UNIVERSIDADE FEDERAL DO RIO GRANDE

Cristhian Lima Froes

Sistema de Execução de Tarefas baseado em Modelos Linguísticos de Larga Escala para Robôs de Serviço Universidade Federal do Rio Grande

Brasil

2024

UNIVERSIDADE FEDERAL DO RIO GRANDE

Cristhian Lima Froes

Sistema de Execução de Tarefas baseado em Modelos Linguísticos de Larga Escala para Robôs de Serviço Universidade Federal do Rio Grande

Trabalho acadêmico apresentado ao Curso de Engenharia de Computação da Universidade Federal do Rio Grande como requisito parcial para a obtenção do grau de Bacharel em Engenheira de Computação.

Orientador: Paulo Lilles Jorge Drews Júnior

Coorientador: Rodrigo da Silva Guerra

Universidade Federal do Rio Grande – FURG

Centro de Ciências Computacionais

Curso de Engenharia de Computação

Brasil

2024

Agradecimentos

Meus agradecimentos a todos os professores com os quais tive a oportunidade de compartilhar conhecimentos durante minha trajetória estudantil e acadêmica, aos meus familiares e amigos, e ao meu falecido avô, Joni Martins Lima, por todo o apoio que sempre me prestaram. O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processo nº 2024/10523-5.

Resumo

Modelos linguísticos podem ser utilizados como planejadores de tarefa, modelos de mundo e modelos de recompensa, no contexto do planejamento de tarefas para robôs de serviço? A presente monografia trata do desenvolvimento de um sistema para planejamento de tarefas para robôs de serviço, baseado em modelos linguísticos de larga escala. É proposto um método para planejamento que consiste na utilização de agentes linguísticos com os papéis de planejador, modelo de mundo e crítico. A eficácia do sistema proposto é validada por meio de experimentos com o simulador de robôs domésticos ALFWorld.

Palavras-chave: modelo linguístico. planejamento de tarefas. robótica.

Abstract

Can language models be used as task planners, world models, and reward models in the context of task planning for service robots? This monograph deals with the development of a task planning system for service robots, based on large-scale language models. A planning method is proposed that consists of using language agents with the roles of planner, world model, and critic. The effectiveness of the proposed system is validated through experiments with the ALFWorld domestic robot simulator.

Keywords: language model. task planning. robotics.

Lista de ilustrações

Figura 1 – Simulador ALFWorld 27
Figura 2 – Modelo de prompt do agente modelo de mundo 29
Figura 3 – Fluxograma de execução do método proposto 30
Figura 4 – Modelo de prompt do agente planejador 31
Figura 5 – Fluxograma de execução do baseline ReAct 32
Figura 6 – Modelo de prompt do agente crítico 32
Figura 7 – Fluxograma de execução do baseline ReAct + re-ranking com crítico . 33

Lista de tabelas

Tabela 1 – Comparação com Trabalhos Relacionados 25
Tabela 2 – ALFWorld val unseen 10 taxa de sucesso 36

Lista de abreviaturas e siglas

LLM	Large Language Model
VLM	Vision Language Model
MCTS	Monte Carlo Tree Search
PDDL	Planning Domain Definition Language

Lista de símbolos

π	Letra grega pi

Sumário

1 INTRODUÇÃO 19
1.1 Motivação 19
1.2 Definição do Problema 20
1.3 Objetivos 21
1.3.1 Objetivo Geral 21
1.3.2 Objetivos Específicos 21

2 TRABALHOS RELACIONADOS 23

3 METODOLOGIA 27
3.1 Simulador 27
3.2 Modelo Linguístico 28
3.3 Agentes Linguísticos 28
3.4 Servidor de Inferência 30

4 RESULTADOS 35

5 CONCLUSÕES E TRABALHOS FUTUROS 37

REFERÊNCIAS 39

1 Introdução

1.1 Motivação

De acordo com as normas ISO (ISO Central Secretary, 2021) um robô de serviço é definido como "um robô em uso pessoal ou profissional que realiza tarefas úteis para humanos ou equipamentos". Robôs de serviço podem ser divididos em robôs de serviço de uso pessoal e robôs de serviço de uso profissional.

Métodos clássicos da literatura sobre planejamento de tarefas para robôs de serviço, como máquinas de estado finito hierárquicas (GONZÁLEZ-SANTAMARTA et al., 2022), árvores de comportamento (COLLEDANCHISE; ÖGREN, 2018), e planejadores de tarefa baseados na linguagem de definição de domínio de planejamento (PDDL (AERONAUTIQUES et al., 1998)), tipicamente pressupõem acesso a primitivas de ação e comportamentos pré-definidos de baixo nível encapsulados em estados, que podem então ser orquestrados em comportamentos mais complexos.

Em sua grande maioria, estes métodos também assumem que o comportamento desejado seja manualmente definido, como no caso de máquinas de estado finito e árvores de comportamento, ou que uma ontologia com todas as classes de objetos com os quais se espera que o robô possa interagir, assim como os requisitos e efeitos de cada ação a ser executada pelo robô, sejam previamente conhecidos, e que o objetivo de cada tarefa seja especificado em uma linguagem de programação lógica, como em sistemas que utilizam a linguagem PDDL (AERONAUTIQUES et al., 1998).

Em ambientes onde se assume a possibilidade de interação humano-robô (HRI), como por exemplo robôs de serviço doméstico no contexto de competições como a RoboCup@Home, ou ainda em aplicações de robôs colaborativos (cobots) a ambientes de trabalho compartilhados com seres humanos, é interessante permitir que novas tarefas possam ser especificadas para dispositivos robóticos através de uma interface em linguagem natural.

O requisito da possibilidade de especificação do objetivo da tarefa em linguagem natural, contudo, não pode ser atendido por meio da aplicação isolada de planejadores baseados em PDDL (AERONAUTIQUES et al., 1998), tendo em vista os requisitos impostos pela linguagem PDDL (AERONAUTIQUES et al., 1998) de um conjunto fechado de categorias semânticas de objetos conhecidas a priori, e a necessidade de especificação da tarefa por meio de proposições lógicas, também utilizando-se de uma linguagem de domínio específica.

Dadas as limitações até então mencionadas a respeito dos métodos clássicos para

planejamento de tarefas para robôs de serviço, este trabalho se propõe a responder a seguinte pergunta: Modelos linguísticos podem ser utilizados como planejadores de tarefa, modelos de mundo e modelos de recompensa, no contexto do planejamento de tarefas para robôs de serviço?

Modelos linguísticos de larga escala (LLM, do termo em inglês Large Language Model) e modelos relacionados, como modelos de visão e linguagem (VLM, do termo em inglês Vision Language Model) têm sido aplicados nos últimos anos como a camada de orquestração de diferentes sistemas baseados em agentes autônomos, como em Wang et al. (2023), He et al. (2024), Yang et al. (2024b).

Como exemplificado em recentes revisões da literatura (ZENG et al., 2023), modelos linguísticos podem ser aplicados ao problema do planejamento de tarefas para robôs de serviço, permitindo uma interface em linguagem natural para a especificação de tarefas ao robô de serviço. O trabalho proposto nesta monografia consiste no desenvolvimento de um sistema baseado em modelos linguísticos para o planejamento de tarefas em robôs de serviço domésticos.

1.2 Definição do Problema

O problema proposto pode ser descrito como um processo de decisão de Markov parcialmente observável (POMDP) condicionado em objetivo, formado pelos seguintes componentes:

- **Estados** (S): O estado interno do ambiente, como por exemplo as poses de cada objeto presente no ambiente habitado pelo robô.
- **Ações** (A): O conjunto de comportamentos atômicos que o robô pode utilizar para interagir com o ambiente.
- **Transições** (T): Define a probabilidade de transição entre os estados $P(s'|s, a)$, que depende do estado atual e da instrução dada ao robô.
- **Observações** (O): Observações parciais sobre o ambiente, adquiridas pelo robô por meio de seus sensores.
- **Recompensa** (R): Uma função de recompensa esparsa e binária, indicando sucesso ou falha de um episódio. No ambiente simulado ALFWorld (SHRIDHAR et al., 2021), cada episódio é caracterizado por um ambiente doméstico simulado, juntamente com um comando em linguagem natural descrevendo o objetivo da tarefa do robô e um limite máximo de 50 ações consecutivas para que a tarefa seja concluída. As tarefas em si tipicamente envolvem realocação de objetos de um lugar para o outro do cenário, potencialmente com a exigência de alguma alteração de estado físico do

objeto antes do mesmo ser realocado (ex.: a tarefa "Coloque uma maçã aquecida na mesa"exige que o robô de serviço simulado utilize senso comum para procurar em diferentes locais do cenário uma maçã, e então utilize um dispositivo de forno de microondas para aquecê-la, para então largar a maçã na mesa). O sucesso é definido como verdadeiro quando o objetivo da tarefa é concluído dentro de no máximo 50 ações executadas, e falso caso contrário.

- **Objetivo** (G): Objetivo especificado por meio de uma instrução em linguagem natural (g), responsável por condicionar a política do robô.

A dinâmica do POMDP pode ser descrita da seguinte maneira:

$$P(s', o|s, a, g) = P(s'|s, a, g)P(o|s', a, g)$$

Sendo $s' \in S$, $o \in O$, $s \in S$, $a \in A$, $g \in G$.

O trabalho desenvolvido nesta monografia assume, com relação ao espaço de observação, que este consiste em uma descrição textual do conteúdo da imagem capturada por uma câmera RGB egocêntrica montada no robô.

Com relação ao espaço de ação, o trabalho desenvolvido assume acesso a um conjunto de comportamentos previamente definidos. O trabalho desenvolvido também assume que cada comportamento possui associado a si uma descrição textual exemplificando em quais situações o comportamento em questão deve ser utilizado. O trabalho desenvolvido assume também que este espaço de ação não é meramente contínuo ou discretizado, mas sim textual e paramétrico.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo do trabalho desenvolvido nesta monografia consiste em responder às seguintes questões de pesquisa: 1) Modelos linguísticos podem ser utilizados como planejadores de tarefa para escolher a próxima ação a ser executada por robôs de serviço? 2) Modelos linguísticos podem ser utilizados como estimadores de valor de recompensa para robôs de serviço? 3) Modelos linguísticos podem ser utilizados como modelos de mundo capazes de prever observações futuras em resposta a ações realizadas por robôs de serviço?

1.3.2 Objetivos Específicos

Investigar a aplicabilidade de modelos linguísticos ao planejamento de tarefas para robôs de serviço, enquanto políticas geradoras de ações no ambiente simulado

ALFWorld(SHRIDHAR et al., 2021).

Investigar a aplicabilidade de modelos linguísticos como modelos de recompensa no contexto do planejamento de tarefas para robôs de serviço no ambiente simulado ALFWorld (SHRIDHAR et al., 2021).

Investigar a aplicabilidade de modelos linguísticos como modelos de mundo no contexto do planejamento de tarefas para robôs de serviço no ambiente simulado ALFWorld (SHRIDHAR et al., 2021).

Comparar diferentes métodos que fazem utilização de modelos linguísticos como políticas de ação, modelos de recompensa e modelos de mundo em relação às suas respectivas taxas de sucesso, por meio de experimentos no simulador ALFWorld (SHRIDHAR et al., 2021)

2 Trabalhos Relacionados

Hu et al. (2023) define planejamento a nível de tarefas como o processo de dividir uma tarefa complexa em etapas menores e acionáveis. Tipicamente, o modelo linguístico é utilizado nestes casos para a tomada de decisões, ao invés de se comportar apenas como um parser.

Ahn et al. (2022) empregam modelos linguísticos para a geração de ações guiadas por afordâncias representadas por funções de valor disponíveis para diferentes habilidades pré-treinadas. Os resultados obtidos por Ahn et al. (2022) demonstram que modelos linguísticos podem ser utilizados como políticas de alto nível para geração de ações aplicadas ao planejamento de tarefas em robôs de serviço. Contudo, diferentemente do trabalho desenvolvido nesta monografia, os autores de Ahn et al. (2022) não investigam a possibilidade da utilização de modelos linguísticos também como funções de recompensa e modelos de mundo.

Wei et al. (2022) introduz uma técnica de engenharia de prompt que induz um modelo linguístico a quebrar um dado problema em um conjunto de subproblemas menores e resolvê-los passo a passo, antes de emitir sua resposta final. Os resultados obtidos por Wei et al. (2022) demonstram que a utilização da técnica Chain-of-Thought contribui para um incremento na capacidade de diferentes modelos linguísticos de resolver problemas que requisitem maior capacidade de raciocínio lógico e matemático. Contudo, (WEI et al., 2022) não investigam, ao contrário do método proposto nesta monografia, a integração do modelo linguístico em um ambiente simulado interativo, e nem a utilização de modelos linguísticos como funções de recompensa ou como modelos de mundo.

Yao et al. (2022) introduz uma extensão do método proposto em Wei et al. (2022), que permite ao modelo linguístico invocar rotinas externas pré-definidas, denominadas "ferramentas", de modo a interagir com o ambiente em ciclos alternados de raciocínio e invocação de ferramentas. Os resultados obtidos por Yao et al. (2022) demonstram que modelos linguísticos podem ser utilizados como planejadores de tarefas para ambientes simulados; contudo, diferentemente do trabalho desenvolvido nesta monografia, Yao et al. (2022) não investiga a utilização de modelos linguísticos como modelos de mundo e de recompensa para o planejamento de tarefas.

Hazra, Martires e Raedt (2024) empregam modelos linguísticos para a geração de ações guiadas por conhecimento de domínio aprendido e recompensa a longo prazo, e busca heurística para a seleção da melhor sequência de ações. Os resultados obtidos por Hazra, Martires e Raedt (2024) também demonstram que modelos linguísticos podem ser utilizados como políticas de alto nível para geração de ações, no contexto do planejamento

de tarefas para robôs de serviço. Embora os autores de Hazra, Martires e Raedt (2024) demonstrem a utilização de modelos de recompensa de modo a auxiliar no processo de pontuação das ações geradas pelo modelo linguístico, eles não avaliam a possibilidade da utilização do próprio modelo linguístico como modelo de recompensa, e também não investigam a possibilidade de utilização do modelo linguístico como modelo de mundo, diferentemente do trabalho desenvolvido nesta monografia.

Yao et al. (2024) propõe a utilização de um algoritmo de busca em árvore sobre o espaço de raciocínio do modelo linguístico guiado por uma heurística de seleção de folhas que se utiliza do próprio modelo para a avaliação dos diferentes caminhos na árvore. Os resultados obtidos por Yao et al. (2024) demonstram que modelos linguísticos podem ser utilizados como funções de recompensa no contexto de um framework para resolução de problemas que demandem capacidades não triviais de planejamento e raciocínio, porém não investigam a utilização de modelos linguísticos como modelos de mundo, e também não se propõem a validar seu método para planejamento de tarefas em ambientes simulados, diferentemente do trabalho desenvolvido nesta monografia.

Zhou et al. (2023) propõe a combinação de um algoritmo de busca em árvore, como proposto em Yao et al. (2024) com a utilização de ferramentas de prompt para interação com o ambiente, como proposto em Yao et al. (2022), promovendo uma sinergia entre raciocínio, planejamento e ação. Os resultados obtidos por Zhou et al. (2023) demonstram que modelos linguísticos podem ser utilizados tanto como planejadores quanto como funções de recompensa, que podem ser utilizados juntamente com um algoritmo de busca em árvore para o planejamento de tarefas. Contudo, o método proposto em Zhou et al. (2023) não investiga a utilização de modelos linguísticos como modelos de mundo, sendo aplicável somente a ambientes determinísticos e com ações reversíveis. O método proposto nesta monografia se diferencia do proposto em Zhou et al. (2023) por utilizar-se também de modelos linguísticos como modelos de mundo, o que elimina a necessidade de que o ambiente seja determinístico e reversível.

Hao et al. (2023) propõe a utilização de modelos linguísticos como políticas geradoras de ações, modelos de recompensa, e também como modelos de mundo. Os autores de Hao et al. (2023) também propõem a utilização de um algoritmo de busca em árvore para a geração de planos de ação a partir da utilização dos modelos de recompensa e de mundo e da política de ação. O trabalho desenvolvido nesta monografia se diferencia de Hao et al. (2023) por investigar a aplicação desses modelos em um ambiente simulado e parcialmente observável, direcionado para a robótica doméstica, enquanto os autores de Hao et al. (2023) validam seu método em um ambiente simulado totalmente observável e voltado para a manipulação de objetos.

A Tabela 1 demonstra uma comparação entre o trabalho desenvolvido nesta monografia e métodos relacionados da literatura.

Tabela 1 – Comparação com Trabalhos Relacionados

Método	Planejador	Recompensa	Modelo de Mundo	Robótica
Ahn et al. (2022)	V	F	F	V
Wei et al. (2022)	F	F	F	F
Yao et al. (2022)	V	F	F	F
Hazra, Martires e Raedt (2024)	V	F	F	V
Yao et al. (2024)	F	V	F	F
Zhou et al. (2023)	V	V	F	F
Hao et al. (2023)	V	V	V	V
Método Proposto	V	V	V	V

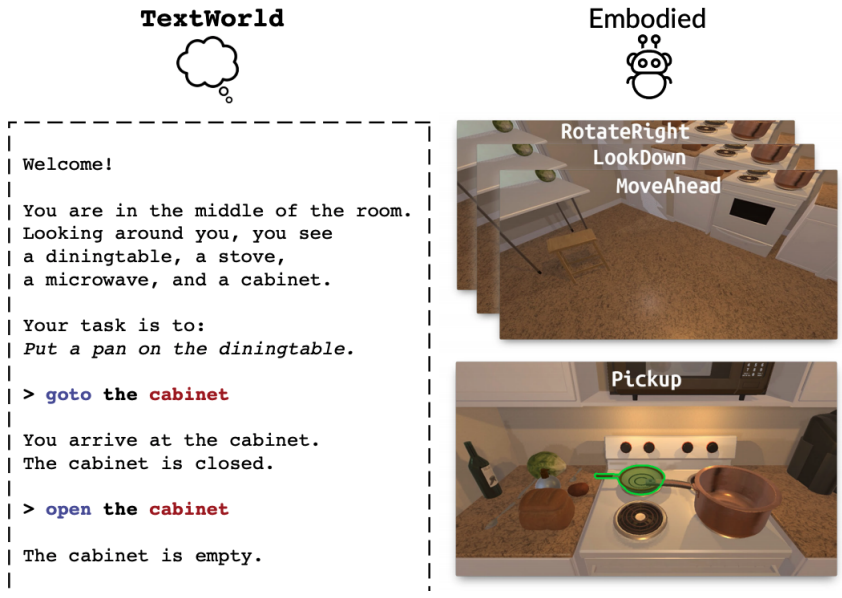
3 Metodologia

A seção metodológica desta monografia discutirá a respeito das escolhas técnicas envolvidas na arquitetura do método proposto.

3.1 Simulador

Como ambiente simulado para ablação dos métodos e baselines propostos, será utilizado o simulador ALFWorld (SHRIDHAR et al., 2021), composto por ambientes domésticos virtuais interagíveis e manualmente autorados. O simulador ALFWorld será utilizado em função do seu espaço de observação textual e espaço de ação abstrato composto por primitivas relevantes para ambientes domésticos, como pick, place e open. A métrica de validação a ser utilizada será a taxa de sucesso com a qual os métodos a serem considerados conseguem utilizar um conjunto pré-definido de primitivas de atuação para reorganizar o ambiente e seus objetos de modo a atender a requisições realizadas em linguagem natural. A Figura 1 ilustra graficamente os espaços de observação e ação do simulador.

Figura 1 – Simulador ALFWorld



Fonte: Shridhar et al. (2021, p. 1)

Simulador ALFWorld

3.2 Modelo Linguístico

A arquitetura transformer proposta em Vaswani et al. (2017), composta por um encoder e um decoder que utilizam camadas repetidas de auto-atenção, recebe como entrada no encoder uma sequência de tokens, e gera um único token de cada vez como saída do decoder. Modelos linguísticos auto-regressivos utilizam um tokenizador para converter texto em uma sequência de tokens, usada como entrada para uma variante da arquitetura transformer composta somente pelo decoder, que então gera em sua saída o próximo token de forma iterativa, sendo que a cada etapa da iteração a entrada da rede consiste nos tokens iniciais do prompt concatenados com os tokens gerados previamente pelo modelo.

Como demonstrado em Brown et al. (2020), ao serem treinados em datasets de texto de larga escala e terem sua contagem de parâmetros aumentada, modelos de linguagem natural adquirem, dentre outras capacidades, aprendizado em contexto com base em exemplos, assim como conhecimento aberto sobre senso comum. O conhecimento sobre senso comum adquirido por estes modelos pode ser reaproveitado no contexto da robótica, auxiliando no encadeamento de comportamentos pré-definidos ou pré-treinados de modo a completar tarefas especificadas em linguagem natural.

Como modelo linguístico a ser utilizado dentro da arquitetura de tomada de decisão do agente, será utilizado o modelo open-source Qwen2.5 14B (YANG et al., 2024a) (Qwen Team, 2024). A escolha deste modelo se deve ao seu desempenho satisfatório, dadas as restrições de capacidade computacional disponível aos autores da presente monografia durante sua execução.

3.3 Agentes Linguísticos

O método proposto consiste na utilização iterativa de três agentes de modelo linguístico com prompts distintos, e encontra-se exemplificado pelo fluxograma da figura Figura 3.

O primeiro agente é um planejador iterativo semelhante ao método ReAct (YAO et al., 2022). Este agente escolhe iterativamente a próxima ação, ou primitiva de atuação, a ser executada pelo robô no simulador. Essa escolha é realizada utilizando Chain-of-Thought (WEI et al., 2022) para induzir raciocínio no modelo sobre as observações a ações previamente executadas, antes de dar a resposta final com a próxima ação a ser executada. A Figura 4 demonstra o modelo de prompt utilizado no agente planejador.

O segundo agente é um substituto para um modelo de mundo. Ele recebe um histórico de ações e observações previamente executadas pelo robô no simulador, além da próxima ação a ser executada, escolhida pelo agente planejador, e usa Chain-of-Thought

(WEI et al., 2022) para induzir um raciocínio sobre esse histórico, antes de responder com uma estimativa da próxima observação resultante da ação selecionada pelo planejador. A Figura 2 demonstra o modelo de prompt utilizado no agente modelo de mundo.

O terceiro agente é um crítico de ações. Ele recebe o histórico de ações e observações previamente executadas, assim como a ação escolhida pelo agente planejador, e a observação estimada pelo agente modelo de mundo, e usa Chain-of-Thought (WEI et al., 2022) para induzir um raciocínio sobre o possível resultado da ação escolhida pelo planejador, e então responder com uma nota de 0 a 10 julgando a ação escolhida pelo planejador. A Figura 6 demonstra o modelo de prompt utilizado no agente crítico de ações.

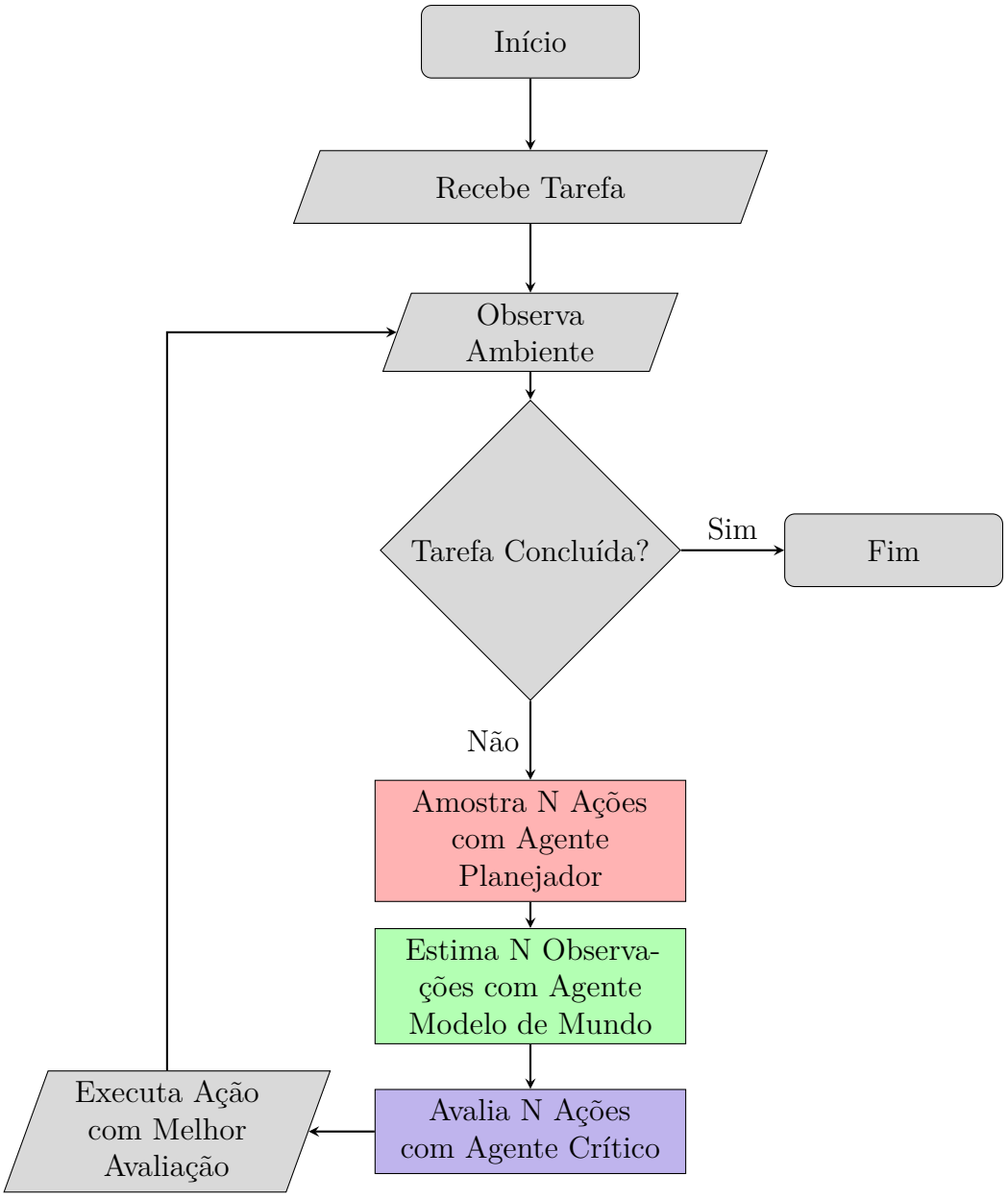
Durante a execução iterativa do método proposto, ao receber uma nova observação do ambiente simulado, são amostradas N potenciais próximas ações do agente planejador. Estas potenciais próximas ações são alimentadas no agente modelo de mundo, e as observações resultantes alimentadas no agente crítico, de modo a obter uma pontuação para cada ação originalmente proposta pelo agente planejador. A ação amostrada com a melhor pontuação é executada no ambiente simulado, e adicionada junto com a observação resultante ao histórico de ações e observações prévias.

Para a comparação com o método proposto, foram escolhidos dois baselines. O primeiro deles, ReAct(YAO et al., 2022), consiste na utilização somente do agente planejador, e na amostragem de uma única ação deste agente. Seu funcionamento encontra-se exemplificado pelo fluxograma na figura Figura 5. O segundo baseline, React(YAO et al., 2022) + re-ranking com crítico, consiste na utilização com amostragem e pontuação de ações somente dos agentes planejador e crítico, sem a utilização de um agente modelo de mundo. Seu funcionamento encontra-se exemplificado pelo fluxograma na figura Figura 7.

Figura 2 – Modelo de prompt do agente modelo de mundo

Role
You are the internal world model of a general purpose service robot inside a house. You answer by predicting the next observation, given previous observations and actions.
Observation/Action History
{*previous_actions*}
Current Action
{*current_action*}

Figura 3 – Fluxograma de execução do método proposto



3.4 Servidor de Inferência

Todos os métodos utilizam o mesmo modelo linguístico quantizado para 4 bits e hospedado localmente utilizando a ferramenta open-source Ollama, com a temperatura de amostragem padrão de 0.8 e top_p padrão de 0.9. De modo a restringir a saída do modelo linguístico a um dicionário JSON com uma estrutura pré-definida composta por um campo para raciocínio seguido por outro campo para a ação, observação e pontuação de recompensa, todos os três agentes utilizam a interface de geração estruturada disponibilizada pela ferramenta Ollama. Esta interface basicamente zera as probabilidades de saída de todos os tokens do modelo que não se adequam a uma gramática livre de contexto

Figura 4 – Modelo de prompt do agente planejador

Role
You are a general purpose service robot operating inside a 3d simulation of a house. You observe the world around you with a head-mounted ego-centric camera. You can only carry at most exactly 1 object in your inventory.
Observation/Action History
{*previous_actions*}
Task
{*task*}
Available Actions
{*available_actions*}
Expected Response Format
{*expected_response_format*}

préviamente estabelecida.

Figura 5 – Fluxograma de execução do baseline ReAct

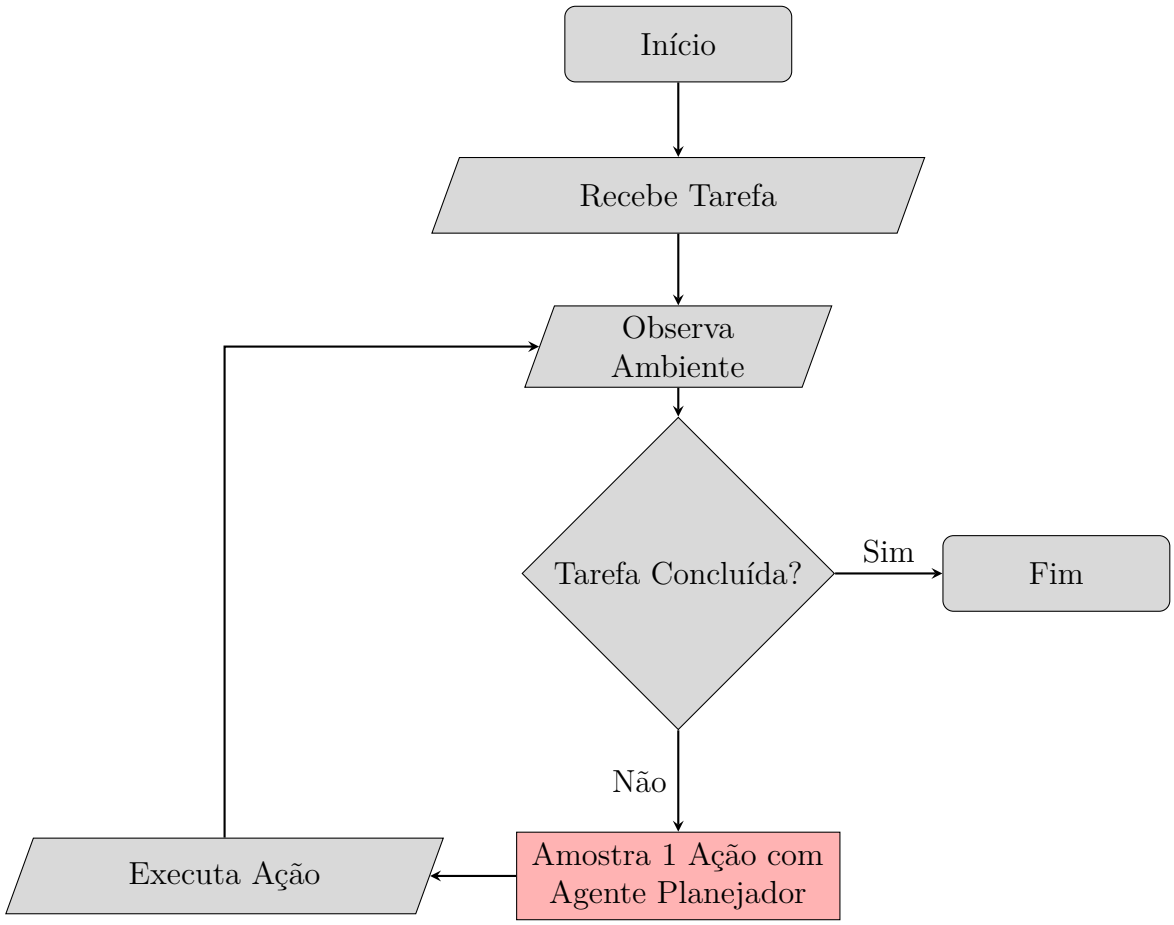
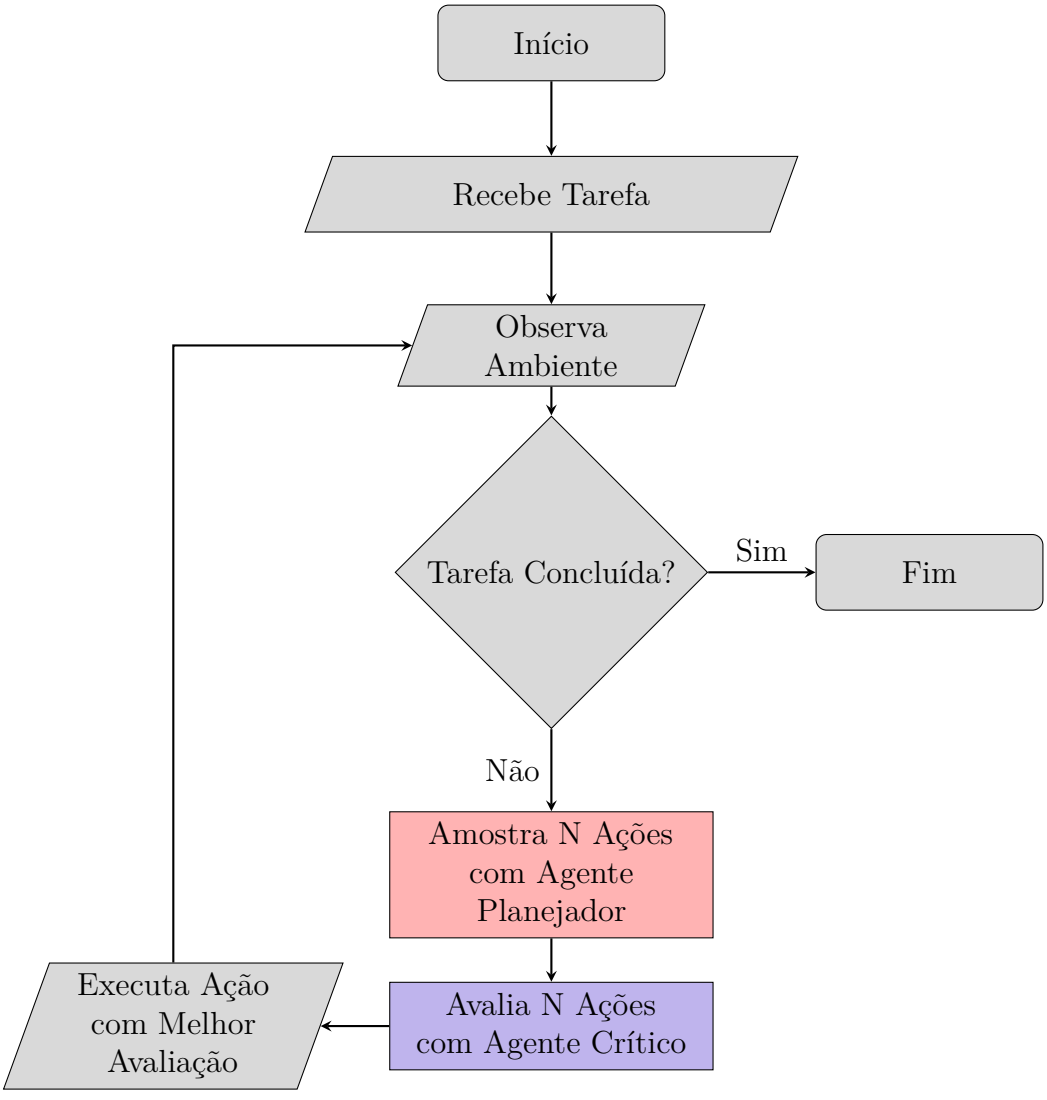


Figura 6 – Modelo de prompt do agente crítico

Role
You are the internal policy value estimator of a general purpose service robot inside a house. You answer by predicting a score from 0 to 10 judging how well the service robot is accomplishing it's task. The robot can carry a maximum of exactly 1 object in it's inventory.
Observation/Action History
{*previous_actions*}

Figura 7 – Fluxograma de execução do baseline ReAct + re-ranking com crítico



4 Resultados

A Tabela 2 mostra a taxa de sucesso dos métodos estudados, ao serem aplicados a um subconjunto de 10 casos de teste amostrados do split *val_unseen* do benchmark ALFWorld (SHRIDHAR et al., 2021). Para os métodos que utilizam amostragem de ações, foi utilizado um valor de $N = 5$ ações amostradas.

Como é possível observar, a utilização da amostragem de ações junto com pontuação e escolha da melhor ação pelo agente crítico, porém sem a simulação com o agente modelo de mundo, contribui para uma queda relativa na taxa de sucesso média, de 43,33% para 36,66%, em relação ao baseline sem amostragem baseado em ReAct (YAO et al., 2022). A hipótese que a presente monografia sugere a respeito deste resultado é a de que, sem o contexto adicional fornecido pela observação gerada pelo agente modelo de mundo, o agente crítico possui maior dificuldade em atribuir uma nota para as ações amostradas do agente planejador, o que resulta em uma estimativa de valor subótima e, consequentemente, em uma queda na taxa de sucesso.

Esta hipótese pode ser corroborada pela observação de que, ao utilizar-se o agente modelo de mundo para a simulação da próxima possível observação antes da avaliação de cada ação pelo modelo crítico, a taxa de sucesso do método deixa de sofrer uma queda, e, ao invés disso, obtém um incremento relativo de 43,66% para 58,51% em relação ao baseline baseado em ReAct.

Uma observação passível de análises futuras, e que foi verificada em todos os três métodos de maneira esporádica, é um fenômeno que o trabalho desenvolvido nesta monografia descreve como "semantic real2sim gap". O trabalho apresentado nesta monografia descreve este fenômeno como sendo o resultado de um ambiente simulado com um conjunto de regras e métricas de validação demasiadamente estritos, e que fazem com que soluções alternativas porém semanticamente corretas encontradas pelo modelo linguístico para a realização de uma tarefa sejam consideradas falhas pela heurística de validação pré-definida do ambiente. Um exemplo verificado neste trabalho ocorreu durante a execução por parte do agente planejador de uma tarefa cujo objetivo era algo semelhante a "put a cloth on the shelf". Após explorar o ambiente, o agente eventualmente encontrou um item da classe "towel", raciocinou que uma toalha é um tipo de tecido (cloth) e então prosseguiu com a tarefa de pegar a toalha e colocá-la na estante. Porém, como a heurística de validação utilizada pelo ambiente simulado ALFWorld considera "towel"e "cloth"como duas categorias distintas, o agente ficou repetidas vezes tentando colocar a toalha na estante, até que o limite de iterações do ambiente foi alcançado, e o episódio foi considerado como uma falha pela métrica de validação do ambiente simulado.

Tabela 2 – ALFWorld val unseen 10 taxa de sucesso

Método	SR(avg)	SR(min)	SR(max)
ReAct	43,33%	40%	50%
ReAct + crítico	36,66%	30%	40%
ReAct + modelo de mundo e crítico	58,51%	50%	70%

5 Conclusões e Trabalhos Futuros

A taxa de sucesso não trivial obtida por todos os métodos no conjunto de validação do benchmark ALFWorld (SHRIDHAR et al., 2021) indica que modelos linguísticos podem ser utilizados como planejadores de tarefa para a orquestração de primitivas externas de pick, place e navigate em robôs de serviço, o que responde à pergunta de pesquisa 1. No entanto, ao menos para modelos linguísticos de pequena escala passíveis de serem executados em dispositivos de borda, como os utilizados neste trabalho, ainda existe espaço para melhorias futuras, como, por exemplo, a destilação e treinamento supervisionado pelas saídas geradas por modelos linguísticos maiores e mais robustos, como realizado em (TAORI et al., 2023).

O decremento na taxa de sucesso observado no método que utiliza somente o agente crítico, sem o agente modelo de mundo, indica que os modelos linguísticos de pequena escala estudados neste trabalho são incapazes de avaliar de forma robusta suas próprias ações no contexto da robótica de serviço em ambientes domésticos, o que responde à pergunta de pesquisa 2. No entanto, é importante ressaltar que os experimentos realizados neste trabalho se limitaram a modelos de pequena escala, sendo que, mais uma vez, uma direção de pesquisa para trabalhos futuros consiste na geração de dados para aprendizado supervisionado a partir de modelos maiores e mais robustos do que os considerados neste trabalho, com o intuito da realização do ajuste-fino de modo a aprimorar as capacidades do agente crítico de ações.

O incremento na taxa de sucesso obtido por meio da introdução do agente modelo de mundo indica que, ao menos para ambientes com observações textuais, modelos linguísticos podem ser utilizados como modelo de mundo, respondendo à pergunta de pesquisa 3. No entanto, uma limitação do método empregado no nosso trabalho é que ele não pode ser utilizado para modelar observações visuais, que são de suma importância para problemas na robótica doméstica e de serviço. O trabalho desenvolvido sugere a hipótese de que a utilização de engenharia de prompt pura, ao contrário de ajuste fino supervisionado sobre observações reais do ambiente no nosso agente modelo de mundo, resulta em uma perda da fidedignidade das observações modeladas pelo agente, o que possivelmente impõe um limite na efetividade do planejamento realizado pelo sistema como um todo envolvendo os três agentes. Trabalhos futuros podem explorar a utilização de modelos treinados para geração de vídeo, como o recentemente anunciado Cosmos (NVIDIA et al., 2025), para a substituição do agente modelo de mundo baseado em modelos linguísticos, por um modelo propriamente desenvolvido para utilização como modelo de mundo.

Por fim, uma direção a ser explorada para trabalhos futuros é a integração do

método apresentado nesta monografia com um robô de serviço manipulador móvel em ambientes da vida real. Tal integração possivelmente exigirá a utilização de modelos de detecção de objetos, como YOLO-World (CHENG et al., 2024), aliados a uma nuvem de pontos proveniente de uma câmera de profundidade, para a localização de objetos manipuláveis no ambiente do robô. Para a implementação de ações como abrir ou fechar as portas de eletrodomésticos e peças de mobília, uma direção promissora é a integração de métodos de aprendizado por imitação, como ACT (ZHAO et al., 2023), VQ-BeT (LEE et al., 2024) e Diffusion Policy (CHI et al., 2023), assim como modelos de visão-linguagem-ação, como OpenVLA (KIM et al., 2024) e π_0. (BLACK et al., 2024).

Referências

AERONAUTIQUES, C. et al. Pddl| the planning domain definition language. *Technical Report, Tech. Rep.*, 1998. Citado na página 19.

AHN, M. et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. Citado 2 vezes nas páginas 23 e 25.

BLACK, K. et al.

\

pi_0 : *Avision − language − actionflowmodelforgeneralrobotcontrol.arXiv preprint arXiv:2410.24164*, 2024.*Citadonapágina*38.

BROWN, T. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020. Citado na página 28.

CHENG, T. et al. Yolo-world: Real-time open-vocabulary object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2024. p. 16901–16911. Citado na página 38.

CHI, C. et al. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. Citado na página 38.

COLLEDANCHISE, M.; ÖGREN, P. *Behavior trees in robotics and AI: An introduction*. [S.l.]: CRC Press, 2018. Citado na página 19.

GONZÁLEZ-SANTAMARTA, M. Á. et al. Yasmin: Yet another state machine. In: SPRINGER. *Iberian Robotics conference*. [S.l.], 2022. p. 528–539. Citado na página 19.

HAO, S. et al. Reasoning with language model is planning with world model. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2023. p. 8154–8173. Citado 2 vezes nas páginas 24 e 25.

HAZRA, R.; MARTIRES, P. Z. D.; RAEDT, L. D. Saycanpay: Heuristic planning with large language models using learnable domain knowledge. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2024. v. 38, n. 18, p. 20123–20133. Citado 3 vezes nas páginas 23, 24 e 25.

HE, H. et al. *WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models*. 2024. Disponível em: <https://arxiv.org/abs/2401.13919>. Citado na página 20.

HU, Y. et al. *Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis*. 2023. Disponível em: <https://arxiv.org/abs/2312.08782>. Citado na página 23.

ISO Central Secretary. *Robotics — Vocabulary*. Geneva, CH, 2021. v. 2021. Citado na página 19.

KIM, M. J. et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. Citado na página 38.

LEE, S. et al. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024. Citado na página 38.

NVIDIA et al. *Cosmos World Foundation Model Platform for Physical AI*. 2025. Disponível em: <https://arxiv.org/abs/2501.03575>. Citado na página 37.

Qwen Team. *Qwen2.5: A Party of Foundation Models*. 2024. Disponível em: <https://qwenlm.github.io/blog/qwen2.5/>. Citado na página 28.

SHRIDHAR, M. et al. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. [s.n.], 2021. Disponível em: <https://arxiv.org/abs/2010.03768>. Citado 5 vezes nas páginas 20, 22, 27, 35 e 37.

TAORI, R. et al. *Stanford Alpaca: An Instruction-following LLaMA model*. [S.l.]: GitHub, 2023. <https://github.com/tatsu-lab/stanford_alpaca>. Citado na página 37.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 28.

WANG, G. et al. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. Citado na página 20.

WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, v. 35, p. 24824–24837, 2022. Citado 4 vezes nas páginas 23, 25, 28 e 29.

YANG, A. et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. Citado na página 28.

YANG, J. et al. *SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering*. 2024. Disponível em: <https://arxiv.org/abs/2405.15793>. Citado na página 20.

YAO, S. et al. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, v. 36, 2024. Citado 2 vezes nas páginas 24 e 25.

YAO, S. et al. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. Citado 6 vezes nas páginas 23, 24, 25, 28, 29 e 35.

ZENG, F. et al. *Large Language Models for Robotics: A Survey*. 2023. Disponível em: <https://arxiv.org/abs/2311.07226>. Citado na página 20.

ZHAO, T. Z. et al. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. Citado na página 38.

ZHOU, A. et al. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023. Citado 2 vezes nas páginas 24 e 25.