

ChatHTN: Interleaving Approximate (LLM) and Symbolic HTN Planning

Héctor Muñoz-Avila

American University

HMUNOZAVILA@AMERICAN.EDU

David W. Aha

Navy Center for Applied Research in AI; Naval Research Laboratory; Washington, DC

DAVID.W.AHA.CIV@US.NAVY.MIL

Paola Rizzo

Interagens s.r.l.

P.RIZZO@INTERAGENS.COM

Editors: G. Pappas, P. Ravikumar, S. A. Seshia

Abstract

We introduce ChatHTN, a Hierarchical Task Network (HTN) planner that combines symbolic HTN planning techniques with queries to ChatGPT to approximate solutions in the form of task decompositions. The resulting hierarchies interleave task decompositions generated by symbolic HTN planning with those generated by ChatGPT. Despite the approximate nature of the results generated by ChatGPT, ChatHTN is provably sound; any plan it generates correctly achieves the input tasks. We demonstrate this property with an open-source implementation of our system.

Keywords: HTN planning, chatGPT, incomplete planning domains

1. Introduction

Hierarchical Task Network (HTN) planning is a paradigm in which high-level tasks are decomposed into simpler ones. HTN planning has been the subject of frequent studies because many applications are amenable to hierarchical modeling, including military planning [Donaldson \(2014\)](#), strategic decision making (e.g., in games [Smith et al. \(1998\)](#); [Verweij \(2007\)](#)), and controlling multiple agents [Cardoso and Bordini \(2017\)](#), including teams of UAVs [Musliner and Goldman \(2010\)](#). Another reason for this recurrent interest are studies reporting that hierarchical modeling is a natural way for automated systems to learn skills of increasing complexity (i.e., by starting with simpler skills and then combining them to learn more complex skills [Langley et al. \(2006\)](#)). Indeed, hierarchical planning is a key component in many cognitive architectures [Laird \(2019\)](#); [Langley and Choi \(2006\)](#) as they aim to create agents that are endowed with advanced cognitive capabilities.

LLMs cannot generate plans that are guaranteed to be *sound*. That is, they cannot guarantee that the solutions they generate correctly solve given planning problems [Valmeekam et al. \(2024b\)](#). Furthermore, they cannot verify that generated plans are indeed solutions for given problems [Valmeekam et al. \(2024a\)](#). Noticeably, they *can* generate plans for problems that have no solution (e.g., because a key resource is not given in the problem description). Nevertheless, they excel at generating *plan approximations* (i.e., plausible plans for solving problems for task domains an LLM has observed in its training data). Since they are trained with vast volumes of information, exploiting these plan approximations is an enticing possibility.

In this paper, we introduce ChatHTN, which combines two ideas: (1) HTN planning with soundness guarantees and (2) approximate hierarchical planning. In a nutshell, ChatHTN is an HTN planner: it maintains a task list, indicating the pending tasks it needs to accomplish, and uses methods

and operators to achieve these tasks. The task list is symbolic; each task is represented as a ground predicate (i.e., a predicate whose arguments are all constants). If ChatHTN attempts to decompose a task for which it has no task decomposition in its knowledge base, then it asks ChatGPT for a plausible decomposition of the task and then reverts back to HTN planning.

We have two objectives in introducing ChatHTN. First, we aim to alleviate one of the most stringent requirements of HTN planning: that a complete knowledge base is given. That is, knowledge (in the form of methods) must be provided to decompose every task in every situation where a decomposition is possible. This is an instance of the well documented knowledge engineering bottleneck that has been a recurrent, long-term thorn for AI systems. Second, we address the inability of LLMs to generate sound plans. This is also a crucial obstacle for the reliable use of LLMs, particularly in the context of mission critical, fail-safe environments such as self-driving vehicles, automated flight assistants, and power grid control.

2. Motivating Example

Consider the logistics transportation domain. In this domain, the main task is to relocate a package. There are trucks that can be used to move packages between locations provided that their locations are in the same city. There are also airplanes that can be flown between any two airports.

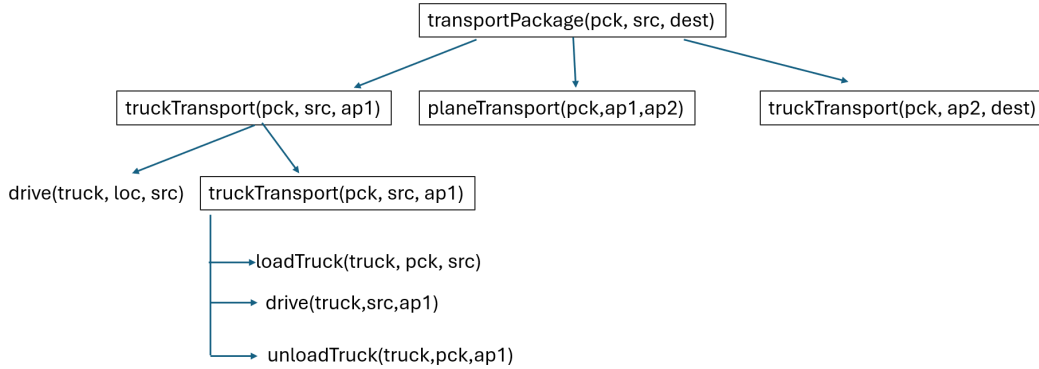


Figure 1: Initial hierarchical decomposition for a task from the logistics transportation domain. Boxed predicates are compound tasks and unboxed tasks are primitive tasks. ChatHTN stops on the planeTransport task because it lacks the knowledge to decompose it.

Consider a task $transportPackage(pck, src, dest)$ to relocate package pck from the source location to the destination location. ChatHTN proceeds with standard HTN planning and generates the hierarchy shown in Figure 1. The $transportPackage$ task is decomposed into 3 compound tasks: $truckTransport$, $planeTransport$ and $truckTransport$ (with parameters as shown in the figure). It decomposes the first $truckTransport$ task, all the way to primitive tasks. However, it stops at the $planeTransport$ task because it has no means to decompose this task given the current state and knowledge base.

In this case, a standard HTN would return \emptyset (signifying that no solutions were found). ChatHTN instead uses ChatGPT to generate a possible decomposition for this task. In doing so, ChatHTN

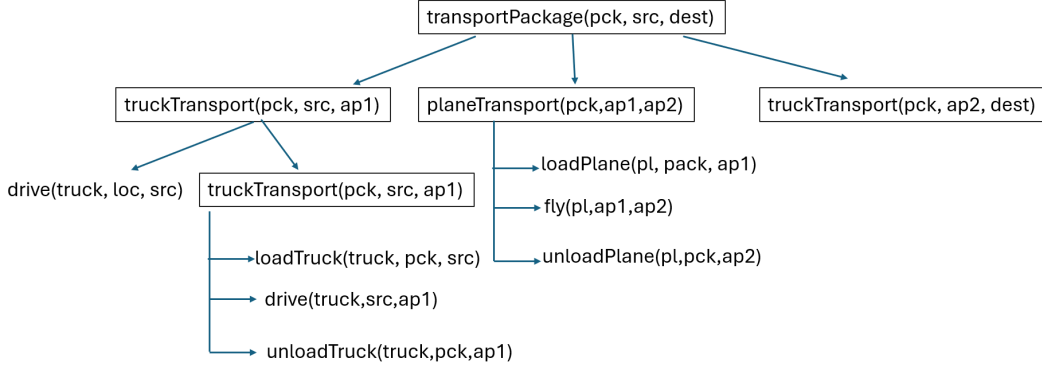


Figure 2: ChatHTN prompts ChatGPT for a decomposition of the planeTransport task.

provides ChatGPT with additional information for context such as the task’s semantics, the state, and part of the knowledge base. ChatGPT is prompted to generate a decomposition of this task into a sequence of primitive tasks. ChatGPT returns the decomposition displayed in Figure 2 for the *planeTransport* task. Then ChatHTN continues the standard HTN planning decomposition process. A complete HTN is shown in Figure 3.

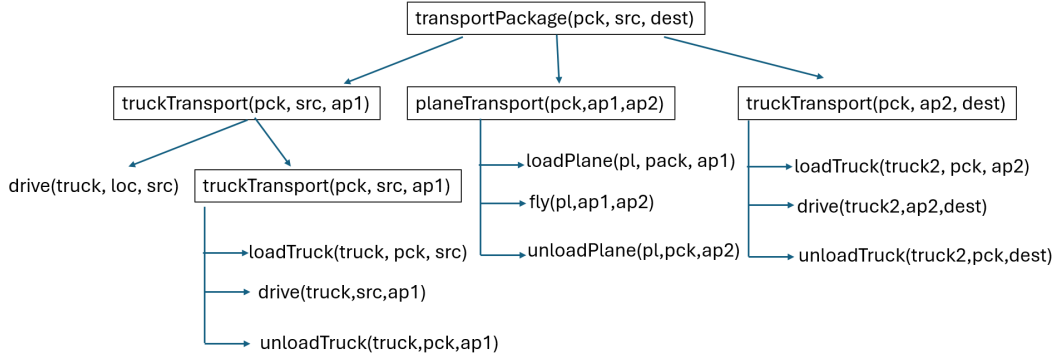


Figure 3: ChatHTN continues hierarchical planning for the remaining task to generate a complete plan.

3. Terminology

The first three rows of Table 1 describe *compound* tasks; they are decomposed into other tasks but will not change the state of the world. In the standard definition of HTN planning, compound tasks have no preconditions and effects Erol et al. (1994). However, some work on learning HTN planning knowledge does assign semantics as this is needed to guide the learning system Langley et al. (2009). Here we define tasks as *(preconditions, effects)* pairs, as in Hogg et al. (2008), which allows ChatHTN to provide ChatGPT with the semantics of the tasks to be decomposed. For instance, the

$truckTransport(p, s, d)$ task (third row) requires that package p to be at location s and the destination location d to be in the same city as location s . The expected effect of the task is for p to be at d .¹

Compound tasks express *what* needs to be accomplished but not *how*. To decompose a task, *methods* are used. They define the subtasks used to decompose a task and the preconditions under which the decomposition is valid. For instance, the $truckTransportM$ method (in the middle of Table 1) has the same preconditions as the $truckTransport$ task plus the precondition that truck t is also in location s . In that case it decomposes the task into three subtasks. The $truckTransportM$ decomposition method is used twice in Figure 3.

Table 1: Some tasks and methods in the logistics transportation task domain (method names have the "M" suffix)

Tasks/ Methods	Preconditions	Effects/Subtasks
transportPackage(p, s, d)	at(p,s)	at(p,d)
planeTransport(p, s, d)	at(p,s), airport(s), airport(d)	at(p,d)
truckTransport(p, s, d)	at(p,s), sameCity(s,d)	at(p,d)
truckTransportM(p, s, d)	at(p,s), sameCity(s,d), at(t,s)	loadTruck(t,p,s),drive(t,s,d),unloadTruck(t,p,d)
drive(t,s,d)	at(t,s), sameCity(s,d)	at(t,d)
loadTruck(t,p,l)	at(t,l), at(p,l)	at(p,t)

The other kinds of tasks are called *primitive*. They have associated actions for transforming the world state. The last two rows of Table 1 list two primitive tasks and the preconditions and effects of their associated actions. For instance, the action associated with $loadTruck(t,p,l)$ requires truck t and package p to be at location l . After the action is executed, p will be inside t .

Formally, we use T to denote the collection of all compound and primitive tasks. Compound tasks have associated (*preconditions*, *effects*) pairs whereas primitive tasks have associated actions, which have their own (*preconditions*, *effects*) pairs.

An action a defines a transition function $a : S \rightarrow S \cup \emptyset$, where S is the set of all states. If $a(s) = \emptyset$, then a is not applicable in $s \in S$. Otherwise, $a(s) \in S$ is the resulting state after taking action a in s . For instance when applying the $loadTruck(truck, pck, src)$ action in a state s where $at(pck, src)$ holds, this will yield a state s' that is identical to s except that $at(pck, src)$ no longer holds in s' and instead $at(pck, truck)$ holds.

A method m is used to decompose a compound task t . A method $m : S \times T \rightarrow \tilde{T} \cup \emptyset$ describes its task decomposition. Here, \tilde{T} is the set of all possible sequences of tasks in T . If $m(s, t) = \emptyset$, then m is not applicable in (s, t) (i.e., cannot be used to decompose task t in state s). Otherwise, $m(s, t) \in \tilde{T}$ yields a sequence of tasks (i.e., t is decomposed into $m(s, t)$). For instance, in the example in Figure 3, if applying $truckTransportM(pck, ap2, dest)$ in a state s where its preconditions hold, it will yield the task list: $load(truck2, pck, ap2)$, $drive(truck2, ap2, dest)$, $unloadTruck(truck2, pck, dest)$.

1. Table 1 displays only the main preconditions for the sake of space. Missing conditions include checking that the arguments are of the correct type, such as $isPackage(p)$, which checks that p is a valid package. We do the same for the effects.

4. ChatHTN

An HTN planning problem is defined by the 4-tuple (s, \tilde{t}, M, A) , where $s \in S$, $\tilde{t} \in \tilde{T}$, M is a collection of methods, and A is a collection of actions. Subsequently we focus on (s, \tilde{t}) as M and A remain fixed.

Figure 1 displays the ChatHTN algorithm. The lines not underlined are standard for HTN planners like SHOP Nau et al. (1999). The underlined lines (and segments) are new additions specific to ChatHTN.

Algorithm 1 The ChatHTN algorithm

```

1: procedure CHATHTN( $s, \tilde{t}$ )
2:   return chatSeekPlan( $s, \tilde{t}, ()$ )  $\triangleright ()$  is the empty plan; a plan with no actions
3: end procedure
4: procedure CHATSEEKPLAN( $s, \tilde{t}, \pi$ )  $\triangleright \pi$  is the plan generated so far
5:   if  $\tilde{t} = ()$  then return  $()$   $\triangleright$  returns the empty plan
6:   let  $\tilde{t} = (t_0, t_1, \dots, t_n)$ 
7:   if  $t_0$  is primitive then
8:     let  $a_0$  be the action associated with  $t_0$ 
9:     if  $a_0(s) \neq \emptyset$  then
10:      return CHATSEEKPLAN( $a_0(s), (t_1, \dots, t_n), \pi \cdot (a_0)$ )  $\triangleright$  continue with remaining tasks
11:    else return  $\emptyset$   $\triangleright$  return no plans
12:   if  $t_0$  is compound then
13:     for  $m_0 \in M$  do  $\triangleright M$  is the list of all methods
14:       if  $m_0(s, t_0) \neq \emptyset$  then
15:         let  $\pi' = \text{CHATSEEKPLAN}(s, m_0(s, t_0) \cdot \underline{(t_0^{ver})} \cdot (t_1, \dots, t_n), \pi) \triangleright t_0^{ver}$  is a verifier task
16:         if  $\pi' \neq \emptyset$  then
17:           return  $\pi'$ 
18:         let  $\tilde{t}_0 = \text{CHATGPTQUERY}(t_0, s, A)$ 
19:         let  $\pi' = \text{CHATSEEKPLAN}(s, \tilde{t}_0 \cdot (t_0^{ver}) \cdot (t_1, \dots, t_n), \pi)$ 
20:         if  $\pi' \neq \emptyset$  then
21:           return  $\pi'$ 
22:       return  $\emptyset$ 
23: end procedure
24: procedure CHATGPTQUERY( $t_0, s, A$ )
25:   let response = CHATGPT("...generate a decomposition ... " +  $t_0 + s + A$ )
26:   let tasks = CHATGPT("... generate a sequence of primitive tasks..." + response +  $t_0 + s + A$ )
27:   return tasks
28: end procedure

```

ChatHTN calls the chatSeekPlan procedure with the state s , the task list \tilde{t} , and a third parameter which is the plan (a list of actions) that will be generated, initially empty (Line 2). If the task list is empty, chatSeekPlan returns the empty plan (Line 5).² The rest of the algorithm is divided into three cases:

2. If chatSeekPlan returns \emptyset , it means no solution was found. If it returns $()$, then a solution was found: the empty plan.

1. (Case 1) if t_0 is primitive and its associated action a_0 is applicable in s (i.e., $a_0(s) \neq \emptyset$), then continue with the new state $a_0(s)$, the rest of the tasks (t_1, \dots, t_n) , and the plan generated so far $\pi \cdot (a_0)$ (Lines 7-10). The \cdot operator is list concatenation. Line 11 returns the empty set (i.e., no plans) if a_0 is not applicable in s .
2. (Case 2) if t_0 is compound, it searches for a method m_0 applicable to s and t_0 (i.e., $m_0(s, t_0) \neq \emptyset$) and continues with the same state s , the updated task list $m_0(s, t_0) \cdot (t_0^{ver}) \cdot (t_1, \dots, t_n)$, and the same plan π (Lines 12-15). If a plan π' is found in the recursive call, it is returned (Lines 16-17). Here is a first change with respect to standard HTN planning: the verifier task t_0^{ver} is concatenated to the task list, after $m_0(s, t_0)$. We will elaborate on verifier tasks below.
3. (Case 3) Standard HTN continues with Line 22, returning no plans since if it reaches that line, then Cases 1 and 2 have both failed. ChatHTN instead calls chatGPTQuery, which returns a sequence of primitive tasks \tilde{t}_0 for t_0 (Line 18). ChatHTN continues with a recursive call with the updated task list (Line 19). If it succeeds, then plan π' is returned (Line 21). If not, then it returns no plan (Line 22).

The procedure ChatGPTQuery is shown in Lines 24-28. It first prompts ChatGPT to generate a task decomposition for t_0 and for context it also provides s and A . Then it performs prompt chaining [Genkina \(2024\)](#) by querying ChatGPT to match the response it provided in Line 25 to a sequence of primitive tasks. In the second query we include the parameters given in the original query for context. Recurrent use of ChatGPT (and other chatbots) have shown that providing piecemeal queries provides better results than a single query with all requirements [Kwak et al. \(2024\)](#). We also observed this when testing our system. The exact prompts we use in ChatHTN can be found in Appendix A.

5. Properties and Other Considerations

ChatHTN is sound. Informally, the tasks in \tilde{t} are satisfied by the plan generated by ChatHTN(s, \tilde{t}). Formally, let Π be an HTN planner. Let $\pi = \Pi(s, \tilde{t}) = (a_0, a_1, \dots, a_m)$. We say that a task list \tilde{t} of **compound** tasks is satisfied by $\Pi(s, \tilde{t})$ if the following recursive conditions are true:³

- (base case) If $\tilde{t} = ()$ and $\Pi(s, \tilde{t}) = ()$, then \tilde{t} is satisfied by $\Pi(s, \tilde{t})$.
- (base case) If $\tilde{t} = (t_0)$, and $s' = a_m(\dots(a_1(a_0(s)))) \neq \emptyset$, then (t_0) is satisfied by $\Pi(s, (t_0))$ if the effects of t_0 are satisfied in s' .
- (recursive case) If $\tilde{t} = (t_0, t_1, \dots, t_n)$, then let $\pi_0 = (a_0 \dots a_x)$ (with $x \leq n$) be the prefix plan of π generated by $\Pi(s, (t_0))$ after recursively decomposing t_0 into primitive tasks. Then \tilde{t} is satisfied by $\Pi(s, \tilde{t})$ if:
 - (t_0) is satisfied by $\Pi(s, (t_0))$, and
 - (t_1, \dots, t_n) is satisfied by $\Pi(s', (t_1, \dots, t_n))$, where $s' = a_x(\dots(a_1(a_0(s)))) \neq \emptyset$.

3. This definition can be extended when \tilde{t} consists of both compound and primitive tasks. We focus on compound tasks because it is both the interesting case and to simplify the definition.

Under these conditions, \tilde{t} is satisfied by $ChatHTN(s, \tilde{t})$. The reason for this is the use of verifier tasks in ChatHTN. A verifier task t_0^{ver} for a compound task t_0 is a primitive task whose associated action has no effects and has as preconditions the effects of t_0 . Therefore, when applying a verifier task t_0^{ver} , the state does not change but it is inapplicable if the effects of t_0 are not valid in the state. Like any other primitive task, they are checked in Line 9 and if the preconditions are not valid, `chatSeekPlan` returns the empty set (line 11). So planning with the rest of the tasks will continue (Line 10) only if the verifier tasks are satisfied.

We added verifier tasks in the two situations when a task is decomposed in ChatHTN. First, in Line 15, when adding the subtask decomposition $m_0(s, t_0)$ for task t_0 , this is followed by adding t_0^{ver} to the task list. This forces ChatHTN to check task t_0 's effects after all tasks in $m_0(s, t_0)$ are achieved. We do the same in Line 19 for the task decomposition \tilde{t}_0 , which was generated by ChatGPTQuery. Standard HTN planning does not include verifier tasks as tasks do not have semantics (e.g., preconditions and effects). We introduced task verifiers in Hogg et al. (2008) in the context of automated learning of HTN methods. When reusing learned methods from a single plan trace, the resulting plans were always correct, satisfying the tasks' effects. However, as more methods were learned from other traces, they interacted in unexpected ways, resulting in plans that sometimes do not satisfy the given task. To our surprise, we experience an analogous phenomenon when ChatGPTQuery introduces its own task decompositions, even when most of the task decompositions were generated by our well-crafted methods. By *well-crafted*, we mean that if ChatGPTQuery is not called, then the plans will be correct even without adding the verifier tasks. This is why we introduce task verifiers for all task decompositions, including those that are generated from manually-crafted methods, a departure from our previous work on learning HTN methods Hogg et al. (2008).

An HTN planner Π is *complete* if, whenever a solvable problem (s, \tilde{t}) is given, then $\Pi(s, \tilde{t})$ returns a plan. HTN planners such as SHOP are not complete. Therefore, ChatHTN is also not complete. For instance, when solving a task, a method might be chosen that results in an infinite loop, whereas selecting a different method might yield a solution. Related to this point, we found that ChatHTN could enter infinite loops due to tasks introduced by ChatGPTQuery. To address this, our implementation of ChatHTN tracks previously visited $(state, task)$ pairs. If such a pair is encountered again, the algorithm returns an empty set (this mechanism is not explicitly shown in Algorithm 1).

6. Empirical Study

We implemented ChatHTN using the PyHop HTN planner Nau (2020) and released our implementation on GitHub⁴. We tested it on three domains that are amenable to hierarchical task decomposition: (1) logistics transportation, (2) a household robot domain, and (3) a search-and-rescue domain. All domains are also released in the GitHub repository. These studies focus on demonstrating our theoretical result regarding the soundness of ChatGPT.

The household robot domain requires a robot to navigate between rooms in a house and perform two distinct tasks: cleaning and organizing them. A key difference from logistics is that the task verifiers for some of the compound tasks, e.g., *houseIsClean(house)*, requires ChatHTN to check if a condition defined by an axiom is satisfied, which depends on multiple state conditions. For instance, "*houseIsClean(house)* is satisfied if every room in the house is clean". We also use axioms to define preconditions, such as the *sameCity* condition in Table 1.

4. <https://github.com/hhhmmmm02/ChatHTN>

The search-and-rescue domain involves a drone scanning areas to check for potential survivors after a calamity in a given region. If it finds survivors in a particular area, it needs to relocate them to a safe location. This domain also requires checking whether conditions defined by axioms are met.

For each domain we conducted the same test. We generated a prototypical problem, one that requires decomposing all possible tasks in the domain. For the logistics transportation domain, this yields a hierarchy that is similar to the one illustrated in Figure 3, with small variations such as a truck relocating before loading a package. We performed the following tests:

- **Full Domain:** We ran ChatHTN on the problem with all tasks, methods, and actions to ensure it is solvable.
- **Unsolvable:** We removed a condition from the problem, making it unsolvable. We then ran ChatHTN on the problem with all tasks, methods, and actions as a demonstration of the theoretical property that it will *not* generate a solution because ChatHTN generates only correct plans.
- **Method Removed:** For each compound task in the domain, we removed one method for decomposing the task while leaving the others available for ChatHTN. We tested whether it could still generate a plan. We repeated this test for every method of every task.
- **No Methods:** We removed all methods for a single compound task and attempt to solve the problem with ChatHTN, which has access to the other tasks and their methods. We repeated this test for each task in the domain.
- **No Model:** We removed all methods from all tasks to test whether ChatHTN could still generate a plan. This is akin to planning with LLMs, with the only difference being that we test whether the top-level task’s effects are still valid (e.g., *transportPackage* for the transportation domain).

One observation from our tests is that sometimes ChatHTN will fail to generate a solution for a given solvable problem whereas other times it generates a solution for that same problem. This occurs because ChatGPT rarely generates the same output when prompted with the same input. Therefore, for each of our tests, we gave ChatHTN five chances to solve a given problem. This happened even if we set ChatGPT’s parameter *temperature* to zero, which reduces randomness. In our experiments, we use the default temperature, which is one. This enables ChatGPT to explore alternatives while still using a coherent model.

The results of our tests are shown in Table 2. We confirmed that, when the **Full Domain** was provided, the task was solvable and no calls were made to ChatGPT. Also as expected, ChatHTN failed to find a solution for the **Unsolvable** test. For each domain, we then list the compound tasks, starting with the simplest first. The final task listed per domain is the top level task, one that when using **Full Domain** will result in a hierarchy that includes all other compound tasks.

Sometimes a solution could be generated without calling `chatGPTQuery` despite a particular method being hidden. The reason for this is that those methods cover corner cases such as when the task is already solved (e.g., the effects are satisfied in the state). Another observation from our tests is that when a problem is unsolvable or the top-level task fails then `chatGPTQuery` will still generate a sequence of tasks, even though this sequence is not a correct solution. In each such case,

Table 2: Results for the three domains. An “X” indicates that no solution was generated and a check mark denotes that a solution was generated. The compound tasks are listed in italics. The number in parentheses indicates the number of calls to CHATGPTQUERY. Thus, 0 means no calls, whereas 4 indicates four calls. If there is more than one number in parentheses next to a checkmark, then it indicates the number of runs until a solution was generated. For instance, (5,2), indicates (1) the first time it failed and there were 5 calls and (2) the second time it succeeded and there were only 2 calls.

Logistics		Household robot		Search and rescue	
Full domain	✓(0)	Full domain	✓(0)	Full domain	✓(0)
Unsolvable	X(4)	Unsolvable	X(3)	Unsolvable	X(4)
<i>truckTrans.</i>	✓(5,2)	<i>sweepTask</i>	✓(1)	<i>scanAreaTask</i>	✓(1)
M1	✓(0)	M1	✓(0)	M1	✓(0)
M2	✓(2)	M2	✓(1)	M2	✓(0)
M3	✓(5,2)	<i>org.Task</i>	✓(1)	M3	✓(1)
<i>planeTrans.</i>	✓(1)	M1	✓(1)	<i>rescueSurvivor</i>	✓(2)
M1	✓(0)	M2	✓(1)	M1	✓(2)
M2	✓(1)	<i>cleanHouse</i>	✓(1)	M2	✓(1)
M3	✓(1)	M1	✓(1)	<i>rescueSurvivor</i>	✓(2)
<i>trans.Pack.</i>	✓(1,1,1,1)	M2	✓(1)	<i>checkSurvivors</i>	✓(1)
M1	✓(0)	<i>org.House</i>	✓(2)	M1	✓(0)
M2	✓(1,1,1,1,1)	M1	✓(1)	M2	✓(1)
		M2	✓(2)	<i>search and rescue</i>	✓(1,1)
		<i>careHouse</i>	✓(1)	M1	✓(1)
		M1	✓(1)	M2	✓(0)
				M3	✓(1)

ChatHTN will catch the error because of its verifier tasks and for which we annotate an “X” in its corresponding table entry.

At the time of this writing and the current pricing for using ChatGPT’s version gpt-4-turbo for each domain, it took around \$30 per domain for tuning and running all the tests. This version of ChatGPT is considered cost-effective. Other versions are more powerful but are also much more costly. If we used a more powerful version of ChatGPT (or other chatbot), presumably more complex domains could be tested, with more complex prompt chaining designs (e.g., the hierarchy generated so far could be given as input to ChatGPT). However, we leave that as a topic for future research.

7. Related Work

There has been substantial work on learning HTNs including learning methods’ preconditions [Zhuo et al. \(2009\)](#), learning methods (i.e., both the preconditions and effects) [Langley et al. \(2009\)](#); [Hogg et al. \(2008\)](#), and learning methods and operators [Zhuo et al. \(2014\)](#). Among these prior efforts, the most closely related is the one using teleoreactive learning [Li et al. \(2011\)](#). In their work, when the algorithm is performing HTN planning and the system finds a gap in its knowledge base, it reverts to

first-principles planning to generate a plan to fill the gap. It then learns methods from the generated plan. In future work, we could do something similar by learning methods from the decompositions generated by ChatGPTQuery.

Efforts on combining planning and LLM have also gained substantial attention. Studies have shown that, while LLMs cannot generate nor verify plans, they could be used to generate approximate solutions [Valmeekam et al. \(2024a\)](#). This is a motivation for our work; to provide soundness guarantees despite using ChatGPT to approximate parts of a solution.

Work has also been conducted on generating a hierarchical planning structure from instructions given to LLMs. This has been done in combination with temporal linear logic to execute the generated hierarchies [Luo et al. \(2023\)](#). In our work, we go in the opposite direction; we generate hierarchies using HTN planning and generate localized task decompositions using LLMs while reverting back to HTN planning when possible. This is done for efficiency reasons and for reliability as the knowledge base has been carefully curated.

Work has explored possible uses of LLMs in combination with hierarchical planning [Puerta-Merino et al. \(2025\)](#). This indicates that LLMs could be used to elicit the problem’s description from the user, help generate the plan, and perform post-processing once a plan is generated. It is possible in future work that we could use LLMs to elicit the tasks that the user wants to achieve (e.g., to select among the compound tasks defined in the knowledge base).

An example of the latter point is explored in [Ding et al. \(2023\)](#). The user gives as input a plain language description of the problem. This work is in the context of underwater unmanned vehicles that need to explore ocean regions. It uses similarity functions to identify which of the tasks the user is aiming to achieve and then proceeds to perform hierarchical planning.

8. Final Remarks

In this paper we introduced ChatHTN, a sound planning algorithm to solve HTN planning problems. At its core ChatHTN is a symbolic HTN planner that performs task decomposition. However, it prompts ChatGPT with queries whenever it encounters situations where there is no applicable method to decompose a task. It uses ChatGPT’s approximate planning capabilities, when needed, to generate a plausible decomposition. It then reverts control back to symbolic HTN planning. Our work leverages sound HTN planning typical of symbolic planners while alleviating the stringent requirement that a complete knowledge base must be provided apriori that covers all cases. It ensures that the plans which are generated are correct while relaxing the requirement that a complete collection of methods must be provided.

There are several potential lines of future research. We would like to expand and test ChatGPT on more complex tasks, including some that may require a hierarchy with dozens of levels. We would also like to test ChatGPT in domains that are not from the planning literature yet are still amenable to hierarchical decomposition. One possible way to tackle these more complex scenarios is to relax the requirement that CHATGPTQUERY returns a sequence of primitive tasks. Instead, we plan to enable CHATGPTQUERY to generate a sequence of primitive and compound tasks. Finally, in some of the problems CHATGPTQUERY is queried repeatedly on the same decomposition. There is an opportunity to learn these decompositions as HTN methods, thereby reducing the calls to CHATGPTQUERY over time.

References

- Rafael Cauê Cardoso and Rafael Heitor Bordini. A multi-agent extension of a hierarchical task network planning formalism. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 2017.
- Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2086–2092. IEEE, 2023.
- M.J. Donaldson. Modeling dynamic tactical behaviors in combatxxi using hierarchical task networks. M.s. thesis, Naval Postgraduate School, 2014. Available at <https://apps.dtic.mil/sti/tr/pdf/ADA607611.pdf>.
- Kutluhan Erol, James Hendler, and Dana S Nau. Htn planning: Complexity and expressivity. In *AAAI*, volume 94, pages 1123–1128, 1994.
- Dina Genkina. Ai prompt engineering is dead. long live ai prompt engineering. *IEEE Spectrum*, 2024.
- Chad Hogg, Héctor Munoz-Avila, and Ugur Kuter. Htn-maker: Learning htms with minimal additional knowledge engineering required. In *AAAI*, pages 950–956, 2008.
- Alice Kwak, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. Classify first, and then extract: Prompt chaining technique for information extraction. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 303–317, 2024.
- John E Laird. *The Soar cognitive architecture*. MIT press, 2019.
- Pat Langley and Dongkyu Choi. A unified cognitive architecture for physical agents. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1469. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- Pat Langley, Dongkyu Choi, Roland Olsson, and Ute Schmid. Learning recursive control programs from problem solving. *Journal of Machine Learning Research*, 7(3), 2006.
- Pat Langley, Nan Li, Negin Nejati, and David J Stracuzzi. Learning hierarchical skills from problem solutions using means-ends analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31, 2009.
- Nan Li, David J Stracuzzi, and Pat Langley. Improving acquisition of teleoreactive logic programs through representation change. In *AAAI Fall Symposium: Advances in Cognitive Systems*, 2011.
- Xusheng Luo, Shaojun Xu, and Changliu Liu. Obtaining hierarchy from human instructions: an llms-based approach. In *CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP)*, 2023.
- David J Musliner and Robert P Goldman. Priority-based meta-control within hierarchical task network planning. In *2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, pages 282–286. IEEE, 2010.

- Dana Nau. A simple htn planning system written in python, 2020. URL <https://bitbucket.org/dananau/pyhop/src/master/>.
- Dana Nau, Yue Cao, Amnon Lotem, and Hector Munoz-Avila. Shop: Simple hierarchical ordered planner. In *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*, pages 968–973, 1999.
- Israel Puerta-Merino, Carlos Núñez-Molina, Pablo Mesejo, and Juan Fernández-Olivares. A roadmap to guide the integration of llms in hierarchical planning. *arXiv preprint arXiv:2501.08068*, 2025.
- Stephen JJ Smith, Dana S Nau, Thomas A Throop, et al. Success in spades: Using ai planning techniques to win the world championship of computer bridge. *AAAI/IAAI*, 98:1079–1086, 1998.
- Karthik Valmeekam, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. Planning in strawberry fields: Evaluating and improving the planning and scheduling capabilities of lrm o1, 2024a. URL <https://arxiv.org/abs/2410.02162>.
- Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. Llms still can’t plan; can lrms? a preliminary evaluation of openai’s o1 on planbench. *arXiv preprint arXiv:2409.13373*, 2024b.
- Tim Verweij. A hierarchically-layered multiplayer bot system for a first-person shooter. Ms thesis, Vrije Universiteit of Amsterdam, 2007. Available at https://www.guerrilla-games.com/media/News/Files/VUA07_Verweij_Hierarchically-Layered-MP-Bot_System.pdf.
- Hankz Hankui Zhuo, Derek Hao Hu, Chad Hogg, Qiang Yang, and Hector Munoz-Avila. Learning htn method preconditions and action models from partial observations. In *IJCAI*, pages 1804–1810, 2009.
- Hankz Hankui Zhuo, Héctor Muñoz-Avila, and Qiang Yang. Learning hierarchical task network domains from partially observed plan traces. *Artificial intelligence*, 212:134–157, 2014.

Appendix A: ChatGPT prompts

In this appendix, we show the exact prompts used by ChatGPT. These prompts are parametrized so that when the domain changes, we pass the domain-specific information such as actions, tasks, etc., while the rest of the prompt remains the same.

The following is the prompt used for ChatGPT in Line 25 of the Algorithm 1. The “+” operator below is text concatenation. Here we are passing text versions (automatically generated by processing the PyHop’s knowledge base) of the task, the task’s preconditions (precs), the task’s effects (effs), state, actions (operatorsText), and axioms. The latter are auxiliary functions to simplify the preconditions. For instance, *samecity(loc, loc2)* is an auxiliary function used by the *drive* action to check that the starting and ending locations are in the same city.

```
conversation=[
{"role": "system", "content": "You are an AI planner specializing in HTN planning."},
{"role": "user", "content": "The domain is defined by the following operators (each defined as
```

a Python function):" + operatorsText + ". Some of the preconditions in the operators are defined by the following python functions: " + axiomsText + ". Provide the Sub-Tasks Breakdown for the following task: " + task + ". Here are the preconditions of the task: " + precs + ". Here are the effects of the task: " + effs + ". Here is the current state: " + stateText + " Provide a complete and logically valid decomposition using the operators and functions provided." + " Do not invent new operators. Your output should be a step-by-step list of sub-tasks in logical order," + " using arguments ground in the current state."}]

The following is the second prompt after the response to the previous prompt has been generated (Line 26 of Algorithm 1). The variable *response* is the response by ChatGPT to the previous prompt.

```
conversation=[
{ "role": "system", "content": "You are an AI planner specializing in HTN planning."},
{ "role": "user", "content": ". You generated the following response:" + chatGPTtext + "to my request to provide the Sub-Tasks Breakdown for the following task: " + task + ". I also gave you the preconditions of the task: " + precs + " and the effects of the task: " + effs + ".and gave you the state: " + state + "and gave you the domain defined by the following operators (each defined as a Python function):" + operatorsText + ". I also gave you the following python functions which are called to check some preconditions and to check some effects: " + axiomsText + ". As a follow-up, can you map the subtasks you generated with the operators I provided, please? please list the operator names as predicates," + " for the match you generate use only the predicate names of the operators and the arguments in your sub-task breakdown" + " Always respond with a compact, machine-readable format using predicate form. " + " Avoid explanations or extra text unless explicitly requested. " + " When generating your output, list only the predicates in the form: " + " predicate(arg1, arg2, ...)" + " where predicate is name of an operator I provided" + " Separate predicates by newlines. Do not include explanations, headings, or descriptions." + " Use only the operator names I provided. Ensure that every predicate corresponds exactly" + " to one of those operators and that all arguments match those in your sub-task breakdown"} ]
```