

m&m's: A Benchmark to Evaluate Tool-Use for *multi-step multi-modal* Tasks

Zixian Ma¹, Weikai Huang¹, Jieyu Zhang¹, Tanmay Gupta², Ranjay Krishna^{1,2}

¹ University of Washington

² Allen Institute of Artificial Intelligence

Abstract. Real-world multi-modal problems are rarely solved by a single machine learning model, and often require multi-step computational plans that involve stitching several models. Tool-augmented LLMs hold tremendous promise for automating the generation of such computational plans. However, the lack of standardized benchmarks for evaluating LLMs as planners for multi-step multi-modal tasks has prevented a systematic study of planner design decisions. Should LLMs generate a full plan in a single shot or step-by-step? Should they invoke tools directly with Python code or through structured data formats like JSON? Does feedback improve planning? To answer these questions and more, we introduce *m&m*'s: a benchmark containing 4K+ *multi-step multi-modal* tasks involving 33 tools that include multi-modal models, (free) public APIs, and image processing modules. For each of these task queries, we provide automatically generated plans using this realistic toolset. We further provide a high-quality subset of 1,565 task plans that are human-verified and correctly executable. With *m&m*'s, we evaluate 10 popular LLMs with 2 planning strategies (multi-step vs. step-by-step planning), 2 plan formats (JSON vs. code), and 3 types of feedback (parsing/verification/execution). Finally, we summarize takeaways from our extensive experiments and provide practical recommendations for designing planners for multi-step multi-modal tasks. Our dataset and evaluation code are available on HuggingFace³ and Github⁴ respectively.

1 Introduction

Planning agents—powered by large language models (LLMs)—are becoming increasingly proficient at decomposing user-specified tasks into a series of subtasks, where each subtask is executed by invoking tools. For example, if a user wants to learn about a new piece of clothing in an image, the LLM planner can create a plan with multiple steps: first, it invokes an image classification tool to identify the clothing as a “kimono”; once identified, it can issue a Wikipedia search query to explain the cultural relevance of the kimono (Figure 1 first row).

LLM planning agents typically consist of an LLM and a set of tools to plan over. Given an LLM and toolset, the design space of planning agents is extremely

³ <https://huggingface.co/datasets/zixianma/mms>

⁴ <https://github.com/RAIVNLab/mms>

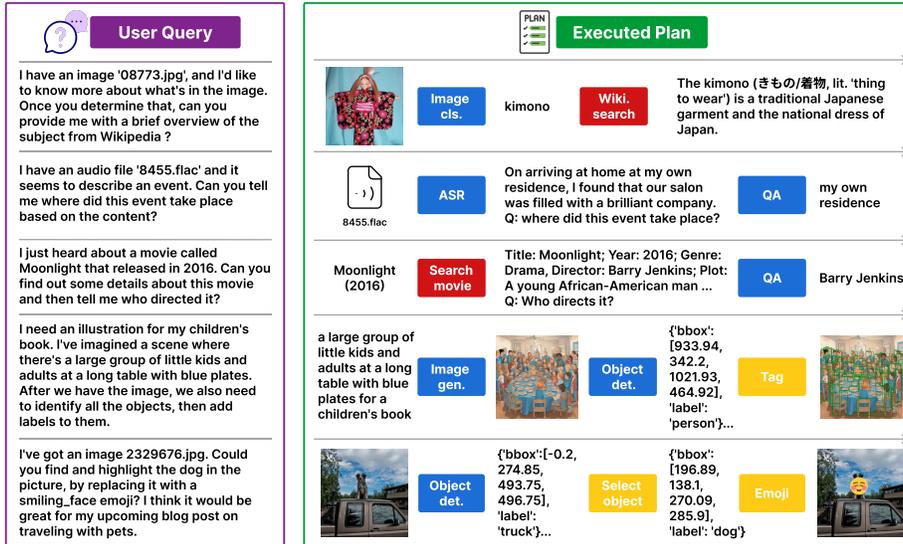


Fig. 1: We present examples of query-plan pairs along with the execution results of the plans in *m&m*'s. Our benchmark contains a large quantity of diverse user queries involving three modalities (i.e. text, image, and audio) as well as human-verified plans that consist of 1 - 3 tools across three categories: **multi-modal machine learning models** (blue), **public APIs** (red) and **image processing modules** (yellow).

rich, involving many decisions such as *planning strategy* (e.g. generation of the whole plan vs one step of the plan at a time), *forms of feedback* (e.g. no feedback or parsing/verification/execution feedback), and *plan format* (e.g. JSON strings that specify tools and arguments vs free-form Python code).

Unfortunately, there is no existing planning benchmark that supports evaluation along this combinatorially rich design space with a realistic set of multimodal tools. Recent concurrent benchmarks such as ToolEmu and TaskBench [21, 24] provide user queries and ground truth plans but lack support for realistic plan execution. For instance, TaskBench assumes that a list of tools is available for planning without providing implementation of them. TaskBench also does not instantiate the queries with actual inputs and uses placeholder input filenames like “example.png” that do not exist. ToolEmu likewise uses LLMs to emulate tool execution instead of providing tool implementations. The lack of actual implementations of tools and real execution feedback while planning makes the study of the design space elucidated above unrealistic at best, if not impossible.

Motivated by this dire need for a standardized benchmark for studying the design space of multi-step multi-modal planning agents, we first propose the *m&m*'s benchmark. *m&m*'s contains 4K+ realistic user tasks and automatically generated task plans. 1565 of these task plans are human-verified and executable with 33 curated tools consisting of multi-modal models, (free) public APIs, and image processing modules.

Table 1: Compared to previous tool planning benchmarks, *m&m*'s contains multi-modal queries that are more realistic and executable. *: MetaTool only considers Open AI plugins as tools. #: The queries of TaskBench contain textural placeholder of other modality data such as images, while queries of *m&m*'s come with real images.

		ToolBench	ToolEmu	TaskBench	MetaTool	<i>m&m</i> 's
		[18]	[21]	[24]	[9]	(ours)
Query	Real multi-modal inputs?	✗	✗	✗ [#]	✗	✓
	Verified by human?	✗	✓	✓	✓	✓
Tools	Are all tools executable?	✓	✗	✗	✓	✓
	Multi-modal models	✗	✗	✓	✗*	✓
Plan	Format	JSON	JSON	JSON	JSON	JSON/Code
Scale	Number of unique tools	3,451	36	103	390	33
	Number of queries	126k	144	17K	20k	1.5k

Next, we use *m&m*'s to systematically study the impact of 2 planning strategies (step-by-step and multi-step), 3 kinds of feedback (parsing, verification and execution), and 2 plan formats (JSON and code). Through extensive experimentation with 10 LLMs – 5 popular open-source LLMs, 2 code LLMs and 3 proprietary LLMs of varying sizes – we provide a series of findings:

First, existing LLMs instructed to perform multi-step planning consistently outperform step-by-step planning on *m&m*'s tasks, although the performance gap is smaller with more capable larger models such as Llama-3-70B and GPT-4. Second, verification and execution feedback improve LLMs' ability to generate overall executable plans and predict the correct argument names but don't necessarily improve their tool selection ability. We also observe a smaller improvement from verification/execution feedback on larger models such as Llama-3-70B and GPT-4, which already obtain relatively high scores with only parsing feedback. Third, LLMs perform comparably on tool prediction with JSON-format and Python code generation, but most models produce more executable plans with JSON-format generation. Nonetheless, this gap in executability is smaller for code LLMs such as CodeLlama-34B and 70B. Taken together, our experiments suggest that for multi-step multi-modal tasks, multi-step planning in JSON with feedback can result in the best overall tool-use performance compared to step-by-step planning, code generation, or the same setup without feedback.

2 Related work

We situate our work amongst the ever-growing number of tool-use research. **Planning evaluations.** Although many tool-use variants have been proposed, evaluating LLMs on tool-use still lacks a standardized protocol. For instance, VisProg and ViperGPT evaluate their plan's *executions* on vision tasks using a Python-like *code* format [7, 28]. HuggingGPT evaluates only the *plan* accuracy (did the agent choose the right tools) without executing the proposed plans [23]. ToolFormer [22] and ToolLLaMA [18] both use *natural language* instead of *code*

to interface with tools; while ToolFormer generates a *multi-step* plan all at once and evaluates the program’s *execution*, ToolLLaMA generates the plan *step-by-step*, with *self-feedback* to correct mistakes. ToolLLaMA evaluates only the *plans* while ToolFormer evaluates both *plans* and executions. Unfortunately, no single benchmark evaluates planning agents along this combinatorial design space, which is what we contribute.

Tool-use benchmarks. Today, tool-use evaluation is spread out across a number of diverse benchmarks, including HotpotQA, WebShop, GQA, RefCOCO, and NLVR [10, 11, 26, 35, 36]. None of these contains ground truth plans, conflating planning errors with execution error. In other words, it is hard to separate whether an LLM failed to propose the correct plan or whether one of the tools used in the plan failed. In response, recent concurrent efforts have proposed new benchmarks, such as ToolEmu, TaskBench, and GAIA [15, 21, 24]. They do contain ground truth plans but fail to support evaluating plans’ execution results.

Planning strategies. There are multiple strategies for planning. For instance, Psychology literature reveals that people rarely plan tasks in their entirety due to the cognitive cost of planning long-range tasks [3]. Instead, they plan the first couple of subtasks, and execute them before planning the rest [1, 3]. In the tool-use literature, we identify two primary forms of planning strategies: *step-by-step planning* [5, 18, 38] and *multi-step planning* [7, 23, 28]. Similar to people, step-by-step planning generates plans sequentially with one subtask at a time. By contrast, multi-step planning creates the entire plan before executing any subtask. Unfortunately, these two strategies have not been systematically compared; we systematically compare both across multiple open-source and close-source LLMs.

Feedback mechanisms. LLM planners make mistakes, stitching together tools that fail to execute or worse, fail to compile. Although human feedback is one mechanism to align plans with human expectations and preferences [2, 31], they require real users, making evaluation stochastic. However, there have been several automatic mechanisms that can improve plans [30, 39]. For instance, syntactic mistakes can easily be detected using external *verifiers* and can guide planners to iterate on their plans [8, 14, 16, 25]. Others require examining the output of individual subtask *executions* [20, 27, 29, 38, 41]. In this work, we compare plan parsing/verification feedback as well as tool execution feedback.

3 *m&m*’s: the benchmark

To facilitate the study of LLM planners for *multi-step multi-modal* tasks, we curate the *m&m*’s benchmark. Before describing the dataset generation process, we first formalize the tool-planning problem in Sec 3.1. We then describe our benchmark creation process in Sec. 3.2 and present dataset statistics in Sec. 3.3.

3.1 Formalizing multi-step multi-modal tool-use

Given a tool set \mathcal{T} , and the user query \mathcal{Q} , a planner is required to produce a plan \mathcal{P} that consists of a sequence of tool invocations $\mathcal{P} = [t_1(\{a_1^k = v_1^k\}_k), t_2(\{a_2^k =$

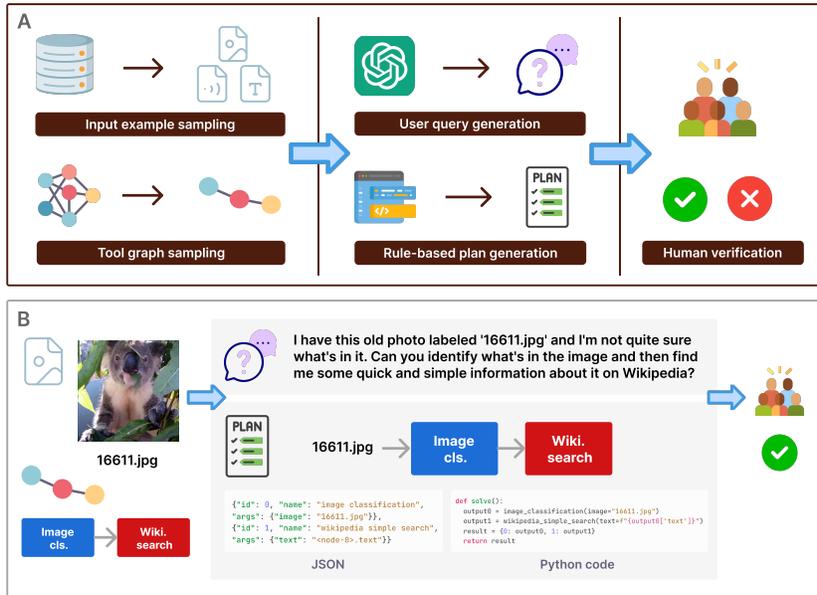


Fig. 2: Data generation pipeline. As shown in panel A, our generation process consists of five main steps: input example sampling, tool graph sampling, user query generation with GPT-4, rule-based plan generation, and human verification. Panel B showcases an instantiation of this process with a real example.

$v_2^k\}_k), \dots, t_m(\{a_m^k = v_m^k\}_k)]$, where t_j represents the j^{th} tool in the plan, and a_j^k , and v_j^k represent tool t_j 's k^{th} argument name and value respectively. Note that the output of t_j may be used as argument values for subsequent tools $t_{j+1:m}$. m&ms contains a set of N query-plan pairs, i.e., $\{(Q_i, P_i)\}_{i \in [N]}$ with each plan composed of executable tools chosen from a curated set of API-calling functions, multi-modal models, and image processing modules.

3.2 Dataset generation

To create such a dataset, our data generation process consists of five major steps: ① We construct a valid tool graph with all the tools and sample a subgraph from it as a sequence of tools in the target plan; ② To instantiate the tool plan with inputs, we pair plans with real inputs (e.g., images) from existing datasets based on the first tool in the plan; ③ To generate the user query, we prompt GPT-4 with the tool graph and input pair to generate realistic user queries with few-shot demonstrations; ④ We transform the tool graph and input pair to a fully specified JSON plan (tool names with corresponding argument names and values) with a rule-based program. Additionally, we map the

JSON-format plans to Python code to support code generation evaluation; ⑤ Finally, three human annotators verify whether each plan can solve the corresponding user query. To obtain the final tool-balanced human-verified subset, we discard some plans from the initial human-verified set to avoid an overwhelming representation of any tool.

① **Tool graph sampling.** We first create a directed graph with all 33 tools as the nodes and edges denoting valid connections between nodes. A connection is valid only if the output of the source tool matches the expected input type of the target tool. For example, there is an edge between `image classification` and `wikipedia simple search`, because the output of `image classification` - a text label - is a valid input type for `wikipedia simple search`. We then sample subgraphs from the full tool graph to obtain tool sequences with valid tool dependencies.

② **Input example sampling.** We now need to instantiate queries with real input examples. To do so, we first collect real-world examples from the validation sets of 11 existing datasets, including ImageNet [4], SQUAD [19], Visual Genome [12], MagicBrush [40], librispeech [17]. Then, to pair a tool graph sampled in the previous step with an input, we randomly sample an input based on the input type needed for the first tool in the graph. For example, if the first tool in a tool sequence is `image classification`, we randomly sample an image (e.g. “16611.jpg”) from ImageNet as the input.

③ **Query generation.** With a set of tool sequences and input examples to the first tools, we prompt GPT-4 to generate realistic user queries. Concretely, we randomly sample five different input examples for each tool sequence and ask GPT-4 to generate two queries for each tool sequence with the same input (See Appendix for the full prompt).

④ **Plan generation.** For plan generation, we write a rule-based program to generate a plan (i.e. an ordered list of tool names with corresponding argument names and values fully specified) for each query. Each step in the plan contains an id, tool name, and an argument dictionary with this tool’s argument names as the keys and argument values as values. We populate each node’s ID and name based on the sampled tool sequence and fill in the argument names for each tool using a pre-defined metadata document. We also fill in the argument values of the first tool using the input examples and those of subsequent tools using a special notation $\langle node - id \rangle.key$, where id refers to the id of the previous node and key refers to the output key. To further refine the plans to be even more faithful to queries, we rewrite the argument values of `text generation` and `image generation` (e.g. from “a shark” to “a child-friendly book cover image of a shark”) by prompting GPT-4 with the queries and original plans.

⑤ **Human verification** Finally, we perform extensive human verification on all 4427 generated query-plan pairs. We ask three expert annotators (who are undergraduate and Ph.D. students in CS) to rate each query-plan pair with 0 or 1, where 1 indicates that the plan can resolve the query perfectly. We obtain a subset of 1500+ examples on which all three annotators rate 1 and perform further filtering of examples where the plan contains much more frequent tools (e.g.

Table 2: The statistics of the *m&m*’s dataset.

Item	Number
Raw examples	4427
Human verified examples	1565
Human verified & balanced examples	882
- 1 / 2 / 3-tool examples	70 / 159 / 653
Tools	33
- ML model / image processing / API	13 / 11 / 9
Tool graphs	317
Avg. # of unique queries per tool graph	2.78

image generation and text generation) to balance the overall distribution of tools (See Appendix for more details on filtering and the tool distribution).

It is worth noting that two of the steps in our dataset generation pipeline draw similarities with the recently released concurrent TaskBench [24]. Similar to them, we also sample a subgraph of tools and query generation steps. However, we want to highlight two major differences: first, we leverage real-world examples as inputs to the tool sequences (in contrast to TaskBench’s “example.jpg”, “example.wav” etc.), which not only leads to a more realistic instantiation of queries but also enables plan execution on actual input which is crucial for studying execution feedback in planning agents. Second, we use a rule-based program instead of GPT-4 to obtain the ground truth plans based on the sampled tool sequences, which eliminates the possibility of hallucinated and incorrect plans.

3.3 Dataset quantity and quality

Overall, *m&m*’s contains a large quantity of **diverse ecologically-valid task** queries (see Figure 1). Each task is associated with **human-verified** and executable plans (Table 2). Concretely, there are a total of 4427 raw examples in *m&m*’s, where 1565 have been verified to be correct by three human annotators. After additional filtering for a **balanced tool distribution** (See Appendix for more details), we select a subset of 882 examples for our evaluation. Tasks are **granular in difficulty** with 70 queries that require a single tool, 159 need two tools, and 653 need three tools. In terms of tools, there are **33 unique tools in total across three different categories**, of which 13 are multi-modal machine learning models on HuggingFace, 11 are image processing modules from VisProg [7], and 9 are free public APIs from RapidAPI⁵. Our final dataset includes 317 representative tool graphs, where each graph has multiple queries. See more examples in the Appendix.

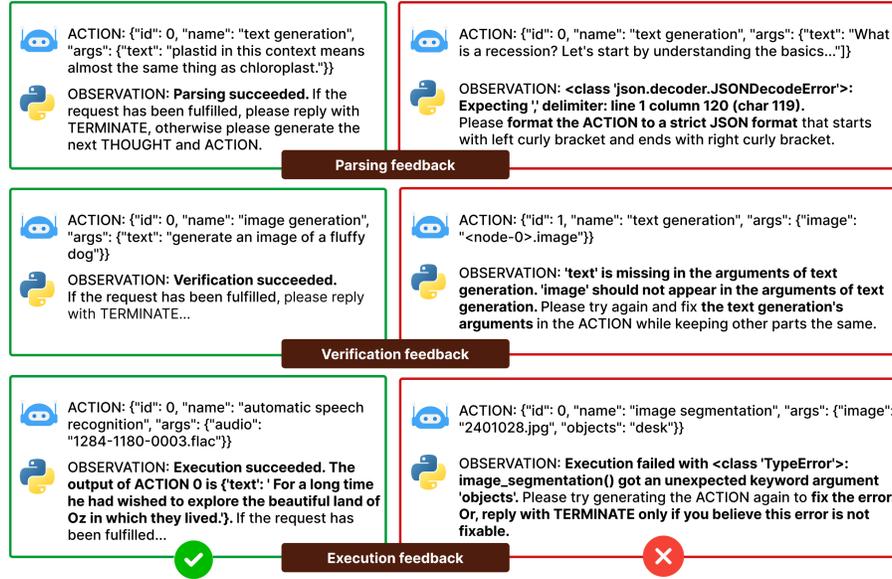


Fig. 3: Types of feedback. We present examples of parsing, verification, and execution feedback in both success and failure cases.

4 Planning agent

To systematically evaluate the design space of planning agents, we design a modular planning system with these components: planning LLM, parser, verifier, and executor. We implement this system with AutoGen’s framework [33]. Given the user query, the LLM must iteratively generate and refine the plan. Each iteration involves generating the whole or a part of the plan and receiving feedback on the generation. Given the raw text output from the LLM at the current iteration, the *m&m*’s agent supports the following 3 kinds of feedback -

Parsing feedback. The parser attempts to parse the LLM text output to either JSON or code formats and returns an error message in case of parsing failures.

Plan verification feedback. The verifier checks the parsed output according to pre-defined rules and returns an error message in case of rule violations. Specifically, the verifier checks if the predicted tool exists in our provided tool list, if it forms a valid connection with the previous tool, and if the predicted argument names match the ones specified in the metadata document.

Plan execution feedback. In the case of JSON output, the executor calls the functions with specified arguments in a Python environment and returns

⁵ <https://rapidapi.com/hub>

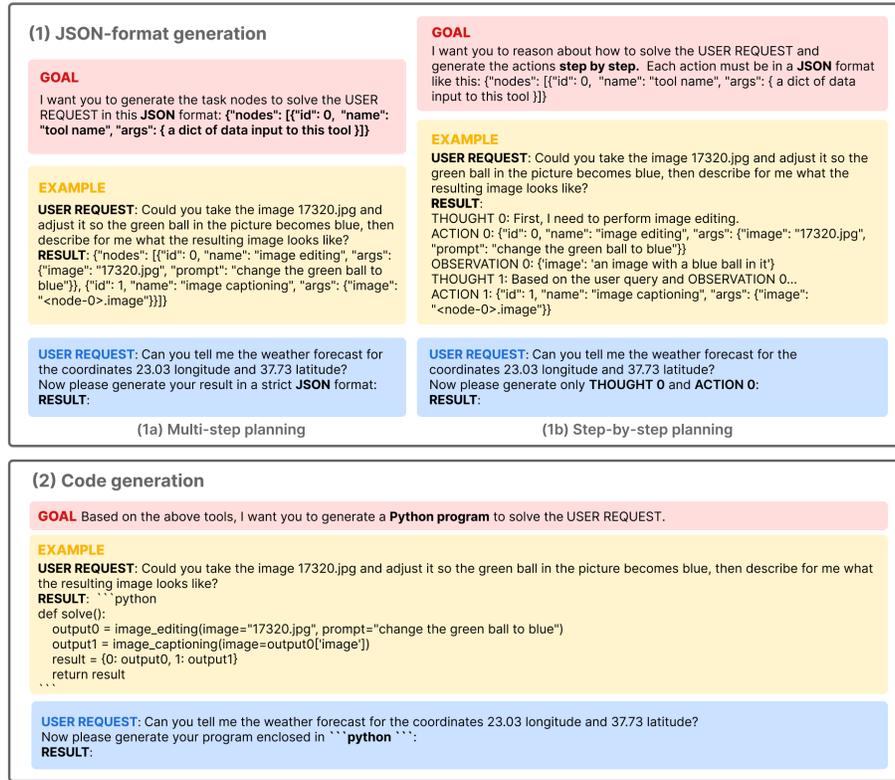


Fig. 4: Illustrating the three main planning setups in our evaluation: (1a) multi-step and (1b) step-by-step JSON-format language generation [38], and (2) code generation. (See the Appendix for the full prompts).

the output or execution errors. In the case of code output, the code is directly executed with outputs or errors returned as feedback.

5 Experiment

Using our benchmark with a flexible agent design, we experiment with 10 instruction-tuned large language models of varying sizes (7 open-source and 3 proprietary) across different planning setups. We describe these evaluation setups in Sec. 5.1, metrics in Sec. 5.2, and our experimental findings in Sec. 5.3.

5.1 Setup

We establish a unified framework to categorize LLMs' task planning setups along the three axes below. **Planning strategy:** Prior works formulate task

planning as either step-by-step or multi-step planning [18, 23, 38]. Step-by-step planning refers to the setup where a language model is instructed to predict only one action at a time (Figure 4 (1b)). On the other hand, in the setting of multi-step planning, a model can predict multiple actions at once (Figure 4 (1a)). **Plan format:** Additionally, existing works have also adopted different plan formats for tool use: often as code, pseudo-code, or predefined structured representations such as JSON [7, 23, 28]. In this work, we primarily focus on two of the common plan formats: JSON and code. **Feedback:** We experiment with three kinds of feedback - feedback from parsers, rule-based verifiers, and execution modules (Figure 3). Nevertheless, our benchmark can be used to study other types of feedback, such as self-feedback [13], which we leave to future work.

5.2 Evaluation metrics

To holistically evaluate planning agents’ performance on our benchmark, we adopt three main metrics: **tool-F1**, **argname-F1**, and **pass rate**. **Tool-F1** is defined as the F1 score of tool name prediction, where we treat each predicted tool name as one example and compare the set of predicted tool names to the groundtruth set of tools in each plan. Similarly, **argname-F1** is defined as the F1 score of argument name prediction for each tool, where we consider each (tool name, argument names) tuple as one example. Our implementation turns each tuple into a string and compares the set of predicted “tool name-argument names” strings to the labels. **Pass rate** is the percentage of predictions that execute successfully without any execution errors. It measures the executability but not the correctness of the predicted plans. We choose these three metrics because they assess two important aspects of planning and tool use: tool selection and tool invocation. A higher tool-F1 indicates better tool selection, whereas higher argname-F1 and pass rate imply improved tool invocation. To evaluate models with the same metrics in the code generation setup, we parse the generated code into an Abstract Syntax Tree (AST) with Python’s AST module and extract the function names and argument names for calculating tool-F1 and argname-F1.

We also provide argvalue-F1 in the Appendix for completeness but caution the reader about the challenges of evaluating argument values due to surface-form or syntactic differences in the values, particularly for free-form text arguments (e.g. the prompts in **image generation** and **text generation**). We report additional results on plan evaluation metrics, including overall plan accuracy, normalized edit distance, edge-F1, code-specific metrics such as AST accuracy and CodeBLEU, and plan execution accuracy in the Appendix.

5.3 Results

We first highlight the key findings from our empirical analysis and then describe each finding in more detail:

1. All planning agents perform better on tool selection with multi-step planning than with step-by-step planning, with a smaller gap for larger models (Fig. 5)

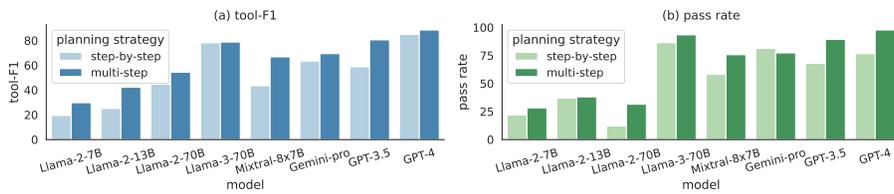


Fig. 5: Comparing planning strategies. We find that models consistently perform better on tool-F1 under multi-step prediction compared to under step-by-step prediction regardless of their sizes. Similarly, all models except for Gemini-pro achieve a higher pass rate with multi-step prediction.

Table 3: We present the tool-F1 and argname-F1 of models with various feedback, where P, V, and E represent parsing, verification, and execution feedback respectively. We use parsing feedback only (P) under multi-step planning and JSON-format language generation as the basis, while showing the Δ of those with other feedback combinations. We find that verification and execution feedback can improve models’ performance on argument name prediction and pass rate, but can hurt tool selection.

model	tool-F1				argname-F1				pass rate			
	P	PV	PE	PVE	P	PV	PE	PVE	P	PV	PE	PVE
Llama-2-7B	29.78	-2.94	-2.59	-2.58	34.03	2.03	1.24	1.15	28.23	18.14	10.32	13.72
Llama-2-13B	42.27	-3.45	-2.78	-4.57	45.07	3.94	3.08	3.29	38.10	29.93	32.99	23.92
Llama-2-70B	54.40	-0.35	-0.49	-0.03	52.49	12.87	8.97	12.60	31.52	39.80	23.13	29.59
Mixtral-8x7B	66.79	1.18	-0.11	-0.04	72.52	2.00	1.89	2.72	75.74	10.32	8.96	10.77
Gemini-pro	69.38	1.18	-0.11	-0.04	73.37	2.00	1.89	2.72	77.32	13.27	14.06	16.67
Llama-3-70B	78.73	1.54	-0.30	0.70	84.97	0.45	-0.68	-0.08	92.40	-0.45	4.31	3.29
GPT-3.5-turbo-0125	80.52	-0.65	-2.80	-2.56	84.86	0.65	-0.92	-0.86	89.46	6.69	7.26	6.92
GPT-4-0125-preview	88.46	-0.60	0.25	-0.91	89.81	-0.18	0.48	0.32	97.73	1.13	-1.25	2.15
GPT-4o-2024-05-13	89.28	-0.22	0.48	-0.21	90.32	1.24	1.00	1.24	96.37	2.61	-0.45	2.15

Note: we use the experiments with parsing feedback as the baseline to highlight external feedback’s effects on tool selection and invocation instead of parsing. We include the results of experiments with no feedback in the Appendix.

2. Verification and execution feedback can help models improve tool invocation by predicting correct argument names and generating executable plans but can lead to worse tool selection due to wrong fixes (Tab. 3 and Fig. 6)
3. While models perform comparably on tool selection with JSON versus code generation, they produce more overall executable plans with JSON-format generation (Fig. 7)

Models consistently perform better on tool-F1 and pass rate when instructed to perform multi-step planning instead of step-by-step planning. We find that all large language models achieve higher tool-F1 when they are instructed to perform multi-step planning compared to when they perform step-by-step prediction (Figure 5), although the gap is smaller for more capable larger models like Llama-3-70B and GPT-4. All models except for Gemini-pro achieve a higher pass rate with multi-step planning. Among the 10 models we evaluated, Llama-2-7B, Llama-2-13B, and GPT-3.5 all showcase a large increase ($>10\%$) in performance with multi-step planning compared to step-by-step prediction, with the greatest increase of 21.8% for GPT-3.5. Through qualitative

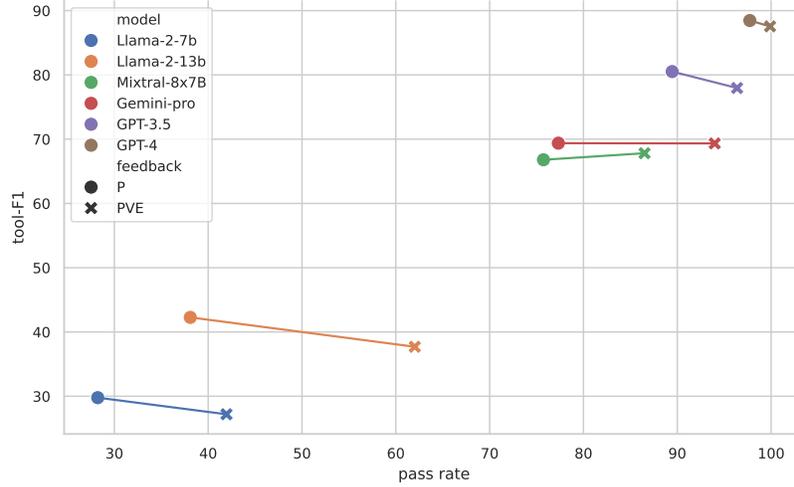


Fig. 6: Comparing tool-F1 and pass rate without vs. with feedback. We find that feedback greatly improves planning agents’ pass rates across different model sizes, especially for Llama-7B/13B and Gemini-pro. However, feedback can also harm models’ tool prediction performance and decrease their tool-F1 by a small amount ($< 5\%$).

analysis, we learn that when models are instructed to perform step-by-step prediction, they tend to “Terminate” after they receive positive feedback (e.g. “verification/execution succeeded”) from the environment, disregarding whether the user request has been fulfilled. This means that they often predict fewer steps than required and miss necessary tools to resolve the requests (See Figure 8 A).

External feedback can improve planning agents’ performance on argument name prediction and pass rate. We find that both verification and execution feedback can lead to slightly better argname-F1 and much higher pass rates (Table 3), indicating that feedback can help models predict correct argument names and generate more executable plans. With feedback, most models can increase argname-F1 by around 1-4% and pass rate by up to 20-30% (Table 3). There are only a few exceptions on GPT-3.5 and GPT-4, which already obtain relatively high performance without feedback (Table 3). Also, verification feedback can be more helpful than execution feedback on argument name prediction. In qualitative analysis, we find that this is because our verifier pinpoints where the error occurs and outputs a targeted and thus more helpful feedback message. On the other hand, the execution module returns the error message as it is, which can be vague and obscure, thus confusing the model and even resulting in wrong fixes (Figure 8 B).

While we see generally positive effects of feedback on argname-F1 and pass rate, we also observe that feedback can lead to a small decrease ($< 5\%$) in models’ tool-F1. We observe that this is mainly because models can change some

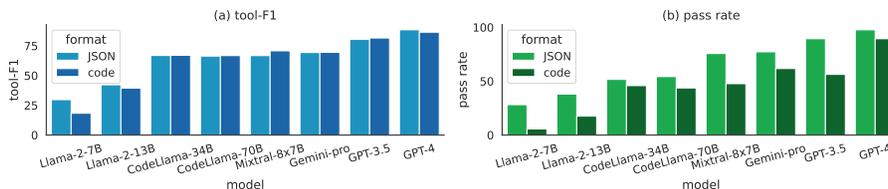


Fig. 7: Comparing plan formats. We find that all models except for Llama-7-b perform comparably on tool-F1 with JSON-format and code generation. However, JSON-format generation leads to a much higher pass rate across all models.

correct tools to the wrong ones or remove them even though the feedback instructs them to only fix the erroneous parts in the plan (Figure 8 B). One way to mitigate this error can be using more fine-grained and localized feedback [34]. Additionally, neither verification feedback nor execution feedback provides useful information on the correctness of the tool selection that can increase tool-F1. Nevertheless, we also note that the decrease in tool-F1 with feedback is a lot smaller compared to the gains in pass rate (Figure 6), which suggests feedback can greatly improve tool invocation at a small cost to tool selection.

Models perform comparably on tool-F1 with JSON-format and code generation but much worse on pass rate with code generation. We learn that plan formats can also influence models’ tool use performance (Figure 7), especially on the executability of the generated plans. Concretely, our experiments show that while all models except for Llama-2-7B achieve similar tool-F1s (<3% difference) with JSON-format generation and code generation, they all suffer from a large drop in pass rate with code generation. Upon qualitative analysis, we find that one common execution error in code generation is failing to access the output from a tool (See Figure 8 C). While the same error also happens to JSON-format generation, it occurs less frequently due to the more rigid structure of JSON. These results suggest that JSON-format generation is preferable to code generation when the executability of generated plans matters.

6 Discussion

6.1 Limitations

There are a few limitations to our work. First, *m&m*’s only considers sequential task plans, which represent a majority of real-world user requests. However, some tasks might require dynamic task plans depending on the output for one subtask [6]. Dynamic plans require a more complex tool graph sampling procedure. Second, as our main goal is to study the effects of planning formulations and feedback, we do not investigate another dimension of planning design: prompt style. We use direct and ReACT-style [38] prompting and exclude more

User request	Label	Prediction
<p>A. Missing tool(s) due to early termination</p> <p>I have an image labeled '86582-input.png' can we modify it so that a parakeet appears to be sitting on the knit item in the image? Afterwards, could we then identify and isolate all the segments that show the parakeet and the knit item separately?</p>	<p>Step-by-step / JSON / Parsing</p> <pre>{ "id": 0, "name": "image editing", "args": { "image": "86582-input.png", "prompt": "a parakeet should be sitting on the knit item" } } { "id": 1, "name": "image segmentation", "args": { "image": "<node-0>-image" } }</pre>	<p>THOUGHT 0: First, I need to perform image editing. ACTION 0: {"id": 0, "name": "image editing", "args": {"image": "86582-input.png", "prompt": "add a parakeet sitting on the knit item"}} OBSERVATION: Parsing succeeded. If the request has been fulfilled, please reply with TERMINATE, otherwise please generate the next THOUGHT and ACTION. Great, I'm glad the parsing was successful! Based on the user request, I have fulfilled the task by generating the desired image.</p>
<p>B. Change/removal of correct tools</p> <p>I'm a bit curious about two characters from my favorite movie, Jack Twist and Ennis Del Mar. Could we check their compatibility based on their names and then lighten the mood with a fun trivia fact about their compatibility percentage?</p>	<p>Multi-step / JSON / Execution</p> <pre>{ "id": 0, "name": "love calculator", "args": { "first_name": "Jack Twist", "second_name": "Ennis Del Mar" } } { "id": 1, "name": "get trivia fact", "args": { "number": "<node-0>-number" } }</pre> <p>Multi-step / JSON / Verification</p> <pre>{ "id": 0, "name": "image classification", "args": { "image": "07600.jpg" } } { "id": 1, "name": "text generation", "args": { "text": "a detailed backstory or scenario involving the identified component" } }</pre>	<pre>{ "id": 0, "name": "love calculator", "args": { "first_name": "Jack Twist", "second_name": "Ennis Del Mar" } } { "id": 1, "name": "get trivia fact", "args": { "number": "<node-0>-output" } }</pre> <p>Execution failed with <class 'KeyError'>: 'output'. Please try generating the # RESULT # again to fix the error. Or, reply with TERMINATE only if you believe this error is not fixable.</p> <pre>{ "id": 0, "name": "love calculator", "args": { "first_name": "Jack Twist", "second_name": "Ennis Del Mar" } } { "id": 1, "name": "love calculator", "args": { "first_name": "Jack Twist", "second_name": "Ennis Del Mar" } }</pre> <p>Execution succeeded. The output is {'number': '79', 'message': 'All the best!'}</p> <pre>{ "id": 0, "name": "image classification", "args": { "image": "07600.jpg" } } { "id": 1, "name": "image captioning", "args": { "image": "<node-0>-text" } }</pre> <p>image captioning should not follow image classification, because it does not make sense to feed 'the text label for the class' into image captioning, which expects the input to be 'an image to be captioned'. Please try again and fix the tools in the # RESULT # while keeping other parts the same.</p> <pre>{ "id": 0, "name": "image captioning", "args": { "image": "07600.jpg" } }</pre> <p>Verification succeeded.</p>
<p>C. Failure to access a tool's output</p> <p>I'm writing a storyline for a new Italian cookbook. Right now, I only have this line: 'The Italian way with the knife is done.' Could you help me generate a few paragraphs that could logically follow this line and then evaluate the tone of the generated text to ensure it instills positivity and excitement in the readers?</p> <p>I'm planning a vacation to Phoenix. Could you tell me what the weather forecast is likely to be for this city?</p>	<p>Multi-step / Code / Parsing</p> <pre>output0 = text_generation(texts="Generate a few paragraphs that could logically follow this line: 'The Italian way with the knife is done.'") output1 = text_classification(text=output0[0][0])</pre> <pre>lon, lat = get_location(city="Phoenix") output0 = get_weather(lon=output0[lon], lat=output0[lat])</pre>	<pre>output0 = text_generation(texts="The Italian way with the knife is done.") output1 = text_classification(text=output0)</pre> <p>Parsing succeeded.</p> <pre>lon, lat = get_location(city="Phoenix") output0 = get_weather(lon=lon, lat=lat)</pre> <p>Parsing succeeded.</p>

Fig. 8: We present examples of three common errors (A-C) in different planning setups.

sophisticated prompting strategies such as tree-of-thoughts prompting [32, 37]. Third, as some tools in our benchmark are suboptimal, generative and/or non-deterministic, we only conducted evaluation of the execution results on a limited subset (See Appendix). Finally, we have only evaluated LLM planners because of their advanced abilities and leave the evaluation of multi-modal planners to future work.

6.2 Conclusion

In conclusion, we highlight three major contributions of our work: first, we introduce a new benchmark $m\&m$'s to support comprehensive and rigorous evaluation of tool-use abilities of planning agents for multi-step multi-modal tasks. $m\&m$'s contains a large and diverse set of queries and human-verified and executable plans; second, we characterize the design space of existing tool-use methods and conducted a systematic study of 10 LLMs with different design choices, including planning formulations, plan formats and various types of feedback; finally, our experiments reveal three takeaways, suggesting that current generation of LLMs demonstrate gains in tool-planning performance on $m\&m$'s when using multi-step planning, outputting plans in JSON format, and using feedback. We hope $m\&m$'s enables further investigation into better planning formulations that incorporate richer and more diverse kinds of feedback for solving multi-step, multi-modal tasks.

Acknowledgement

This work was partially funded by a Sony grant. It was also made possible because of OpenAI’s credit grant. We also thank Zeyu Tang for his help with figures and annotations, and Jiafei Duan, Chenhao Zheng, and Dylan Bunarto for their help with data annotations.

References

1. Allen, K.R., Smith, K.A., Tenenbaum, J.B.: Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences* **117**(47), 29302–29310 (2020)
2. Chen, P.L., Chang, C.S.: Interact: Exploring the potentials of chatgpt as a cooperative agent. *arXiv preprint arXiv:2308.01552* (2023)
3. Correa, C.G., Ho, M.K., Callaway, F., Daw, N.D., Griffiths, T.L.: Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology* **19**(6), e1011087 (2023)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. *Ieee* (2009)
5. Gao, D., Ji, L., Zhou, L., Lin, K.Q., Chen, J., Fan, Z., Shou, M.Z.: Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640* (2023)
6. Grunde-McLaughlin, M., Lam, M.S., Krishna, R., Weld, D.S., Heer, J.: Designing llm chains by adapting techniques from crowdsourcing workflows. *arXiv preprint arXiv:2312.11681* (2023)
7. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training (2022)
8. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al.: Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608* (2022)
9. Huang, Y., Shi, J., Li, Y., Fan, C., Wu, S., Zhang, Q., Liu, Y., Zhou, P., Wan, Y., Gong, N.Z., et al.: Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128* (2023)
10. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6700–6709 (2019)
11. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 787–798. Association for Computational Linguistics, Doha, Qatar (Oct 2014). <https://doi.org/10.3115/v1/D14-1086>, <https://aclanthology.org/D14-1086>
12. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**, 32–73 (2017)

13. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023)
14. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* **36** (2024)
15. Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., Scialom, T.: Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983* (2023)
16. Miao, N., Teh, Y.W., Rainforth, T.: Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436* (2023)
17. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210 (2015). <https://doi.org/10.1109/ICASSP.2015.7178964>
18. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., Sun, M.: Toolllm: Facilitating large language models to master 16000+ real-world apis (2023)
19. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). <https://doi.org/10.18653/v1/D16-1264>, <https://aclanthology.org/D16-1264>
20. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: Say-plan: Grounding large language models using 3d scene graphs for scalable task planning. *arXiv preprint arXiv:2307.06135* (2023)
21. Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C.J., Hashimoto, T.: Identifying the risks of lm agents with an lm-emulated sandbox (2023)
22. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* **36** (2024)
23. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face (2023)
24. Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., Zhuang, Y.: Taskbench: Benchmarking large language models for task automation. *arXiv preprint arXiv:2311.18760* (2023)
25. Shinn, N., Labash, B., Gopinath, A.: Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366* (2023)
26. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Barzilay, R., Kan, M.Y. (eds.) *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 217–223. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). <https://doi.org/10.18653/v1/P17-2034>, <https://aclanthology.org/P17-2034>
27. Sun, H., Zhuang, Y., Kong, L., Dai, B., Zhang, C.: Adapllan: Adaptive planning from feedback with language models. *Advances in Neural Information Processing Systems* **36** (2024)

28. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. arXiv preprint arXiv:2303.08128 (2023)
29. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023)
30. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432 (2023)
31. Wang, X., Wang, Z., Liu, J., Chen, Y., Yuan, L., Peng, H., Ji, H.: Mint: Evaluating llms in multi-turn interaction with tools and language feedback. arXiv preprint arXiv:2309.10691 (2023)
32. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models (2023)
33. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., Wang, C.: Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155 (2023)
34. Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N.A., Ostendorf, M., Hajishirzi, H.: Fine-grained human feedback gives better rewards for language model training (2023)
35. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct–Nov 2018). <https://doi.org/10.18653/v1/D18-1259>, <https://aclanthology.org/D18-1259>
36. Yao, S., Chen, H., Yang, J., Narasimhan, K.: Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems* **35**, 20744–20757 (2022)
37. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models (2023)
38. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models (2023)
39. Zhang, J., Krishna, R., Awadallah, A.H., Wang, C.: Ecoassistant: Using llm assistant more affordably and accurately. arXiv preprint arXiv:2310.03046 (2023)
40. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* **36** (2024)
41. Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., et al.: Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. arXiv preprint arXiv:2305.17144 (2023)