A Details of Numerical Reasoning

In this section, we describe the data processing workflow (including building the testing datasets and synthesizing training data) and list the prompts used for different methods. For a fair comparison, we use the same prompts for CoT and the reasoning mode of ToolkenGPT. With the same example questions, we label the calculation process in place to get the prompts for ReAct[2]. For the tool mode of ToolkenGPT, we randomly sample 4 examples of the specified tool from the training set, and transform them into ReAct-style prompts. Because a large number of tools are used, we show the prompts in the supplementary file instead of listing them here.

A.1 GSM8K-XL

A.1.1 Level-up Strategy

To magnify the numbers for the GSM8K to build our dataset GSM8K-XL, we follow the steps outlined below:

1. We prompt the `gpt-3.5-turbo` with two examples to replace the numbers with appropriate placeholders. The prompt is presented below.
2. In order to validate the correctness of the number replacements, we develop a solving function for the GSM8k dataset, which utilizes the formulas presented in the provided answers and produces the expected solutions. By substituting the original numbers back into the rewritten question-answer pairs and executing the solving function, we can identify any erroneous rewrites and exclude them from our dataset. This ensures the reliability and accuracy of the number replacements.
3. Once we have confirmed the accuracy of the rewriting process, we can proceed with magnifying the numbers. We systematically replace the original numbers with their respective cubic values, exponentially amplifying their magnitudes. However, it is essential to exercise caution in certain scenarios. For instance, when dealing with percentage values (e.g., 10%), in most cases, a direct cubic magnification to 1000% would be inappropriate. In such cases, we maintain the original values to preserve their intended meaning.
4. With the magnified numbers in place, we seamlessly substitute the placeholders within the rewritten question-answer pairs. By employing the solving function developed earlier, we can process the magnified formulas and obtain accurate solutions, leading to the GSM8K-XL dataset.

```
{
    "role": "system",
    "content": "Your task is to rewrite the given question and answer in a
        more general form using variables instead of specific numbers."
},
{
    "role": "user",
    "content": "Question: Janet's ducks lay 16 eggs per day. She eats three
        for breakfast every morning and bakes muffins for her friends every
        day with four. She sells the remainder at the farmers' market daily
        for $2 per fresh duck egg. How much in dollars does she make every
        day at the farmers' market\nAnswer: Janet sells 16 - 3 - 4 =
        <<16-3-4=9>>9 duck eggs a day.\nShe makes 9 * 2 = $<<9*2=18>>18
        every day at the farmer's market.\n#### 18"
},
{
    "role": "assistant",
    "content": "Question: Janet's ducks lay {v_1} eggs per day. She eats
        {v_2} for breakfast every morning and bakes muffins for her friends
```

[2]As Yao et al. [62] doesn't evaluate ReAct on numerical reasoning, we don't follow the exact format (i.e. trajectories marked by `[THINK]` and `[ACT]`), but design a more natural way to combine CoT reasoning and tool calling (as shown by our prompts).

```
        every day with {v_3}. She sells the remainder at the farmers'
        market daily for ${v_4} per fresh duck egg. How much in dollars
        does she make every day at the farmers' market?\nAnswer: Janet
        sells {v_1} - {v_2} - {v_3} = <<{v_1}-{v_2}-{v_3}={r_1}>>{r_1} duck
        eggs a day.\nShe makes {r_1} * {v_4} = $<<{r_1}*{v_4}={r_2}>>{r_2}
        every day at the farmer's market.\n#### {r_2}"
},
{
    "role": "user",
    "content": "Question: A wooden bridge can carry no more than 5000
        pounds. A delivery truck filled with identical boxes, each weighing
        15 pounds, will pass over the bridge. The combined weight of the
        driver and the empty truck is 3755 pounds. What is the maximum
        number of boxes which can be loaded onto the truck while not
        exceeding the bridge's weight limit?\nAnswer: The boxes can weigh
        up to 5000 pounds - 3755 pounds = <<5000-3755=1245>>1245 pounds in
        total.\nThere can be 1245 / 15 = <<1245/15=83>>83 boxes loaded onto
        the truck without exceeding the bridge's weight limit.\n#### 83"
},
{
    "role": "assistant",
    "content": "Question: A wooden bridge can carry no more than {v_1}
        pounds. A delivery truck filled with identical boxes, each weighing
        {v_2} pounds, will pass over the bridge. The combined weight of the
        driver and the empty truck is {v_3} pounds. What is the maximum
        number of boxes which can be loaded onto the truck while not
        exceeding the bridge's weight limit?\nAnswer: The boxes can weigh
        up to {v_1} pounds - {v_3} pounds = <<{v_1}-{v_3}={r_1}>>{r_1}
        pounds in total.\nThere can be {r_1} / {v_2} =
        <<{r_1}/{v_2}={r_2}>>{r_2} boxes loaded onto the truck without
        exceeding the bridge's weight limit.\n#### {r_2}"
},
{
    "role": "user",
    "content": [INPUT]
}
```

A.1.2 Training Details

As mentioned in Section 4.1, the Toolken embeddings are trained with a subset of 5,063 examples. An additional 1,000 examples are reserved for validation. The embeddings was trained with a learning rate of 5e-4, performing early stopping based on the development set, with a maximum of 10 epochs.

A.1.3 Prompt for GSM8K-XL Dataset

Prompt for Direct Prompting with ChatGPT:

```
Solve the following math problem step by step, and then provide the final
answer in the format: 'So, the answer is xxx.'

[QUESTION]
```

Prompt for Chain of Thought (CoT) and ToolkenGPT:

```
Answer the following questions step by step.

Question: Mark has 3 tanks for pregnant fish. Each tank has 4 pregnant
fish and each fish gives birth to 20 young. How many young fish does he
have at the end?
```

```
Answer: He has 4*3=12 pregnant fish They give birth to 12*20=240 fish ####
240

Question: The math questions in a contest are divided into three rounds:
easy, average, and hard. There are corresponding points given for each
round. That is 2, 3, and 5 points for every correct answer in the easy,
average, and hard rounds, respectively. Suppose Kim got 6 correct answers
in the easy; 2 correct answers in the average; and 4 correct answers in
the difficult round, what are her total points in the contest?
Answer: Kim got 6 points/round x 2 round = 12 points in the easy round.
She got 2 points/round x 3 rounds = 6 points in the average round. She got
4 points/round x 5 rounds = 20 points in the difficult round. So her total
points is 12 points + 6 points + 20 points = 38 points. #### 38

Question: A clothing store sells 20 shirts and 10 pairs of jeans. A shirt
costs $10 each and a pair of jeans costs twice as much. How much will the
clothing store earn if all shirts and jeans are sold?
Answer: Twenty shirts amount to $10 x 20 = $200. The cost of each pair of
jeans is $10 x 2 = $20. So 10 pairs of jeans amount to $20 x 10 = $200.
Therefore, the store will earn $200 + $200 = $400 if all shirts and jeans
are sold. #### 400

Question: Arnold's collagen powder has 18 grams of protein for every 2
scoops. His protein powder has 21 grams of protein per scoop. And his
steak has 56 grams of protein. If he has 1 scoop of collagen powder, 1
scoop of protein powder and his steak, how many grams of protein will he
consume?
Answer: 2 scoops of collagen powder have 18 grams of protein and he only
has 1 scoop so he consumes 18/2 = 9 grams of protein He has 9 grams
collagen powder, 21 grams of protein powder and 56 grams in his steak for
a total of 9+21+56 = 86 grams of protein #### 86

Question: [QUESTION]
Answer:
```

Prompt for ReAct:

```
Answer the following questions with <add>, <subtract>, <multiply>,
<divide> operators

Question: Mark has 3 tanks for pregnant fish. Each tank has 4 pregnant
fish and each fish gives birth to 20 young. How many young fish does he
have at the end?
Answer: He has 4*3=<multiply>(4, 3)=12 pregnant fish They give birth to
12*20=<multiply>(12, 20)=240 fish #### 240

Question: The math questions in a contest are divided into three rounds:
easy, average, and hard. There are corresponding points given for each
round. That is 2, 3, and 5 points for every correct answer in the easy,
average, and hard rounds, respectively. Suppose Kim got 6 correct answers
in the easy; 2 correct answers in the average; and 4 correct answers in
the difficult round, what are her total points in the contest?
Answer: Kim got 6 points/round x 2 round = <multiply>(6, 2)=12 points in
the easy round. She got 2 points/round x 3 rounds = <multiply>(2, 3)=6
points in the average round. She got 4 points/round x 5 rounds =
<multiply>(4, 5)=20 points in the difficult round. So her total points is
12 points + 6 points + 20 points = <add>(12, 6, 20)=38 points. #### 38

```

```
Question: A clothing store sells 20 shirts and 10 pairs of jeans. A shirt
costs $10 each and a pair of jeans costs twice as much. How much will the
clothing store earn if all shirts and jeans are sold?
Answer: Twenty shirts amount to $10 x 20 = $<multiply>(10, 20)=200. The
cost of each pair of jeans is $10 x 2 = $<multiply>(10, 2)=20. So 10 pairs
of jeans amount to $20 x 10 = $<multiply>(20, 10)=200. Therefore, the
store will earn $200 + $200 = $<add>(200, 200)=400 if all shirts and jeans
are sold. #### 400

Question: Arnold's collagen powder has 18 grams of protein for every 2
scoops. His protein powder has 21 grams of protein per scoop. And his
steak has 56 grams of protein. If he has 1 scoop of collagen powder, 1
scoop of protein powder and his steak, how many grams of protein will he
consume?
Answer: 2 scoops of collagen powder have 18 grams of protein and he only
has 1 scoop so he consumes 18/2 = <divide>(18, 2)=9 grams of protein He
has 9 grams collagen powder, 21 grams of protein powder and 56 grams in
his steak for a total of 9+21+56 = <add>(9, 21, 56)=86 grams of protein
#### 86

Question: [QUESTION]
Answer:
```

A.2 FuncQA

A.2.1 Training Details

As mentioned in Section 4.1, Toolken embeddings are trained using a subset of 611 examples, with an additional 39 examples reserved for development/validation purposes. The learning rate we use is 1e-4, and we perform early stopping based on the development set, with the maximal training epochs to be 20.

A.2.2 Prompt for Synthetic Training Data

In Section 4.1, we discussed the utilization of ChatGPT for synthesizing the training set. To create the training data, we begin by manually crafting two examples that adhere to the desired format, and then use the follow specific prompt to generate more examples. However, it is important to acknowledge that the prompt does not guarantee the generation of examples that strictly conform to the required format. So, we apply a filtering process to remove any non-conforming instances. Furthermore, the generation process often produces duplicate examples, necessitating a subsequent de-duplication step.

```
You are a math question generator for teachers, and your task is to
generate some questions and answers using function [FUNC] to solve and can
be solved within one single step. You do not need to give specific
numbers, so that the teachers can fill any numbers they want. Here are two
examples that use the function [FUNC].

[EXAMPLE_1]

[EXAMPLE_2]

[FUNC] is a function to [DESCRIPTION]. Now, let's mimic the format of
examples to generate various real world QA pairs using the function [FUNC]
that can be solved within one step. The numbers should be replaced by
[ARG] and [ANSWER] as the examples given above.
```

A.2.3 Prompt for FuncQA One-Hop

Prompt for Zero-Shot with ChatGPT:

```
Solve the following math problem, and then provide the final answer in the
format: ‘So, the answer is xxx.’

[QUESTION]
```

Prompt for Chain of Thought (CoT) and ToolkenGPT:

```
Q: If Amy’s income increases by 4% annually, how many times will it
multiply in 11 years?
A: In 11 years, Amy’s income will increase by 1.04^11=1.54 times. So, the
answer is 1.54.

Q: If a store sells 147 bananas today and 354 more bananas tomorrow, how
many bananas does the store sell in total?
A: The store sells 147 bananas today and 354 more bananas tomorrow, so the
total number of bananas sold is 147+354=501. So, the answer is 501.

Q: A man had 789.4 dollars in his wallet. He spent 11.99 dollars on a
movie ticket. How much money does he have left now?
A: The man had 789.4 dollars in his wallet and spent 11.99 dollars on a
movie ticket, so he has 789.4-11.99=777.41 dollars left. So, the answer is
777.41 dollars.

Q: If a cake weighs 3.77 pounds and is divided into 13 equal pieces, how
much does each piece weight?
A: Each piece of the cake weighs 3.77/13=0.29 pounds. So, the answer is
0.29 pounds.

Q: [QUESTION]
A:
```

Prompt for ReAct:

```
Answer the following question with <add>, <subtract>, <multiply>,
<divide>, <power>, <sqrt>, <log>, <lcm>, <gcd>, <ln>, <choose>,
<remainder>, <permutate>:

Q: If Amy’s income increases by 4% annually, how many times will it
multiply in 11 years?
A: In 11 years, Amy’s income will increase by 1.04^11 =
<power>(1.04,11)=1.54 times. So, the answer is 1.54.

Q: If a store sells 147 bananas today and 354 more bananas tomorrow, how
many bananas does the store sell in total?
A: The store sells 147 bananas today and 354 more bananas tomorrow, so the
total number of bananas sold is 147+354=<add>(147,354)=501. So, the answer
is 501.

Q: A man had 789.4 dollars in his wallet. He spent 11.99 dollars on a
movie ticket. How much money does he have left now?
A: The man had 789.4 dollars in his wallet and spent 11.99 dollars on a
movie ticket, so he has 789.4-11.99=<subtract>(789.4,11.99)=777.41 dollars
left. So, the answer is 777.41.

Q: If a cake weighs 3.77 pounds and is divided into 13 equal pieces, how
much does each piece weight?
A: Each piece of the cake weighs 3.77/13=<divide>(3.77,13)=0.29 pounds.
So, the answer is 0.29.

Q: [QUESTION]
```

```
A:
```

A.2.4 Prompt for FuncQA Multi-Hop

Prompt for Zero-Shot with ChatGPT:

```
Solve the following math problem step by step, and then provide the final
answer in the format: 'So, the answer is xxx.'

[QUESTION]
```

Prompt for Chain of Thought (CoT) and ToolkenGPT:

```
Answer the following questions step by step:

Question: A coin is tossed 8 times, what is the probability of getting
exactly 7 heads ?
Answer: The total number of possible outcomes to toss a coin 8 times is
2^8=256. The number of ways of getting exactly 7 heads is 8C7=8. The
probability of getting exactly 7 heads is 8/256=0.03125. #### 0.03125

Question: If paint costs $3.2 per quart, and a quart covers 12 square
feet, how much will it cost to paint the outside of a cube 10 feet on each
edge?
Answer: The total surface area of the 10 ft cube is 6*10^2=6*100=600
square feet. The number of quarts needed is 600/12=50. The cost is
50*3.2=160. #### 160

Question: log(x)=2, log(y)=0.1, what is the value of log(x-y) ?
Answer: log(x)=2, so x=10^2=100; log(y)=0.1, so y=10^0.1=1.26;
x-y=100-1.26=98.74, so log(x-y)=log(98.74)=1.99. #### 1.99

Question: How many degrees does the hour hand travel when the clock goes
246 minutes?
Answer: The hour hand travels 360 degrees in 12 hours, so every hour it
travels 360/12=30 degrees. 246 minutes is 246/60=4.1 hours. The hour hand
travels 4.1*30=123 degrees. #### 123

Question: [QUESTION]
Answer:
```

Prompt for ReAct:

```
Answer the following questions with <add>, <subtract>, <multiply>,
<divide>, <power>, <sqrt>, <log>, <lcm>, <gcd>, <ln>, <choose>,
<remainder>, and <permutate>:

Question: A coin is tossed 8 times, what is the probability of getting
exactly 7 heads?
Answer: The total number of possible outcomes to toss a coin 8 times is
2^8=<power>(2,8)=256. The number of ways of getting exactly 7 heads is
8C7=<choose>(8,7)=8. The probability of getting exactly 7 heads is
8/256=<divide>(8,256)=0.03125. #### 0.03125

Question: If paint costs $3.2 per quart, and a quart covers 12 square
feet, how much will it cost to paint the outside of a cube 10 feet on each
edge?
Answer: The total surface area of the 10 ft cube is
6*10^2=6*<power>(10,2)=100=<multiply>(6,100)=600 square feet. The number
of quarts needed is 600/12=<divide>(600,12)=50. The cost is
50*3.2=<multiply>(50,3.2)=160. #### 160
```

```
Question: log(x)=2, log(y)=0.1, what is the value of log(x-y) ?
Answer: log(x)=2, so x=10^2=<power>(10,2)=100; log(y)=0.1, so
y=10^0.1=<power>(10,0.1)=1.26; x-y=100-1.26=<subtract>(10,1.26)=98.74, so
log(x-y)=log(98.74)=<log>(98.74)=1.99. #### 1.99

Question: How many degrees does the hour hand travel when the clock goes
246 minutes?
Answer: The hour hand travels 360 degrees in 12 hours, so every hour it
travels 360/12=<divide>(360,12)=30 degrees. 246 minutes is
246/60=<divide>(246,60)=4.1 hours. The hour hand travels
4.1*30=<multiply>(4.1,30)=123 degrees. #### 123

Question: [QUESTION]
Answer:
```

B Details of Knowledge-based QA

For KAMEL, we first transform each wikidata relation identifier (e.g. `P1346`) into a natural language description. Note that the natural language descriptions are not necessary for ToolkenGPT which is directly trained with demonstrations, but they are crucial for the in-context learning baselines, especially for ICL (desc), which can only understand tools through descriptions. We then describe the process of synthesizing training data.

B.1 Getting Text Description

KAMEL provides a question template for each relations. We randomly sample 3 facts from the dataset and instantiate them into question-answer pair, and use the following prompt to generate a description for them with ChatGPT:

```
Given a question template and some example answer, you need to define an
API that can help you answer the question.
Q 1.1: What is the original language of The Wonderful Galaxy of Oz
A 1.1: Japanese
Q 1.2: What is the original language of Wild Field?
A 1.2: Russian
Q 1.3: What is the original language of Nadigan?
A 1.3: Tamil
API 1: original_language(title): gets the original language of an art work
Q 2.1: What languages does Judah Maccabee speak?
A 2.1: Hebrew
Q 2.2: What languages does Ronelda Kamfer speak?
A 2.2: Afrikaans
Q 2.3: What languages does Leibush Lehrer speak?
A 2.3: Yiddish
API 2: spoken_languages(name): gets the spoken languages of a person
Q 3.1: [Q1]
A 3.1: [A1]
Q 3.2: [Q2]
A 3.2: [A2]
Q 3.3: [Q3]
A 3.3: [A3]
API 3:
```

B.2 Synthetic Data

We use two prompts to synthesize diverse training data, and aggregate the samples from each prompt.

```
Here are some examples of using functions for text generation (after the
function call, the sentence should continue with the returned value of the
function call):
1. star_rating(product): gets the star rating of the product on a scale
from 0 to 5.
Example 1.1: The new iPhone 12 Pro Max is already generating a lot of
buzz, thanks to its <f>star_rating("iPhone 12 Pro Max")="4.7"</f>4.7 star
rating.
2. literary_genre(book): gets the literary genre of a book
Example 2.1: Literature is often categorized by genre, such as drama,
romance, or science fiction. The Harry Potter series is a popular example
of the <f>literary_genre("Harry Potter")="fantasy"</f>fantasy genre, which
features imaginary worlds and magical elements.
3. current_location(user_id): gets the current location of a user.
Example 3.1: If you're trying to coordinate plans with a friend, it's
helpful to know their current location. You can ask the question "Where
are you right now?" and use the function <f>current_location("1234")="New
York"</f>New York as an example response.
4. number_of_movies(director): gets the number of movies directed by a
specific director.
Example 4.1: Martin Scorsese is one of the most celebrated movie directors
of all time. He has directed a total of <f>number_of_movies("Martin
Scorsese")="78"</f>78 movies throughout his career.
5. word_definition(word): gets the definition of a particular word
Example 5.1: Writers and English language learners can enhance their
vocabulary by knowing the definition of unfamiliar words. The definition
of the word "eccentric" is <f>word_definition("eccentric")="unconventional
and slightly strange"</f>unconventional and slightly strange.
6. number_of_spotify_followers(artist): gets the number of Spotify
followers for the artist.
Example 6.1: Taylor Swift's latest album is a hit and her fan base is
growing rapidly. In fact, her number of Spotify followers as of today is
<f>number_of_spotify_followers("Taylor Swift")="49,879,220"</f>49,879,220.
6. [DESCRIPTION]
Please continue to generate 10 examples using the function [NAME],
starting with 6.1 to 6.10.
```

```
Here are some examples of using functions for text generation (after the
function call, the sentence should continue with the returned value of the
function call):
1. current_weather(city): gets the current weather of a city.
Example 1.1: What's the weather in Beijing now? The weather is
<f>current_weather("Beijing")="sunny"</f>sunny now. Example 1.2: Do you
know what's the weather in San Diego now? The weather is
<f>current_weather("San Diego")="cloudy"</f>cloudy now.
2. calculator(formula): gets the calculation result of a formula.
Example 2.1: What's sum of 213 and 5032? The answer is
<f>calculator("213+5032")="5245"</f>5245.
Example 2.2: What's difference between 2015 and 33? The answer is
<f>calculator("2015-33")="1982"</f>1982.
3. [DESCRIPTION]
Please continue to generate 10 examples using the function [NAME],
starting with 3.1 to 3.10.
```

B.3 Training Details

Toolken embeddings are trained with a learning rate of 1e-4, performing early stopping based on the development set, and trained for a maximum of 5 epochs.

C Details of Embodied AI

C.1 Preprocessing

We collect all scripts from ActivityPrograms [21] and filter the dataset with the following steps: (1) filter out all the scripts that are not executable, or don't cause any state changes in VirtualHome (2) deduplicate the scripts with the same goal and instruction. (3) discard the script that involves two different objects of the same name (4) find the verbs and objects that occur more than 10 times in the data, and keep the scripts composed of only these verbs and objects.

Note that our preprocessing is different from Huang et al. [21], where they regard a high-level goal as a task. We treat two scripts with the same goal but different instructions as distinct tasks because different instructions often indicate different action sequences, which may lead to different final state graphs, e.g., for a high-level goal "Reading", some of the instructions mention "Turn on desk lamp" while others don't. Huang et al. [21] relies on human annotation to evaluate the correctness of the generated script, which actually lowers the difficulties of learning the environment, because humans may assign a correct label as long as the plan looks "reasonable". On the contrary, we can use the Evolving Graph [3] to strictly match the resulting state and ground truth state. This serves as an automatic and more objective evaluation.

C.2 Prompts

We show the prompts for LLMs to generate plans below. Note that all methods use the same prompts in this experiment.

```
I am a household robot and I can take actions from '[FIND]', '[SIT]',
'[SWITCHON]', '[TURNTO]', '[LOOKAT]', '[TYPE]', '[WALK]', '[LIE]',
'[GRAB]', '[READ]', '[WATCH]', '[POINTAT]', '[TOUCH]', '[SWITCHOFF]',
'[OPEN]', '[PUSH]', '[PUTOBJBACK]', '[CLOSE]', '[DRINK]', '[RUN]',
'[DROP]', '[PULL]'.

Task 1:
I am in ['bathroom']. The objects I can manipulate are ['faucet',
'keyboard', 'television', 'coffe_maker', 'chair', 'button', 'pillow',
'phone', 'cup', 'couch', 'freezer', 'desk', 'oven', 'light', 'table',
'bedroom', 'dining_room', 'cupboard', 'computer', 'sink', 'mail', 'bed',
'mouse', 'home_office'].
Goal:
Write an email
Hint:
i went near the computer and turned it on. then sent the mail
Plan:
[WALK] <home_office>
[WALK] <table>
[FIND] <table>
[WALK] <table>
[FIND] <computer>
[TURNTO] <computer>
[LOOKAT] <computer>
[TURNTO] <computer>
[SWITCHON] <computer>
[FIND] <mail>
[TURNTO] <mail>

Task 2:
I am in ['home_office']. The objects I can manipulate are ['faucet',
'novel', 'keyboard', 'television', 'newspaper', 'chair', 'coffe_maker',
'pillow', 'phone', 'check', 'couch', 'freezer', 'desk', 'toothbrush',
```

[3] https://github.com/xavierpuigf/virtualhome

```
'oven', 'light', 'food_food', 'table', 'bookmark', 'bedroom',
'dining_room', 'computer', 'sink', 'mail', 'bed', 'cat', 'mouse',
'home_office', 'pot'].
Goal:
Work
Hint:
Find the computer. Turn it on by pressing the on button. Wait for it to
load. Use the mouse and keyboard to perform your tasks on screen.
Plan:
[FIND] <computer>
[SWITCHON] <computer>
[FIND] <mouse>
[TOUCH] <mouse>
[FIND] <keyboard>
[TOUCH] <keyboard>

Task 3:
I am in ['bathroom']. The objects I can manipulate are ['dishwasher',
'faucet', 'keyboard', 'television', 'newspaper', 'chair', 'coffe_maker',
'pillow', 'phone', 'cup', 'check', 'couch', 'freezer', 'desk', 'oven',
'light', 'food_food', 'plate', 'table', 'bookmark', 'bedroom',
'dining_room', 'cupboard', 'computer', 'sink', 'bed', 'cat', 'mouse',
'home_office', 'pot'].
Goal:
Pick up phone
Hint:
first when i hear the ringing sound i will run to my living room and picks
up and i will say hello
Plan:
[RUN] <home_office>
[WALK] <chair>
[FIND] <chair>
[SIT] <chair>
[FIND] <phone>
[GRAB] <phone>

Task 4:
[QUESTION]
```

C.3 Training Details

Toolken embeddings are trained with a learning rate of 1e-4, performing early stopping based on the development set, with a maximum of 10 epochs.

D Computational Resources

In terms of computational resources, we train and test ToolkenGPT based on LLaMA-13B and LLaMA-30B using 2 and 4 Nvidia A3090 GPUs, respectively.

E The License of the Assets

In this study, we would like to emphasize that all data used are exclusively from official public sources. We ensure strict compliance with ethical guidelines, data access rights, and intellectual property regulations. We adhere to the license agreements and terms of use associated with each dataset. In addition, we acknowledge and thank the original creators and providers of the datasets for their valuable contributions to the research community.

F Safeguard Statement

In this paper, we primarily focus on the applications of mathematical, knowledge-based, and embodied planning problems, posing no significant ethical or harmful concerns. We recognize that future research on border applications of tool learning may pose a risk of misuse, and we recommend careful considerations of all aspects of safety before it's applied in the real world.