{Z}} \pi^*_{z,h}\left(g^t_{h,k}\,|\,\tau^t_h,\omega^t\right)\cdot \mathbb{P}_{\mathcal{D}}\left(z\,|\,\texttt{prompt}^t_h\right)\right),
\end{aligned}
\tag{E.20}
$$

where the first equation arises from the autoregressive manner that subgoals are simultaneously generated, and the last equation follows generating distribution $\pi^*_{z,h}(\mathbf{g}_h\,|\,\tau_h,\omega)=\prod_{k\in\mathcal{K}}\pi^*_{z,h,k}(g_{h,k}\,|\,\tau_h,\omega)$. Recall that the `Planner` takes a mixture policy of $\pi_{\texttt{exp}}$ and $\pi_{\texttt{LLM}}$ such that

$$
\pi^t_h(\mathbf{g}^t_h\,|\,\tau^t_h,\omega^t)\sim(1-\epsilon)\cdot\pi^t_{h,\texttt{LLM}}(\mathbf{g}^t_h\,|\,\tau^t_h,\omega^t)+\epsilon\cdot\pi_{h,\texttt{exp}}(\mathbf{g}^t_h\,|\,\tau^t_h),
\tag{E.21}
$$

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

$$
\mathrm{Reg}(T)\le\sum_{t=1}^T\sum_{h=1}^H \mathbb{E}_{\mathcal{H}_t\sim\mathbb{P}_z^{\pi_{1:t-1}}}\mathbb{E}_{(s^t_h,\tau^t_h)\sim\mathbb{P}_z^{\pi^t}}\left[\left(\pi^*_{z,h}-\pi^t_{h,\texttt{LLM}}\right)Q^*_{z,h}(s^t_h,\tau^t_h,\omega^t)\right]+HT\epsilon,
\tag{E.22}
$$

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 $(h,t)\in[H]\times[T]$,

$$
\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}(z'\,|\,\texttt{prompt}^t_h)\le\mathcal{O}\left(\min\left\{\log\left(c_{\mathcal{Z}}|\mathcal{Z}|/\delta\right)\eta^{-1}/|\mathcal{X}^{t-1}_{\texttt{exp}}|,1\right\}\right),
\tag{E.23}
$$

where $\mathcal{X}^t_{\texttt{exp}}=\{i\in[t]:\pi^i=\pi_{\texttt{exp}}\}$ is the set of exploration episodes. Furthermore, we have

$$
\begin{aligned}
&\left(\pi^*_{z,h}-\pi^t_{h,\texttt{LLM}}\right)Q^*_{z,h}(s^t_h,\tau^t_h,\omega^t)\\
&\quad\le H\int_{\mathbf{g}^t_h}\left\|\prod_{k\in\mathcal{K}}\left(\sum_{z'\in\mathcal{Z}}\pi^*_{z',h}\left(g^t_{h,k}\,|\,\tau^t_h,\omega^t\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z'\,|\,\texttt{prompt}^t_h\right)\right)-\prod_{k\in\mathcal{K}}\pi^*_{z,h}(g^t_{h,k}\,|\,\tau^t_h,\omega^t)\right\|_1\mathrm{d}\mathbf{g}^t_h\\
&\quad\le H\int_{\mathbf{g}^t_h}\sum_{k\in\mathcal{K}}\left\|\sum_{z'\in\mathcal{Z}}\pi^*_{z',h}\left(g^t_{h,k}\,|\,\tau^t_h,\omega^t\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z'\,|\,\texttt{prompt}^t_h\right)-\pi^*_{z,h}(g^t_{h,k}\,|\,\tau^t_h,\omega^t)\right\|_1\\
&\qquad\cdot\prod_{k'=1}^{k-1}\left(\sum_{z'\in\mathcal{Z}}\pi^*_{z',h}\left(g^t_{h,k}\,|\,\tau^t_h,\omega^t\right)\cdot\mathbb{P}_{\mathcal{D}}\left(z'\,|\,\texttt{prompt}^t_h\right)\right)\cdot\prod_{k'=k+1}^{K}\pi^*_{z,h}(g^t_{h,k}\,|\,\tau^t_h,\omega^t)\,\mathrm{d}\mathbf{g}^t_h\\
&\quad= HK\cdot\sum_{z'\neq z}\mathbb{P}_{\mathcal{D}}\left(z\,|\,\texttt{prompt}^t_h\right)
\end{aligned}
\tag{E.24}
$$

for all $(h,t) \in [H] \times [T]$. Based on (E.24), Proposition 4.2 and conditioned on $\mathcal{E}_1$, it holds that

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \mathbb{P}_z^{\boldsymbol{\pi}_{1:t-1}}} \mathbb{E}_{(s_h^t, \tau_h^t) \sim \mathbb{P}_z^{\pi^t}} \left[\left(\pi_{z,h}^* - \pi_{h,\texttt{LLM}}^t\right) Q_{z,h}^*(s_h^t, \tau_h^t, \omega^t)\right] \\
&\quad \le HK \cdot \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{z' \neq z} \mathbb{E}_{\mathcal{H}_t \sim \mathbb{P}_z^{\boldsymbol{\pi}_{1:t-1}}} \mathbb{E}_{\tau_h^t \sim \mathbb{P}_z^{\pi^t}} \left[\mathbb{P}_{\mathcal{D}}(z' \,|\, \texttt{prompt}_h^t)\right] \\
&\quad \le \mathcal{O}\left(\log(c_{\mathcal{Z}}|\mathcal{Z}|/\delta) HK\eta^{-1} \cdot \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\mathcal{H}_t \sim \mathbb{P}_z^{\boldsymbol{\pi}_{1:t-1}}} \mathbb{E}_{\tau_h^t \sim \mathbb{P}_z^{\pi^t}} \left[\min\left\{1/|\mathcal{X}_{\texttt{exp}}^{t-1}|, 1\right\}\right]\right),
\end{aligned} \tag{E.25}
$$

Note that $\mathbb{1}(\pi^t = \pi_{\texttt{exp}}) \overset{\text{iid}}{\sim} \text{Bernuolli}(\epsilon)$ for all episode $t \in [T]$. Thus, Bernstein's inequality indicates that with probability at least $1-\delta$, the following event $\mathcal{E}_2$ holds:

$$
\sum_{t=1}^{T} \min\left\{1/|\mathcal{X}_{\texttt{exp}}^{t-1}|, 1\right\} \le \mathcal{O}\left(\log T \cdot \epsilon^{-1} + \sqrt{T}\log(1/\delta)\right). \tag{E.26}
$$

Combine (E.22), (E.25) and (E.26), it follows that

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

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

F Technical Lemmas

Lemma F.1 (Exponentialized Concentration)**.** Let $X_1, \dots, X_T$ be a sequence of real-valued random variables adapted to a filter $(\mathscr{F}_t)_{t \le T}$, with probability at most $1-\delta$ for all $T' \le T$ it holds

$$
\left|\sum_{t=1}^{T'} X_t - \sum_{t=1}^{T'} \log \mathbb{E}\left[\exp(X_t)|\mathscr{F}_{t-1}\right]\right| \le \log\left(1/\delta\right).
$$

Proof of Lemma F.1. See Lemma A.4 in Foster et al. (2021) for detailed proof.

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

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

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

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

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

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

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

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

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