

# AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers

Yongchao Chen<sup>1,2</sup>, Jacob Arkin<sup>1</sup>, Charles Dawson<sup>1</sup>, Yang Zhang<sup>3</sup>, Nicholas Roy<sup>1</sup>, and Chuchu Fan<sup>1</sup>

**Abstract**—For effective human-robot interaction, robots need to understand, plan, and execute complex, long-horizon tasks described by natural language. Recent advances in large language models (LLMs) have shown promise for translating natural language into robot action sequences for complex tasks. However, existing approaches either translate the natural language directly into robot trajectories or factor the inference process by decomposing language into task sub-goals and relying on a motion planner to execute each sub-goal. When complex environmental and temporal constraints are involved, inference over planning tasks must be performed jointly with motion plans using traditional task-and-motion planning (TAMP) algorithms, making factorization into subgoals untenable. Rather than using LLMs to directly plan task sub-goals, we instead perform few-shot translation from natural language task descriptions to an intermediate task representation that can then be consumed by a TAMP algorithm to jointly solve the task and motion plan. To improve translation, we automatically detect and correct both syntactic and semantic errors via autoregressive re-prompting, resulting in significant improvements in task completion. We show that our approach outperforms several methods using LLMs as planners in complex task domains. See our project website<sup>§</sup> for prompts, videos, and code.

## I. INTRODUCTION

Providing agents with the ability to find and execute optimal plans for complex tasks is a long-standing goal in robotics. Robots need to not only reason about the task in the environment and find a satisfying sequence of actions but also verify the feasibility of executing those actions given the robot’s motion capabilities. This problem is referred to as task and motion planning (TAMP), and there has been considerable research on efficient algorithms [1]. Classic solutions rely on specifying tasks in a dedicated planning representation, such as PDDL [2] or Temporal logics [3], that is both sufficiently expressive to specify task complexities (e.g. constraints on task execution) and amenable to such algorithms [2], [3], [4], [5].

While this approach to task specification has been quite successful, directly using these representations requires training and experience, making them poor interfaces for non-expert users. As an alternative, natural language (NL) provides an intuitive and flexible way to describe tasks. Pre-trained large language models (LLMs) have demonstrated surprisingly good performance on many language-related tasks [6], and there has been an associated burst of research

applying them to task execution [7], task planning [8], [9], [10], [11] and TAMP [12], [13].

Promising early efforts used LLMs as direct task planners [8] generating a sequence of sub-tasks based on a set of natural language instructions, but these approaches were limited by a lack of feedback and inability to verify whether sub-task sequences are executable. Further research addressed executability by connecting sub-tasks to control policy affordance functions [9], providing environmental feedback of robot actions [11], and interleaving action feasibility checking with LLM action proposals [12]; this last work also addressed long-horizon action dependencies. However, these approaches struggle with complex tasks involving temporally-dependent multi-step actions, action sequence optimization [9], [11], and constraints on task execution [12]. Furthermore, these frameworks factor the planning problem and use LLMs to infer a task plan separately from the motion plan. In many situations, the task and motion plan must be optimized together to fulfill the task. For instance, when the task is ‘reach all locations via the shortest path’, the order of places to be visited (task planning) depends on the geometry of the environment and the related motion optimization. Unfortunately, we find that LLMs do not seem capable of directly generating trajectories, possibly due to limitations in complex spatial and numerical reasoning [14], [15].

To benefit from both the user-friendliness of NL and the capabilities of existing TAMP algorithms, we approach the problem by using LLMs to translate from high-level task descriptions to formal task specifications. We are not the first to use LLMs in this way [16], [17], but our work addresses some limitations of prior approaches. Previous work translated natural language to Linear Temporal Logics (LTL) [18], which only considered the problem of task planning, and PDDL problem descriptions [16] or PDDL goals [17]. Here we utilize Signal Temporal Logic (STL) as the intermediary representation, allowing for more expressive constraints than LTL and facilitating integrated task and motion planning as with PDDL [19].

The LLM translation process can produce malformed (*syntax errors*) and semantically misaligned (*semantic errors*) formal task specifications. To address syntax errors, we adopt an existing iterative re-prompting technique that relies on an external syntax verifier to prompt the LLM with the specific syntactic error for correction [20]. Unfortunately, the lack of an external verifier makes this technique inapplicable for a semantic misalignment between the original natural language instruction and the translated specification. To address this problem, we contribute a novel autoregressive re-prompting

<sup>1</sup>Massachusetts Institute of Technology. jarkin@mit.edu, cbd@mit.edu, nickroy@csail.mit.edu, chuchu@mit.edu

<sup>2</sup>Harvard University. yongchaochen@fas.harvard.edu

<sup>3</sup>MIT-IBM Watson AI Lab. yang.zhang2@ibm.com

<sup>§</sup><https://yongchao98.github.io/MIT-REALM-AutoTAMP/>

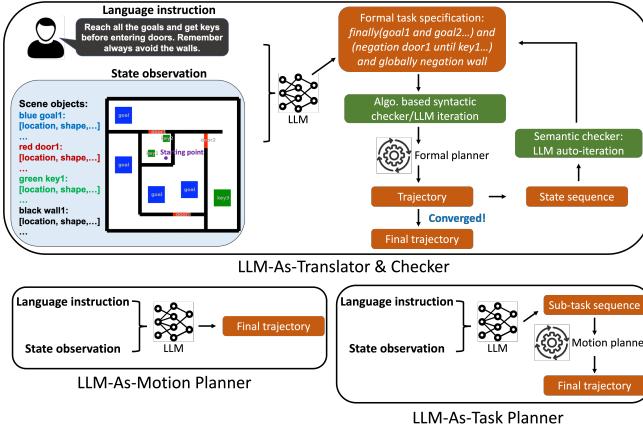


Fig. 1. Illustration of different approaches applying LLMs for task and motion planning; our work contributes the LLM-As-Translator & Checker approach. Each approach accepts a natural language instruction and environment state as input and outputs a robot trajectory.

technique that uses an LLM to evaluate whether the generated plan is semantically consistent with the original instruction. We re-prompt the model to check the alignment between the original instruction and the generated plan by providing the context of the instruction, the generated STL, and the output of the planner. We conduct comprehensive experiments in challenging 2D task domains, including several multi-agent tasks, and find that our approach outperforms direct LLM planning for tasks with hard geometric and temporal constraints. We show that, when combined with automatic syntactic correction, our technique significantly improves task success rates. We conduct an ablation study over the translation step by integrating a fine-tuned NL-to-STL model [21] with the AutoTAMP framework and show that GPT-4 few-shot learning is competitive with fine-tuning. In addition to our code, we publish a dataset of 1400 test cases consisting of the language instructions, environments, generated STL, and planner trajectory outputs. We conclude that in-context learning with pre-trained LLMs is well suited for language-to-task-specification translation for solving TAMP problems.

## II. PROBLEM DESCRIPTION

As shown in Figure 1, we aim to convert a natural language instruction, including spatial and temporal constraints, into a motion plan for a robot encoded as a set of timed waypoints, e.g.,  $(x_i, y_i, t_i)$ . The environment state is encoded as set of named obstacles described as polygons and is provided as additional context. Our task is to generate a constraint-satisfying trajectory based on the given instruction and the environment state. The robot must not surpass its maximum velocity, and the total operation time should not exceed the task time limit. We assume that the full trajectory is a linear interpolation between the timed waypoints; complex trajectories can be specified by dense waypoint sequences.

## III. METHODS

Figure 1 illustrates three of the approaches we compare in our work, each using LLMs in some capacity. Each takes as

input (1) a text-based representation of the global environment state, (2) in-context examples for few-shot learning, and (3) a natural language instruction. The LLM-As-Translator & Checker approach is the contribution of this paper. Details and examples of context for prompting and re-prompting can be found in our code repository<sup>8</sup>.

### A. LLM End-to-end Motion Planning

One natural idea is to use an LLM for both task and motion planning by directly generating a trajectory for a given language instruction; we refer to this as LLM End-to-end Motion Planning. In cases where the generated trajectory violates constraints, we re-prompt the model with the constraint violation to produce another trajectory, allowing up to five such re-prompts. Figure 2 shows this pipeline, including a specific failure case with two constraint-violating trajectories. The LLM End-to-end Motion Planning violates the constraints multiple times even with direct correction re-prompts. This results from the poor spatial and numerical reasoning abilities of LLMs.

### B. LLM Task Planning

A more common approach is to use an LLM to handle the task planning by directly generating a sequence of sub-tasks from a given language instruction; we refer to this as LLM Task Planning. To generate a final trajectory, the sub-tasks are handled by an independent motion planner. In this work, these sub-tasks are limited to navigation actions, and the motion planning is handled by the STL planner used by our proposed approach; this permits fair comparison of results across methods. Each sub-task is converted to STL to be consumed by the planner. We evaluate and compare against three methods that each use LLMs for task planning: (1) Naive Task Planning, (2) SayCan, and (3) LLM Task Planning + Feedback.

**Naive Task Planning** As proposed by [8], we evaluate using LLMs to generate the entire sub-task sequence without checking for executability.

**SayCan** Alternatively, an LLM can be iteratively prompted to generate each subsequent sub-task conditioned on the previous sub-tasks in the sequence. The next sub-task can be selected from the top K candidates by combining the language model likelihood with a feasibility likelihood of the candidate action and choosing the most-likely next sub-task, as proposed by [9]. We set K to 5 in our evaluations.

**LLM Task Planning + Feedback** A third task planning method combines full sequence generation with feasibility checking to both find sub-task sequences that satisfy the full task and verify their feasibility before execution. For any infeasible sub-tasks, the LLM can be re-prompted with feedback about the infeasible actions to generate a new sub-task sequence. This is similar to the hierarchical method proposed by [12] but with feedback for re-prompting.

### C. Autoregressive LLM Specification Translation&Checking + Formal Planner

Unlike LLM Task Planning, our approach translates NL to STL with an LLM and then plans the trajectory with an

```

User prompt1:
Task explanation: 'Hope you can help me plan
the trajectory to fulfill the instruction ...' +
Few-shot examples +
Environment settings: '[name: room1, color:
red, position and size: [0, 0.9, -1, -0.5],
function: kitchen] ...' +
Instruction: 'At some point go to the yellow
box, and at some point go to the red box, and
then enter the green box, and always do not
enter the blue area.'

GPT-4 response1:
[(-1.3, -1.3, 0), (-0.95, -0.2, 0.5), (0.45,
0.95, 0.9), (0.4, 0.95, 1.5)...]

User prompt2:
Your trajectory between subpoint3 and subpoint4
enters the blue box, you should avoid it.

GPT-4 response2:
[(-1.3, -1.3, 0), (-0.95, -0.2, 0.5), (0.45,
0.95, 0.9), (-0.35, 0.55, 1.45)...]

```

⋮

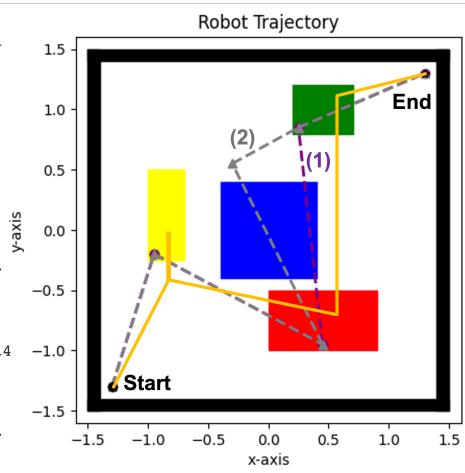


Fig. 2. GPT-4 failure case for direct end-to-end trajectory planning. The orange line shows the correct path obeying the instruction. The purple and gray dashed lines show the trajectories from GPT-4 after first and second prompts, respectively. GPT-4 generates a list of  $(x, y)$  locations with associated timestamps. The initial prompt describes the language modeling task, environment state, and instruction. Each object is a rectangle described by  $(x, y)$  boundaries.

STL planner, as shown in Figure 1. We include two re-prompting techniques to improve translation performance: one for syntactic errors and another for semantic errors. By “semantic error”, we mean a misalignment between the intended task described in natural language and the STL expression to which it is translated. Figure 3 shows the structure of the context for re-prompting the model for semantic error correction; we include a full prompt example in our code repository<sup>§</sup>.

**Signal Temporal Logic Syntax** In this work, we use STL [22] as a formal task specification that supports continuous real-time constraints suitable for time-critical missions. An STL formula is defined recursively according to the following syntax:

$$\phi ::= \pi^\mu \mid \neg\phi \mid \phi \wedge \varphi \mid \phi \vee \varphi \mid \mathbf{F}_{[a,b]} \phi \mid \mathbf{G}_{[a,b]} \phi \mid \phi \mathbf{U}_{[a,b]} \varphi \quad (1)$$

where  $\phi$  and  $\varphi$  are STL formulas, and  $\pi^\mu$  is an atomic predicate.  $\neg$  (negation),  $\wedge$  (and),  $\vee$  (or),  $\Rightarrow$  (imply), and  $\Leftrightarrow$  (equal) are logical operators.  $\mathbf{F}_{[a,b]}$  (eventually/finally),  $\mathbf{G}_{[a,b]}$  (always/globally), and  $\mathbf{U}_{[a,b]}$  (until) are temporal operators with real-time constraints  $t \in [a, b]$ . The action primitives in this work are ‘enter(room.name)’ and ‘not\_enter(room.name)’.

**STL Trajectory Planner** We use a state-of-the-art multi-agent STL planner [23] that uses piece-wise linear reference paths defined by timed waypoints to recursively encode the constraints expressed in the provided STL expression. It defines the validity of an STL formula with respect to a trajectory and then optimizes the trajectory to maximize the validity. The planner not only searches for a sub-task sequence but also optimizes the time efficiency under dynamical constraints of robot maximum velocity. Here we assume that the locations and shapes of all the

objects/rooms in the whole environment are known, which serves as the environment information to the STL planner.

**Syntactic Checking & Semantic Checking** Open-loop translation can suffer from syntactic and semantic errors. We use two re-prompting techniques to automatically correct such errors. Like [20], we use a verifier to check for syntax errors (we use a simple rules-based STL syntax checker); any errors are provided as feedback when re-prompting the LLM to generate corrected STL. We repeat until no errors are found (up to five iterations). For semantic errors, we propose a novel autoregressive re-prompting technique; we provide the STL planner’s generated state sequence (i.e., [[in(road), 0], [in(red kitchen), 0.5], [in(blue restroom2), 1.2], ...]) as context alongside the original instruction and ask the LLM to check whether the plan aligns with the instruction’s semantics. If it does not, the LLM is prompted to modify the STL, which repeats the syntactic and semantic re-prompting. This process terminates in the case of no detected errors or no change in STL (up to three iterations). The structure of the semantic error prompt is shown in Figure 3; full example prompts can be found in our code repository<sup>§</sup>.

#### IV. EXPERIMENTAL DESIGN

Each task scenario is set in a 2D environment and entails navigation of one or more robots; the robots have extent in the environment and are initialized with varying start positions. Each environment consists of regions with shapes, locations, and properties (e.g., color, name, function). For each method, the LLM is initially prompted with a description of the language task (e.g. task planning or translation) and five in-context examples for that task. To mitigate variance across prompts, we initially tested six different sets of examples for each method and chose the one that performed best. Through this testing, we found that the variance over prompts was insignificant relative to overall performance.

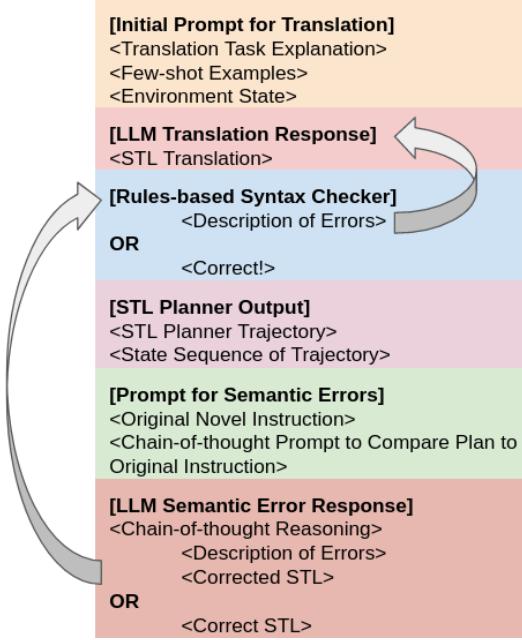


Fig. 3. High-level structure of the prompt used for AutoTAMP. The arrow on the right indicates re-prompting for syntax error correction. The arrow on the left indicates re-prompting in cases of semantic errors.

We evaluated the different methods described in Section III across six different task scenarios (three single-agent and three multi-agent) with different combinations of geometric and temporal constraints. For each scenario description below, we indicate the presence of these constraints below with G and T, respectively. For each method, we evaluate performance with both GPT-3 and GPT-4 as the LLM. Note that in multi-agent scenarios, we do not test SayCan or LLM Task Planning + Feedback because these methods are not straight-forwardly adaptable for multiple agents. For multi-agent tasks, the agents are assigned a subtask and a time for completion at each time step; since the time for completion is often different, it is not obvious how/when to check and provide feedback. We also terminate and report failure for test cases that take more than 90 minutes. We automatically check resulting trajectories via hard-coded checkers. The full set of experiments took two weeks using four 16-core CPUs; the cost of LLM API calls for evaluating all of the approaches was  $\sim 1500$  USD.

**HouseWorld1 (single-agent)** As shown in Figure 4(a), this is a house environment from [24]. We first manually constructed 10 different instructions of varying complexity before prompting GPT-4 to paraphrase each into 9 differently worded instructions with the same meaning, resulting in 100 total instructions for this environment. For each instruction, we randomly initialize between two start-end position pairs for 200 total test cases. For this scenario, we do not impose a hard time constraint for the planned trajectory.

**HouseWorld2 (T, single-agent)** This scenario is identical to HouseWorld1, but each planned trajectory is subjected to a hard time constraint. This time limit is pre-determined by

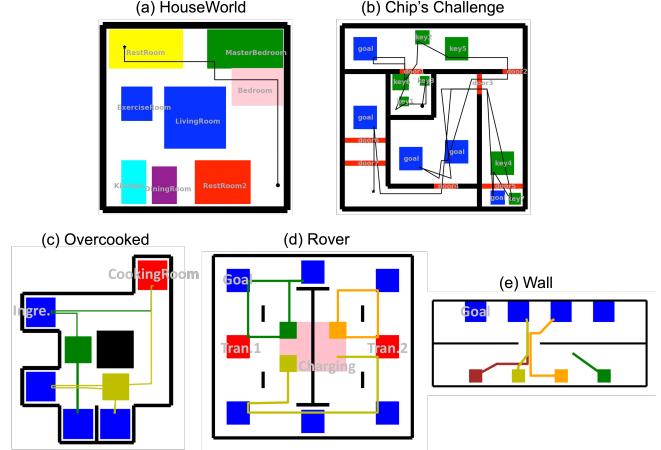


Fig. 4. HouseWorld and Chip’s Challenge are single-agent scenarios. Overcooked, Rover, and Wall are multi-agent scenarios. The black square in Overcooked is inadmissible. The lines indicate the correct trajectories following the instructions. For the HouseWorld and Chip’s Challenge environments, the black round dot and pentagonal dot indicate the start and end positions, respectively.

completing the correct trajectory with 0.8 maximum velocity.

The remaining task scenarios were designed with specific rules and goals for the agent(s) to follow. For each scenario, GPT-4 was used to paraphrase the original description into 20 unique variants with the same meaning, which are further checked by humans. We instantiate three different instances of the environment for each scenario and randomize five different start/end location pairs for a total of 300 test cases.

**Chip’s Challenge (G, single-agent)** Figure 4(b) shows a scenario inspired by Chip’s Challenge, a classic puzzle solving game with strict geometric and logical constraints. The robot must reach all goal regions (blue) but must acquire a unique key to pass through the corresponding door.

**Overcooked (G & T, multi-agent)** Figure 4(c) shows a scenario inspired by Overcooked, a popular cooking simulation game with strict time constraints. The agents must cooperatively gather ingredients and return to CookingRoom in a limited time. The multi-agent motion planning is challenged by limited space for agents to maneuver.

**Rover (G & T, multi-agent)** Figure 4(d) is a scenario used by [23]. Multiple agents must reach each observation region (blue) before transmitting their observations from a red region, all while subjected to time and energy constraints.

**Wall (G & T, multi-agent)** Figure 4(e) is also from [23]. Multiple agents must occupy each goal region (blue) while subject to a time constraint and a maneuver bottleneck.

## V. RESULTS

We report the task success rates for the single-agent and multi-agent scenarios in Table I and Table II, respectively. For HouseWorld1 (Figure 4(a)) with no hard time constraint, we find that all methods using LLMs as task planners outperform our approach; whereas our approach can fail due to translation errors, this environment permits direct trajectories between any two positions and thus lacks geometric challenges that direct task planning methods will struggle with.

TABLE I  
TASK SUCCESS RATES FOR SINGLE-AGENT SCENARIOS. EACH SCENARIO'S CONSTRAINTS ARE LISTED IN THE TABLE.

|       |                                | <b>HouseWorld1</b>            | <b>HouseWorld2</b>   | <b>Chip's Challenge</b>    |
|-------|--------------------------------|-------------------------------|----------------------|----------------------------|
|       |                                | Soft Time Constraint          | Hard Time Constraint | Hard Geometric Constraints |
| GPT-3 | <b>LLMs as Motion Planners</b> | End-to-end Motion Planning    | 0.0%                 | 0.0%                       |
|       |                                | Task Planning (naive)         | 74.0%                | 36.0%                      |
|       | <b>LLMs as Task Planners</b>   | SayCan                        | 75.5%                | 36.0%                      |
|       |                                | Task Planning (feedback)      | <b>79.0%</b>         | 40.0%                      |
|       | <b>LLMs as Translators</b>     | No Corrections                | 28.0%                | 27.0%                      |
|       |                                | Syntax                        | 49.0%                | 47.0%                      |
|       |                                | Syntax + Semantics (AutoTAMP) | 62.0%                | <b>62.0%</b>               |
| GPT-4 | <b>LLMs as Motion Planners</b> | End-to-end Motion Planning    | 9.5%                 | 9.5%                       |
|       |                                | Task Planning (naive)         | 90.0%                | 45.0%                      |
|       | <b>LLMs as Task Planners</b>   | Saycan                        | 90.0%                | 47.5%                      |
|       |                                | Task Planning (feedback)      | <b>92.0%</b>         | 49.0%                      |
|       | <b>LLMs as Translators</b>     | No Corrections                | 43.5%                | 42.0%                      |
|       |                                | Syntax                        | 59.5%                | 59.0%                      |
|       |                                | Syntax + Semantics (AutoTAMP) | 82.5%                | 82.0%                      |
|       |                                | NL2TL + Syntax + Semantics    | -                    | <b>83.5%</b>               |
|       |                                |                               |                      | 86.0%                      |

TABLE II  
TASK SUCCESS RATES FOR MULTI-AGENT SCENARIOS. EACH SITUATION HAS HARD CONSTRAINTS ON TIME AND GEOMETRY.

|       |                                | <b>Overcooked</b>                        | <b>Rover</b>  | <b>Wall</b>   |
|-------|--------------------------------|--|---------------|---------------|
|       |                                | Hard Time & Geometric Constraints        |               |               |
| GPT-3 | <b>LLMs as Motion Planners</b> | End-to-end Motion Planning               | 0.0%          | 0.0%          |
|       | <b>LLMs as Task Planners</b>   | Task Planning (naive)                    | 13.3%         | 0.0%          |
|       | <b>LLMs as Translators</b>     | No Corrections                           | 25.0%         | 22.0%         |
|       |                                | Syntax Corrections                       | 70.0%         | 35.0%         |
|       |                                | Syntax + Semantic Corrections (AutoTAMP) | <b>89.0%</b>  | <b>60.7%</b>  |
| GPT-4 | <b>LLMs as Motion Planners</b> | End-to-end Motion Planning               | 5.0%          | 0.0%          |
|       | <b>LLMs as Task Planners</b>   | Task Planning (naive)                    | 17.0%         | 0.0%          |
|       | <b>LLMs as Translators</b>     | No Corrections                           | 85.0%         | 46.0%         |
|       |                                | Syntax Corrections                       | 94.0%         | 67.0%         |
|       |                                | Syntax + Semantic Corrections (AutoTAMP) | <b>100.0%</b> | 79.0%         |
|       |                                | NL2TL + Syntax + Semantic Corrections    | <b>100.0%</b> | <b>79.7%</b>  |
|       |                                |  |               | <b>100.0%</b> |

When adding a strict time constraint (HouseWorld2), we see that such methods perform much worse while AutoTAMP's success rate persists. For the other tasks that include geometric constraints, LLM End-to-end Motion Planning and Naive Task Planning both perform quite poorly. Unsurprisingly, we observe a general trend that GPT-4 outperforms GPT-3.

We find that most failures for LLM Task Planning methods result from task execution time violation and sequencing of actions for long-horizon tasks. For example, Chip's Challenge requires the robot to efficiently collect keys for future doors. Also, the Naive Task Planning method fails to avoid collisions in the multi-agent scenarios. Failures for methods that translate to STL primarily are due to incorrect translation; while our re-prompting techniques help address this issue, there remain cases of poor translation.

**Ablation Studies** In Table I and Table II, we evaluate the impact of syntactic and semantic error correction on using LLMs to translate to STL. The results show that translation with no error correction has modest success across task scenarios, but both syntactic and semantic error correction significantly improve performance; this trend is present across all scenarios. We also evaluate replacing a pre-trained LLM for translation with a state-of-the-art modular translation pipeline, NL2TL, that uses a smaller LLM (T5-large) fine-tuned on a multi-domain corpus of 30K examples

of instructions paired with their corresponding temporal logic expressions [21]; the error correction steps were still performed by GPT-4. Integrating NL2TL performs similarly to using a pre-trained LLM for translation, providing a modest improvement in HouseWorld2 and Rover. We note that incorporating the two re-prompting techniques for error correction is competitive with fine-tuning since we do not rely on additional data or training.

**3D Simulation** In supplemental videos, we demonstrate plans generated via AutoTAMP in two 3D simulated environments: a drone navigation scenario that requires reasoning about height, and a tabletop color sorting manipulation scenario. We did not incorporate the semantic check for these demos. The STL planner is directly applicable to the drone scenario using timed waypoints, as done in the 2D experiments. For manipulation tasks, we integrated a simple discrete planner to handle the dynamics mode transitions. We discuss this more in Section VII.

**Physical Demonstrations** We demonstrate AutoTAMP on physical differential-drive robots via the remotely-accessible Robotarium platform [25] for the Overcooked, Rover, Wall, and Chip's Challenge scenarios. We track the planned trajectories using a low-level controller that also includes a control barrier function to prevent collisions between robots. This controller and the underlying nonlinear dynamics in-

duce a tracking error; we account for this by padding obstacles at planning time. Obstacles are displayed in the robot workspace using an overhead projector. These physical demos provide evidence that our method can be applied to real-world navigation task and motion planning. They are included as part of supplemental videos.

## VI. RELATED WORK

**Task and Motion Planning** Planning for robotics involves both high-level, discrete planning of tasks [5] and low-level continuous planning of motions [26]; solving these simultaneously is referred to as task and motion planning [1]. Modern approaches either attempt to satisfy the motion constraints prior to action sequencing [27], [28], [29], find action sequences then satisfy the motion constraints [30], [31], [32], [19], or interleave these steps [33], [34], [35]. For tasks specified in temporal logic, existing methods either use multi-layer planning [36], like the aforementioned approaches, or direct optimization via a mixed-integer linear program [37], [23] or a non-linear program [38]. Our work focuses on translating natural language to STL, relying on [23] as a TAMP solver, but can be integrated with other STL-based planners.

**LLMs for TAMP** Recent claims about the impressive reasoning capabilities of LLMs [6], [39] have led to interest in such models for task and motion planning. One approach is to directly use LLMs as planners [8], [9], [12], [11], [7], [10], [13]. Initial work showed that zero-shot generation of an action sequence from a high-level task description had relatively poor executability, but few-shot in-context learning, constraining output to admissible actions, and iterative action generation significantly improved performance [8]. Subsequent efforts grounded the primitive actions to motion control policies, using affordance functions to guide LLM-based task planning [9] and TAMP [12], also adding feedback [11]. Other work focused on how prompting can inform task execution [7], [13]. Despite these successes, however, there is evidence that LLMs perform poorly on more realistic tasks [15], [40], motivating different approaches. While we are interested in LLMs for TAMP, our work does not directly use LLMs as planners.

**Translating Language to Task Representations** A natural alternative is to rely on dedicated planners by mapping from natural language to a planning representation. There is a rich history of parsing natural language into formal semantic representations [41], [42], [43], of which we only provide a relatively small sampling. The robotics community adopted parsing and other techniques to map language to such representations as lambda calculus [44], [45], motion planning constraints [46], linear temporal logic [47], [48], [49], [50], and signal temporal logic [51], [52], among others [53]. We refer readers to [54] for a more thorough review. To address challenges of data availability, task generalization, linguistic complexity, common sense reasoning, and more, recent work has applied LLMs to this translation problem. Modular approaches have used LLMs to extract referring expressions with corresponding logic propositions

to then construct a full temporal logic specification [55], [21]. Relying on LLMs for direct translation, other work has mapped from language to PDDL goals [17] or full PDDL problems [56], [16]. Our work similarly translates to a task specification, but we can represent complex constraints (e.g. temporal), and we introduce a novel mechanism for automatic detection and correction of semantic errors. An interesting alternative maps language to code [57], which is highly expressive but does not easily optimize or provide behavior guarantees for long-horizon tasks.

**Re-prompting of LLMs** The quality of LLM output is greatly improved with useful context, such as few-shot in-context learning for novel tasks [6]. LLMs for TAMP are typically also provided task-relevant information, such as environment state or admissible actions [10]. Re-prompting with additional context based on LLM output has been shown to be extremely beneficial, such as with iterative action generation [8], environmental feedback [11], inadmissible actions [8], [9], [12], unmet action preconditions [58], [56], code execution errors [59], and syntactic errors in structured output [20]. Our work uses the same syntactic correction re-prompting technique as [20], but we also introduce automatic detection and correction of semantic errors via re-prompting.

## VII. CONCLUSION

This paper presented AutoTAMP, a framework for using pre-trained LLMs as both (1) translators from language task descriptions to formal task specifications (e.g. STL) via few-shot in-context learning and (2) checkers of syntactic and semantic errors via corrective re-prompting, in which we contributed a novel autoregressive re-prompting technique for semantic errors. Our experimental results show using LLMs to translate to task specifications that can be solved via a formal planner outperforms approaches that use LLMs directly as planners when handling tasks with complex geometric and temporal constraints.

We note a few limitations. First, though our results rely on using the best prompt out of several candidates, alternatives may elicit better performance. However, we expect the trends between methods to persist even with better prompts, supporting the conclusion that LLMs are not well suited for directly solving complex TAMP. Second, the cost of planning time is high, especially when there are multiple iterations of re-prompting. Further work is needed to address the runtime of formal planners and LLM inference. Third, the STL planner used in this work is not immediately applicable to manipulation tasks due to the optimization methods used in the planner; however, our approach does not depend on this specific planner, and we believe it can be integrated with STL planners more suitable for such TAMP domains.

## VIII. ACKNOWLEDGEMENTS

This work was supported by ONR under Award N00014-22-1-2478, Army Research Laboratory under DCIST CRA W911NF-17-2-0181, and MIT-IBM Watson AI Lab. This article solely reflects the conclusions of its authors.

## REFERENCES

- [1] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, “Integrated task and motion planning,” *Annual review of control, robotics, and autonomous systems*, vol. 4, pp. 265–293, 2021.
- [2] M. Fox and D. Long, “Pddl2. 1: An extension to pddl for expressing temporal planning domains,” *Journal of artificial intelligence research*, vol. 20, pp. 61–124, 2003.
- [3] E. A. Emerson, “Temporal and modal logic,” in *Formal Models and Semantics*. Elsevier, 1990, pp. 995–1072.
- [4] K. He, M. Lahijanian, L. E. Kavraki, and M. Y. Vardi, “Towards manipulation planning with temporal logic specifications,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 346–352.
- [5] R. E. Fikes and N. J. Nilsson, “Strips: A new approach to the application of theorem proving to problem solving,” *Artificial Intelligence*, vol. 2, no. 3, pp. 189–208, 1971. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0004370271900105>
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [7] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, “Tidybot: Personalized robot assistance with large language models,” *arXiv preprint arXiv:2305.05658*, 2023.
- [8] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 9118–9147.
- [9] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al., “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv preprint arXiv:2204.01691*, 2022.
- [10] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thompson, and A. Garg, “ProgPrompt: Generating situated robot task plans using large language models,” in *International Conference on Robotics and Automation (ICRA)*, 2023. [Online]. Available: <https://arxiv.org/abs/2209.11302>
- [11] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al., “Inner monologue: Embodied reasoning through planning with language models,” *arXiv preprint arXiv:2207.05608*, 2022.
- [12] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, “Text2motion: From natural language instructions to feasible plans,” *arXiv preprint arXiv:2303.12153*, 2023.
- [13] Y. Ding, X. Zhang, C. Paxton, and S. Zhang, “Task and motion planning with large language models for object rearrangement,” *arXiv preprint arXiv:2303.06247*, 2023.
- [14] N. Wake, A. Kanekihira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, “Chatgpt empowered long-step robot control in various environments: A case application,” *arXiv preprint arXiv:2304.03893*, 2023.
- [15] K. Valmecikam, A. Olmo, S. Sreedharan, and S. Kambhampati, “Large language models still can’t plan (a benchmark for llms on planning and reasoning about change),” *arXiv preprint arXiv:2206.10498*, 2022.
- [16] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, “Llm+ p: Empowering large language models with optimal planning proficiency,” *arXiv preprint arXiv:2304.11477*, 2023.
- [17] Y. Xie, C. Yu, T. Zhu, J. Bai, Z. Gong, and H. Soh, “Translating natural language to planning goals with large-language models,” *arXiv preprint arXiv:2302.05128*, 2023.
- [18] J. Pan, G. Chou, and D. Berenson, “Data-efficient learning of natural language to linear temporal logic translators for robot task specification,” *arXiv preprint arXiv:2303.08006*, 2023.
- [19] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, “Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning,” in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, 2020, pp. 440–448.
- [20] M. Skreta, N. Yoshikawa, S. Arellano-Rubach, Z. Ji, L. B. Kristensen, K. Darvish, A. Aspuru-Guzik, F. Shkurti, and A. Garg, “Errors are useful prompts: Instruction guided task programming with verifier-assisted iterative prompting,” *arXiv preprint arXiv:2303.14100*, 2023.
- [21] Y. Chen, R. Gandhi, Y. Zhang, and C. Fan, “Nl2tl: Transforming natural languages to temporal logics using large language models,” *arXiv preprint arXiv:2305.07766*, 2023.
- [22] O. Maler and D. Nickovic, “Monitoring temporal properties of continuous signals,” in *Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems: Joint International Conferences on Formal Modeling and Analysis of Timed Systems, FORMATS 2004, and Formal Techniques in Real-Time and Fault-Tolerant Systems, FTRTFT 2004, Grenoble, France, September 22–24, 2004. Proceedings*. Springer, 2004, pp. 152–166.
- [23] D. Sun, J. Chen, S. Mitra, and C. Fan, “Multi-agent motion planning from signal temporal logic specifications,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3451–3458, 2022.
- [24] C. Finucane, G. Jing, and H. Kress-Gazit, “Ltlmop: Experimenting with language, temporal logic and robot control,” in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 1988–1993.
- [25] S. Wilson, P. Glotfelter, L. Wang, S. Mayya, G. Notomista, M. Mote, and M. Egerstedt, “The robotarium: Globally impactful opportunities, challenges, and lessons learned in remote-access, distributed control of multirobot systems,” *IEEE Control Systems Magazine*, vol. 40, no. 1, pp. 26–44, 2020.
- [26] S. M. LaValle, *Planning algorithms*. Cambridge university press, 2006.
- [27] J. Ferrer-Mestres, G. Frances, and H. Geffner, “Combined task and motion planning as classical ai planning,” *arXiv preprint arXiv:1706.06927*, 2017.
- [28] C. R. Garrett, T. Lozano-Perez, and L. P. Kaelbling, “Ffrob: Leveraging symbolic planning for efficient task and motion planning,” *The International Journal of Robotics Research*, vol. 37, no. 1, pp. 104–136, 2018.
- [29] A. Akbari, J. Rosell, et al., “Task planning using physics-based heuristics on manipulation actions,” in *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2016, pp. 1–8.
- [30] F. Lagriffoul and B. Andres, “Combining task and motion planning: A culprit detection problem,” *The International Journal of Robotics Research*, vol. 35, no. 8, pp. 890–927, 2016.
- [31] J. Wolfe, B. Marthi, and S. Russell, “Combined task and motion planning for mobile manipulation,” in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 20, 2010, pp. 254–257.
- [32] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel, “Combined task and motion planning through an extensible planner-independent interface layer,” in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 639–646.
- [33] M. Colledanchise, D. Almeida, and P. Ögren, “Towards blended reactive planning and acting using behavior trees,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8839–8845.
- [34] L. P. Kaelbling and T. Lozano-Pérez, “Integrated task and motion planning in belief space,” *The International Journal of Robotics Research*, vol. 32, no. 9–10, pp. 1194–1227, 2013.
- [35] E. Fernandez-Gonzalez, B. Williams, and E. Karpas, “Scottyactivity: Mixed discrete-continuous planning with convex optimization,” *Journal of Artificial Intelligence Research*, vol. 62, pp. 579–664, 2018.
- [36] K. He, M. Lahijanian, L. E. Kavraki, and M. Y. Vardi, “Towards manipulation planning with temporal logic specifications,” in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 346–352.
- [37] M. Katayama, S. Tokuda, M. Yamakita, and H. Oyama, “Fast ltl-based flexible planning for dual-arm manipulation,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 6605–6612.
- [38] R. Takano, H. Oyama, and M. Yamakita, “Continuous optimization-based task and motion planning with signal temporal logic specifications for sequential manipulation,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 8409–8415.
- [39] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” in *ICML 2022 Workshop on Knowledge Retrieval and Language Models*, 2022. [Online]. Available: <https://openreview.net/forum?id=6p3AuAHAFin>
- [40] T. Silver, V. Hariprasad, R. S. Shuttleworth, N. Kumar, T. Lozano-Pérez, and L. P. Kaelbling, “PDDL planning with pretrained large language models,” in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. [Online]. Available: <https://openreview.net/forum?id=1QMMUB4zf1>

- [41] L. S. Zettlemoyer and M. Collins, “Learning to map sentences to logical form: structured classification with probabilistic categorial grammars,” in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005, pp. 658–666.
- [42] L. Zettlemoyer and M. Collins, “Online learning of relaxed ccg grammars for parsing to logical form,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 678–687.
- [43] Y. W. Wong and R. J. Mooney, “Learning for semantic parsing with statistical machine translation,” in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 439–446.
- [44] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn, “What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution,” in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 4163–4168.
- [45] Y. Artzi and L. Zettlemoyer, “Weakly supervised learning of semantic parsers for mapping instructions to actions,” *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 49–62, 2013.
- [46] T. M. Howard, S. Tellex, and N. Roy, “A natural language planner interface for mobile manipulators,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 6652–6659.
- [47] A. Boteanu, J. Arkin, T. Howard, and H. Kress-Gazit, “A model for verifiable grounding and execution of complex language instructions,” in *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2016.
- [48] N. Gopalan, D. Arumugam, L. L. Wong, and S. Tellex, “Sequence-to-sequence language grounding of non-markovian task specifications.” in *Robotics: Science and Systems*, vol. 2018, 2018.
- [49] R. Patel, E. Pavlick, and S. Tellex, “Grounding language to non-markovian tasks with no supervision of task specifications.” in *Robotics: Science and Systems*, vol. 2020, 2020.
- [50] H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas, “Translating structured english to robot controllers,” *Advanced Robotics*, vol. 22, no. 12, pp. 1343–1359, 2008.
- [51] J. He, E. Bartocci, D. Ničković, H. Isakovic, and R. Grosu, “Deepstl: from english requirements to signal temporal logic,” in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 610–622.
- [52] S. Mohammadinejad, J. Thomason, and J. V. Deshmukh, “Interactive learning from natural language and demonstrations using signal temporal logic,” *arXiv preprint arXiv:2207.00627*, 2022.
- [53] C. N. Bonial, L. Donatelli, J. Ervin, and C. R. Voss, “Abstract meaning representation for human-robot dialogue,” *Proceedings of the Society for Computation in Linguistics*, vol. 2, no. 1, pp. 236–246, 2019.
- [54] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, “Robots that use language,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.
- [55] J. X. Liu, Z. Yang, B. Schornstein, S. Liang, I. Idrees, S. Tellex, and A. Shah, “Lang2LTL: Translating natural language commands to temporal specification with large language models,” in *Workshop on Language and Robotics at CoRL 2022*, 2022. [Online]. Available: <https://openreview.net/forum?id=VxfjGZzrdn>
- [56] L. Guan, K. Valmeekam, S. Sreedharan, and S. Kambhampati, “Leveraging pre-trained large language models to construct and utilize world models for model-based task planning,” *arXiv preprint arXiv:2305.14909*, 2023.
- [57] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, “Code as policies: Language model programs for embodied control,” *arXiv preprint arXiv:2209.07753*, 2022.
- [58] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, “Planning with large language models via corrective re-prompting,” in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. [Online]. Available: <https://openreview.net/forum?id=cMDMRBe1TKs>
- [59] T. Silver, S. Dan, K. Srinivas, J. B. Tenenbaum, L. P. Kaelbling, and M. Katz, “Generalized planning in pdll domains with pretrained large language models,” *arXiv preprint arXiv:2305.11014*, 2023.