

# TapeAgents: a Holistic Framework for Agent Development and Optimization

Anonymous authors  
Paper under double-blind review

## Abstract

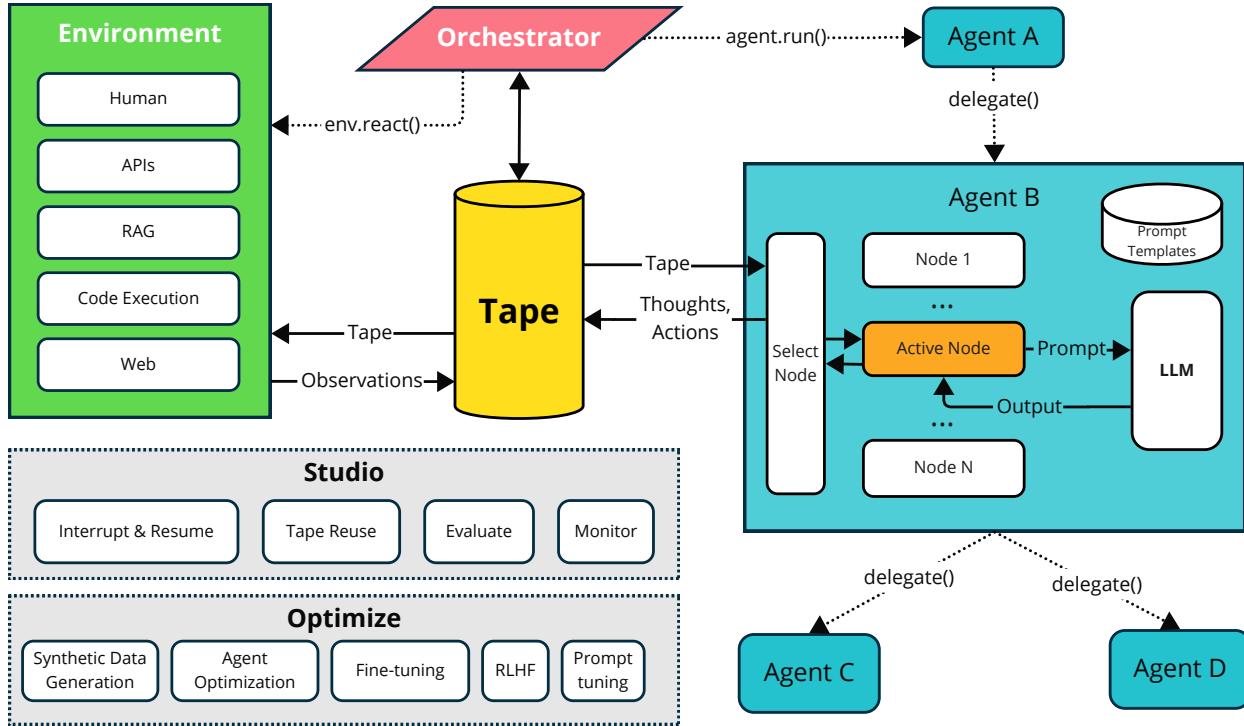
We present TapeAgents, an agent framework that leverages a structured, replayable log (**tape**) of the agent session to facilitate all stages of the LLM Agent development lifecycle. In TapeAgents, the agent reasons by processing the tape and the LLM output to produce new thought and action steps and append them to the tape. The environment then reacts to the agent’s actions by likewise appending observation steps to the tape. By virtue of this tape-centred design, TapeAgents can provide AI practitioners with holistic end-to-end support. At the development stage, tapes facilitate session persistence, agent audit, and step-by-step debugging. Post-deployment, one can reuse tapes for evaluation, fine-tuning, and prompt-tuning; crucially, one can adapt tapes from other agents or use revised historical tapes. In this report, we explain the TapeAgent design in detail. We demonstrate possible applications of TapeAgents with several concrete examples of building and optimizing agents using diverse paradigms. We present tooling prototypes and report a case study where we use TapeAgents to finetune a Llama-3.1-8B form-filling assistant to perform as well as GPT-4o while being orders of magnitude cheaper. We compare TapeAgents to prior agent frameworks and find that TapeAgents is the first one that helps the practitioner to build, debug, serve and optimize their agent.

Manuscript version: October 15, 2024

## 1 Introduction

In the coming years, we are likely to witness widespread deployments of Large Language Model (LLM) Agents: complex user-facing and background workflows that interleave traditional programming with LLM-based intelligence. This big paradigm shift in software architecture will greatly challenge AI practitioners who put LLM agents to work. The agent developers and applied scientists will have to troubleshoot and improve systems that operate in non-stationary environments and deal with non-deterministic LLM behavior and the LLM’s often fragile instruction following. For the LLM agent adoption to go smoothly and lead to good outcomes, it is crucial that agent developers and applied scientists operate in appropriate frameworks that enable effective tooling. Developers and researchers have recently proposed many agentic frameworks that support practitioners at different stages of the agent development lifecycle. Several frameworks, like LangChain (Chase, 2022), CrewAI and AutoGen (Wu et al., 2024a), help developers quickly build an agent using low-code paradigms, such as prompt-chaining or multi-agent teams. Others, like LangGraph (Chase, 2023), offer low-level support in achieving resumability, asynchronous execution, concurrency and instrumentation. At the other end of the spectrum are frameworks built by researchers like DSPy (Khattab et al., 2023a), TextGrad (Yuksekgonul et al., 2024) and Agents (Zhou et al., 2023), that usually focus on data-driven optimization of the agent performance with model finetuning and prompt-tuning algorithms, while putting less emphasis on the needs of the agent developers.

In this technical report, we present **TapeAgents** — a new holistic agent framework that supports practitioners at both the agent development and data-driven agent optimization stages. We achieve both objectives by building the framework around a comprehensive, structured, replayable, semantic-level log of the agent session that we call a **tape**, a term that also gives the framework its name (see Figure 1 for an illustration). The agents in TapeAgents read the tape to make the LLM prompt and then process the LLM output to append new **steps** to the tape: **thought** steps to express reasoning and **action** steps to request external



**Figure 1: TapeAgents at a glance.** The orchestrator alternates between running the agent and the environment who interact with each other via adding steps to the tape: a comprehensive, replayable semantic-level log. Agents are composed from basic reasoning units that we call nodes. The agents are organized in hierarchical teams with one agent being active at a time. The tape and the agent configurations are highly structured and linked with rich metadata that supports implementation of broadly usable developer tools (collectively called *Studio*) and optimization methods (collectively called *Optimize*).

inputs. The **environment** responds to the action steps at the end of the tape with **observation** steps that it likewise appends to the tape. The **orchestrator** invokes the agent and the environment in an alternate fashion and maintains full control over their interactions. By design, the orchestrator can resume from any intermediate tape, which enables session persistence and step-by-step debugging, both key developer requirements for an agent framework. For data-driven algorithms, tapes record the attribution of each step to the respective part of the agent configuration, which facilitates training, data generation and automatic prompt-tuning. Crucially, for both manual debugging and algorithms, agents can reuse lightly adapted tapes from other agents and revise their own tapes. This allows practitioners to maximally benefit from imperfect historical tapes by earlier versions of the agent, both for evaluating the newer versions and for improving them algorithmically. Last but not least, agents stream their intermediate events to the orchestrator to enable delightful interactive experiences.

We invite the reader to start their TapeAgents journey with the technical presentation of the framework in Section 2. There, we cover the details of agent architecture, agent-environment orchestration, tape content and structure. Section 3 describes three low-code agent-building framework prototypes on top of TapeAgents: one for monolithic agents, another for multi-agent teams and the third one with easy-to-tune function-like prompts. The same section also covers early versions of our Studio toolsuite for development and debugging and our Optimize toolsuite for agent optimization. In Section 4, we present diverse examples of building and optimizing agents using TapeAgents framework and tooling. Section 5 presents a deeper case study of a key practical TapeAgents use-case: optimizing the quality of a cost-effective conversational assistant using tapes from an expensive multi-step teacher agent. After presenting the framework and the examples we offer the reader a detailed comparison of TapeAgents with prior work in Section 6. Lastly, Section 7 discusses possible extensions and applications of TapeAgents.

## 2 TapeAgents: foundations

Our TapeAgents framework proposes an agent-building paradigm that facilitates all stages of the AI Agent development lifecycle. This section presents a detailed bottom-up approach to present the framework. First, we introduce the building blocks: the nodes, the agents, and the environment. Then, we explore how these parts can be composed and orchestrated to build a tape-centered system. In this section, we also describe the tape structure and metadata.

### 2.1 Nodes and Steps

As outlined in Figure 1, in TapeAgents, one builds the agent from **nodes**: the basic atoms of intelligence. A node describes one LLM call and the classical symbolic processing of the call’s output. The agent will dynamically determine which node to run next based on the tape. Nodes generate new tape entries that we call **steps**: basic atoms of the agent’s memory. Examples of what an agent can do in a step include making a long-term plan, reasoning about how to fulfill the plan or how to use a tool, requesting a tool call. Among these examples, the last one is an **action** step as it requests interaction with or impact on the agent’s environment. The first three examples are **thoughts**: the agent’s inner reasoning steps. The remaining step type in TapeAgents are **observations** that the agent receives from its environment in response to the agent’s actions. The reader can find an example tape with color-coded actions, thoughts and observations in Figure 3. In TapeAgents we often define a tape type by specifying what specific actions, thought and observations classes it can contain, though all such tapes are merely aliases for the one and only `tapeagents.core.Tape` class.

A typical node generates tape steps based on the stream of tokens that the agent receives from the last LLM call. To trigger the LLM call, the node constructs the LLM prompt through its `make_prompt` method that has the following Python signature:

```
def make_prompt(self, agent, tape) -> Prompt
```

(1)

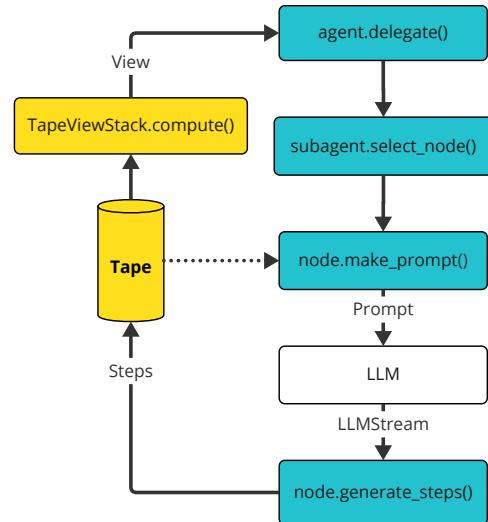
Some nodes perform only the conventional non-neural computation, like taking a branching decision. These nodes can use the default `make_prompt` implementation that produces a null prompt. Note that the node does not call the LLM directly but only makes a prompt. This is a deliberate design decision to keep all node methods pure functions, i.e. deterministic functions with no side effects.

By default, the agent will be calling its nodes sequentially and append the steps they create to the tape that the agent is asked to continue. One defines the step-generating behavior of a new node class in its `generate_steps` method:

```
def generate_steps(self, agent, tape, llm_stream) -> Generator[Step | PartialStep]
```

(2)

All nodes must generate `Step` objects; some can also parse the LLM token stream incrementally to produce partial steps which the agent will pass through to the application without adding them to the tape. Figure 2 shows how the agent runs one node and adds the resulting steps to the tape, along with the relationship between `make_prompt` and `generate_steps`.



**Figure 2:** A reasoning loop of an agent in TapeAgents. The root agent delegates to a subagent, the subagent selects the node, the node makes the prompt. The subagent calls the LLM with the prompt and lets the node process the resulting stream of tokens (LLM-Stream) that the root agent will then append to the tape.

The reader can find an example tape with color-coded actions, thoughts and observations in Figure 3. In TapeAgents we often define a tape type by specifying what specific actions, thought and observations classes it can contain, though all such tapes are merely aliases for the one and only `tapeagents.core.Tape` class.

A typical node generates tape steps based on the stream of tokens that the agent receives from the last LLM call. To trigger the LLM call, the node constructs the LLM prompt through its `make_prompt` method that has the following Python signature:

```
def make_prompt(self, agent, tape) -> Prompt
```

(1)

Some nodes perform only the conventional non-neural computation, like taking a branching decision. These nodes can use the default `make_prompt` implementation that produces a null prompt. Note that the node does not call the LLM directly but only makes a prompt. This is a deliberate design decision to keep all node methods pure functions, i.e. deterministic functions with no side effects.

By default, the agent will be calling its nodes sequentially and append the steps they create to the tape that the agent is asked to continue. One defines the step-generating behavior of a new node class in its `generate_steps` method:

```
def generate_steps(self, agent, tape, llm_stream) -> Generator[Step | PartialStep]
```

(2)

All nodes must generate `Step` objects; some can also parse the LLM token stream incrementally to produce partial steps which the agent will pass through to the application without adding them to the tape. Figure 2 shows how the agent runs one node and adds the resulting steps to the tape, along with the relationship between `make_prompt` and `generate_steps`.

### 2.1.1 Nodes That Can Make Training Data

Some nodes also implement the reverse direction — make the LLM output that would be required to produce the steps at a given index in the tape. The respective node method is

```
def make_llm_output(self, agent, tape, index) -> LLMOuput
```

(3)

This method is crucial for making fine-tuning data.

## 2.2 Agents

Like nodes, a TapeAgent agent generates steps and makes a new tape by appending the generated steps to the input tape. Specifically, `agent.run(tape)` runs an iterative reasoning loop that, at every iteration, selects a node, lets it make the prompt and generates the next steps (see Figure 2). By default, the agent will run its nodes sequentially (see next paragraph for more details). The loop continues as long as the nodes only generate thoughts. When a node produces an action, the agent stops and returns a new tape with the generated steps from all iterations appended to it. More precisely, `agent.run(tape)` returns an `AgentStream` object for streaming events like partial tapes and steps, but the final new agent tape is easy to extract from the stream object using `AgentStream.get_final_tape()` method.

An agent may have **subagents** for whom this agent is the **manager**. The subagents can have further subagents, which gives rise to a hierarchical agent organization with a single manager-free root agent on top. Given an input tape, the root agent determines the next active organization member to which delegate the generation of next steps. By default, the root agent makes the delegation decision by looking at the special `Call` and `Respond` thoughts. When an agent A wants the root to delegate to an agent B, A will append `Call(agent_name="B", content=...)` thought to the tape with an optional free-form message in the `content` field. When B responds by appending `Respond(content=...)`, A becomes active again. Note that both `Call` and `Respond` will affect the delegation logic *at the next agent iteration*. To sum up the delegation description, the root delegates to the agent that was called last and has not responded yet. See Figure 3 for an example of communication between a financial analyst agent and its web search helper.

### 2.2.1 Tape Views

In many cases, a middle-level agent A will make its prompts only using a subset of the tape’s steps starting from A’s last `Call` message and excluding the inner steps of the subagents that A called. To keep track of the steps each subagent can see, and to select the active agent and node, most agents compute the **tape view stack** from the tape and delegate to the agent whose view is at the top. For each agent that has not responded yet, the **view** contains the steps that this agent can see and the next node that the agent should run when it becomes active (see Figure 3). To track the next node, the `TapeViewStack.compute` function increments the top view’s next node pointer when it sees (from step metadata) that the current node has started running (see more on step metadata in Section 2.5). Exceptionally when `TapeViewStack.compute` encounters a special `SetNextNode(next_node=...)` thought, it changes the next node pointer to the value that the thought carries. A common usecase for `SetNextNode` is to implement looping within an agent.

A reader familiar with how Python interpreter works can find agents similar to Python functions, node similar to lines of Python code, steps similar to Python bytecode instructions, the tape view stack similar to the Python call stack and tape views similar to Python frames.

### 2.2.2 Optimizable Agents

Agent optimization algorithms tune agent prompts or alter agent structure in order to maximize the agent’s performance. To make such algorithms applicable to as many agents as possible, we standardize the structure of the agent configuration. We achieve this by making `tapeagents.agent.Agent` a Pydantic model<sup>1</sup> with the

<sup>1</sup>Pydantic models, from the Python `pydantic` package, provide robust data validation and parsing, ensuring that input data is properly structured and typed. This not only improves reliability by catching errors early but also simplifies code by automatically handling serialization, deserialization, and type conversions.

following mandatory fields: `.llms` for the LLM configurations, `.templates` for the prompt templates, `.nodes` for the nodes, and `.subagents` for the subagents.

Agents can also make training data for the LLM that they use. An agent's `agent.make_training_text(tape)` method reconstructs the LLM calls from a given tape, validates the reconstruction by replaying the step generation and returns training text characters. Internally, `agent.make_training_text` uses `node.make_llm_output` method introduced in Section 2.1.1; hence all nodes must implement this method for the agent to be trainable.

## 2.3 Environment

Just like nodes and agents, the environment in TapeAgents makes a new tape by adding steps to an existing tape. The main method of an environment object is:

```
def react(self, tape) -> Tape. (4)
```

The `environment.react` searches for the unfulfilled actions in the tape and adds the corresponding **observation** steps to the tape. Unlike nodes and agents, the environment may be non-deterministic and have side effects. We encourage agent developers to put all the deterministic and pure-function aspects of the system in the agent part, isolating only non-deterministic, computationally heavy or transactional aspects in the environment part.

## 2.4 Orchestration

To run a TapeAgent-based agentic application, one must alternate between running the root agent (which handles the delegation internally) and calling the environment to react to the agent's actions (see Figure 1). While we provide a default `tapeagents.orchestrator.main_loop` orchestrator for this purpose, we expect many application developers to build their custom orchestrators to closely control the agent-environment communication and ensure safety or enhance iteration logic.

### 2.4.1 Resumption and Replay

We designed TapeAgents with resumption and replay as key priorities. To resume, one can just restart the orchestration from an intermediate tape. For testing purposes, one can run an agent with replayed observations and LLM outputs and verify that this process leads to the same tape or print the diff otherwise. We found the replay tests to be incredibly helpful in our development work. When applicable, one can also replay the tape's observations (or even some of the agent's steps) in a new session to evaluate a new agent, though the old observations can be implausible if the new tape deviates too much from the old one.

## 2.5 Tape Metadata and LLM Call Database

Regardless of the orchestration method, the implementations of `agent.run()` and `environment.react()` ensure that the tape and its steps contain rich metadata, including these fields:

- `tape.metadata.author`: which agent or environment made this tape; either by authoring it, or by adding steps to it, or by making a revision of another tape.
- `tape.metadata.parent_id`: the ID of the parent tape of which the current tape is a continuation (when applicable).
- `step.metadata.agent`: the hierarchical name of the agent that generated the step.
- `step.metadata.node`: the name of the node that generated the step.
- `step.metadata.prompt_id`: the identifier (id) of the prompt that led to the generation of this step, see the explanation below.

When an agent runs a node, the node generates a unique ID for the prompt that it builds at this iteration. The prompt ID thus serves as the unique identifier of a node execution, i.e., of a specific iteration when the node was active. The ID also links the step to the LLM call from the node run so we can trace the origin of each step down to the specific prompt and LLM output. We store the prompt and the output for all LLM calls in an SQLite database. One can view LLM calls as an effective part of the tape in that they are always easily accessible; we don't include them in the tape to keep the latter lightweight.

The metadata is crucial for building the tooling and the algorithms that empower the agent developer. Figure 3 shows a visualization of some metadata fields.

### 3 TapeAgents: tooling

The TapeAgents foundation that we covered in Section 2 allows the creation of a wide range of reusable agent components, tooling and learning algorithms. What the right building blocks and tooling are often depends on the application area. In our initial release, we provide several prototypes to jump-start future open-source collaborations.

#### 3.1 Low-code Mini-Frameworks

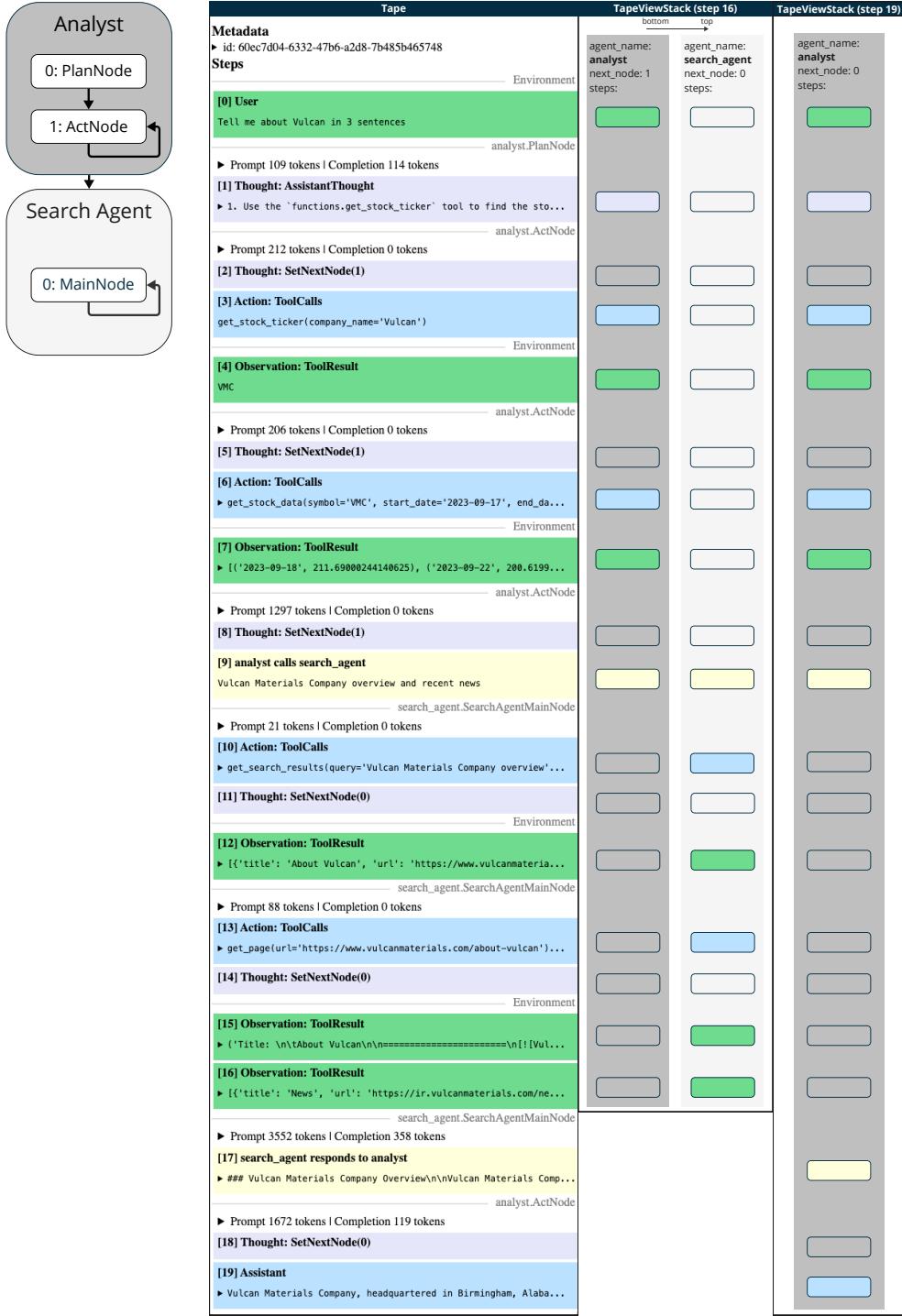
Building agents requires implementing many similar template rendering (`node.make_prompt`) and text parsing (`node.generate_steps`) routines. As a part of TapeAgents, we provide three examples of low-code mini-frameworks for building agents by composing and configuring off-the-shelf components:

1. **MonoAgent** exemplifies the most straightforward way to implement a monolithic agent: make a comprehensive prompt from all the data from the tape and the possible step schemas, then parse the LLM output using the schemas. One creates a MonoAgent from MonoNode nodes whose prompts are the same except for the final user message instruction. A MonoAgent also requires the agent developer to provide Pydantic models for all possible steps that the agent can generate.
2. **TeamAgent** shows how an AutoGen-style agent team can work in one tape. One can create three different kinds of team agents: (a) an initiator that send the first `Call` message, (b) a manager that chooses the next active agent, (c) a worker agent that responds using its system prompt.
3. **LLMFunction** demonstrates how one can build agents using function-style prompt templates, akin to DSPy signatures. These prompt templates are particularly easy to optimize by adding demonstrations.

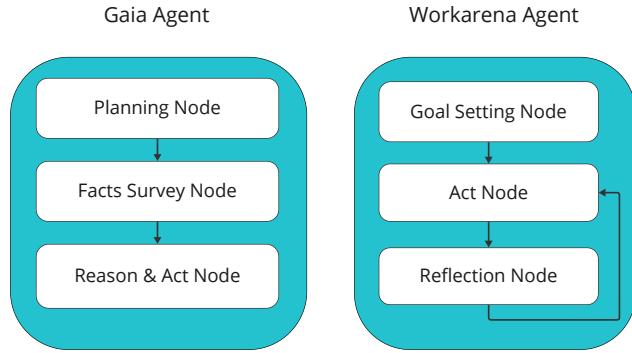
We include the mini-frameworks mostly for demonstration purposes, as it is hard to offer a high-level programming paradigm without a good knowledge of the intended application domain. The TapeAgent paradigm makes it easy to build such mini-frameworks thanks to the agent's double compositionality (agents and nodes).

#### 3.2 Tooling

In TapeAgents, the agent configuration and the tape are highly structured and linked with metadata. This allows us to offer developer tooling for a broad range of possible TapeAgents. In the initial release, we include several app prototypes. We offer *TapeAgents Studio* (see Figure 9 in Appendix B), an app to interact with the agent and its tape, *Tape Browser* (Figure 10 in Appendix B), an app to inspect a batch of tapes, and *Tape Diff* (Figure 11 in Appendix B), which compares two batches of tape. Furthermore, for agent optimization, we provide algorithms for auto-prompting, LLM fine-tuning, and a modular Reinforcement Learning orchestrator. The finetuning component uses `Accelerate` and `DeepSpeed` libraries and supports tuning resumption, experiment tracking, reproducibility, LoRA tuning, and distributed training. The above apps and algorithms represent the first steps towards the fully fledged Studio and Optimize modules that we envision in Figure 1.



**Figure 3:** A multi-agent tree structure (left) and a tape resulting from their work (middle) with the TapeViewStack at specific steps (right). At step 16 the stack's top view is the SearchAgent's tape view. At step 19, only the Analyst's view exists. Note how the Analyst's view does not include the Search Agent's steps except for its response. Steps are color-coded: yellow for communication thoughts, purple for internal agent thoughts, blue for actions, and green for observations. The step's author is indicated in grey using the “agent.node” format.



**Figure 4:** Agent structures for GAIA and WorkArena experiments (see Section 4.2 for details).

## 4 Examples

In an initial set of examples, we demonstrate agents that represent different agent-building paradigms, as well case-studies of using different agent optimization methods.

### 4.1 Financial Analyst and Their Web Search Helper

To offer an example with the maximal educational value, we have implemented a user-facing financial analyst agent that can delegate searching the web to its subagent. We show the structure of the analyst agent and an example tape in Figure 3. Our introductory hands-on notebook<sup>2</sup> takes the reader through a journey from TapeAgents basic concepts to building this agent.

For illustrative purposes, we implemented the nodes in this example from scratch, without using mini-frameworks from Section 3.1. We offered the analyst agent an environment with several tools: one to get the company ticker, another to download stock data, as well as several tools to search and browse the web. We inform the analyst and their web search helper of the tools that they can use by including their tools’ schemas in the prompts that the respective agent’s nodes make. The agent operates on a tape type called `DialogTape`, which can only contain two kinds of actions: `ToolCalls` to call one or more tools and `AssistantStep` to respond to the user. The agent uses the same `ToolCalls` step with different content to call different tools. This is the use we intend for `DialogTape`: quick agent prototyping without declaring usecase-specific step schemas, though we believe most TapeAgents users will find it useful to declare their own action and thought types.

### 4.2 Open-domain Question Answering and Web Browsing With Monolithic Agents

To validate TapeAgents quantitatively, we build two agents that target existing benchmarks. The first one is a question-answering (QA) agent that targets the GAIA benchmark (Mialon et al., 2024). The QA agent can search the web, run Python code, read multiple file types. To meet the GAIA evaluation requirements, we prompt the agent to output the precise short answer only. We build the QA agent from `MonoNode` nodes, with two planning nodes and one acting node in which the agent loops (see Figure 4). Table 1 shows that the agent performs well for a simple monolithic agent of its kind. Implementing this agent mostly requires gathering the tools, declaring corresponding action steps (like `ReadDocumentAction` and `UseCalculatorAction`) and declaring usecase-specific thoughts for reasoning (like `ListOfFactsThought` and `NewFactThought`).

We used a similar approach to build a web-browsing agent that targets the WorkArena benchmark (Drouin et al., 2024). We likewise declare action classes like `HoverAction` and `PressAction`. Figure 4 illustrates the exact agent structure. We benchmark our web agent and find that it performs competitively (see Table 2).

<sup>2</sup><https://github.com/ServiceNow/TapeAgents/blob/main/intro.ipynb>

**Table 1:** Gaia Agent evaluation results.

Framework & Model	Val accuracy, %	Test accuracy, %
das_agent v0.4 GPT-4o (SOTA)	—	38.2
HuggingFace Agents GPT-4o	44.2	33.3
TapeAgents GPT-4o	33.9	27.6
FRIDAY GPT-4-turbo	34.5	24.3
TapeAgents GPT-4o-mini	25.5	—
GPT-4 + manually selected plugins	14.6	14.6

**Table 2:** Workarena Agent evaluation result on Workarena L1 tasks.

Framework & Model	Accuracy, %
TapeAgents GPT-4o	44.2
Agentlab GPT-4o	42.7
TapeAgents GPT-4o-mini	29.1
Agentlab GPT-4o-mini	23.0

### 4.3 Data Science With a Team of Agents

To demonstrate that TapeAgents natively supports the multi-agent paradigm, we implement a “data science” agent team that consists of the Requestor, Manager, Software Engineer, Code Executor and Asset Reviewer agent. Figure 8 (in Appendix A) shows the team in action as it builds a stock price comparison plot. We drew inspiration from the popular AutoGen framework for the multi-agent communication pattern in this example. Benefits of the TapeAgents implementation of this agent team include that one can easily resume the team from an intermediate tape or use tapes to optimize the entire agent organization algorithmically.

### 4.4 Finetuning a Cheap Math Agent

We test-drive TapeAgents fine-tuning component with an example of training a LLAMA-8B-based math agent using tapes by its teacher counterpart with LLAMA-70B under the hood. We equip each agent with a reasoning node and run the in environment with a calculator tool. After finetuning on 3,000 samples from 1,000 teacher tapes, the student performance rises significantly from 66.2% to 77.5%, though the teacher’s performance, at 93.1%, remains much higher still.

### 4.5 Prompt-Tuning for Agentic RAG

In our last example we show how the tape, the agent configuration and the metadata linking them, can serve as a medium to implement data-driven agent optimization algorithms. In this example we use `LLMFunction` prompt templates that describe the intended behavior of a transformation that the LLM should perform, including the instruction, the input / output format, and optionally, a few demonstrations. We designed `LLMFunction` to make it possible to implement DSPy-like algorithms in TapeAgents. Below we describe how we used TapeAgents components to closely reimplement the DSPy introductory notebook.

We compose a Retrieval-Augmented Generation (RAG) agent that performs two rounds of query generation and Wikipedia retrieval and then produces a short factual answer. We build this agent mostly from `LLMFunctionNode` nodes that describe how the input fields in their respective `LLMFunction` templates should be filled with the steps from the tape. The only different kind of a node is a null-prompt node that deduplicates the retrieved paragraphs. We tune the prompts of the resulting 5-node agent by adding demonstrations to the function prompt templates. We obtain demonstrations by running the agent on training examples from HotPotQA and filtering the tapes with the wrong answer or duplicate queries. In this setting, prompt-tuning leads to modest gains but significant gains in retrieval accuracy, raising it from 50% to 56%. The optimized agent is still a TapeAgent that can be resumed from any intermediate tape, unlike a free-form Python program that uses DSPy. Notably, the implementation of the actual demonstration selection algorithm

took just 12 lines of code, highlighting how the TapeAgent structures and metadata facilitate algorithm implementation.

## 5 Case Study: Building a Cost-Effective Enterprise Form-Filling Assistant

A key use case of TapeAgents is optimizing LLM Agents to offer great quality services at a fraction of the cost. In this section, we present a fleshed-out example of how these goals can be achieved for a conversational assistant that can help fill a request form and submit the request.

### 5.1 Problem Setting

Employees in large enterprises often fill forms to request resources, assistance or access. A conversational assistant can make the form-filling experience smoother by guiding its user to the right form, by accepting the user’s free-form inputs, and by answering the questions that the user may have in the process. For a great experience, the assistant must also gracefully handle the “unhappy-path” situations, such as when the user’s ask is impossible to fulfill or when the assistant cannot answer the user’s question. In this case study, we show one can use TapeAgents to train a cost-effective assistant that scores high according to a formal metric of user experience that we call the *GREADTH* score. GREADTH stands for Grounded, REsponsive, Accurate, Disciplined, Transparent, and Helpful. We will explain these metrics in Section 5.2.

For simplicity, we consider building a restricted assistant:

- The assistant should answer questions solely based on the form documentation; it does not have to retrieve any additional documents.
- The assistant can only help with one form at a time.
- At the start of the conversation the assistant converses with the user to guide them to the correct form.
- After the form is chosen, the assistant will help the user fill out the form correctly. The agent will not allow the user to switch to a different form after this point.
- During the form-filling process, the assistant maintains the field values that the user has provided so far.
- The interaction ends with either the form submission or the agent exiting the conversation after the user confirmation.

The resulting form-filling setup is reminiscent of the Task-Oriented Dialogue setting that has been widely discussed in the literature (Rastogi et al., 2020; Budzianowski et al., 2018). Following this body of work, we will refer to form fields as *slots*.

### 5.2 Evaluation Criteria: GREADTH Experience

Despite the apparent simplicity of the form-filling setup, it can be non-trivial to develop a form-filler assistant that balances an excellent conversational experience with low hallucination rate and reasonable cost. To balance these desiderata, one must first define them in a measurable way. In our case study, we train our assistant to have maximum **GREADTH**: **G**rounded, **R**Eresponsive, **A**ccurate, **D**isciplined, **T**ransparent, and **H**elpful. We define these aspects as follows:

- Everything a **Grounded** assistant says must be fully supported by the form documentation, the conversation history and the *grounding statement*. The latter defines the assistant’s identity and purpose and constrains the assistant to form-filling. Small talk is considered ungrounded.

- A **Helpful** assistant must actively take the conversation forward by asking for the user’s intent, requesting the next slot to fill, or asking for confirmation before making the request once all slots have been filled. It should also (a) provide all relevant information regarding a slot when asking for it (default value, allowed values, optionality), (b) answer any user question if the form documentation provides the relevant information, (c) exit the conversation at any time if the user desires so.
- An **Accurate** assistant must correctly identify the user’s intent and import the relevant form documentation, fill the slots correctly based on user messages, update the slots if the user changes their mind, or skip the slots if relevant.
- A **Transparent** assistant acknowledges all changes made to the partially filled form. This includes *slot-filling* or *skipping* slots that are optional or have a default value. The summary of slots changes can be concise, yet the user must be able to understand how the slots were affected. While in a mixed modality interaction the user may visually see the form changes, in a purely voice interaction such as talking over the phone the transparent behavior is essential.
- A **Disciplined** assistant must follow its planning thoughts, such as requesting a slot, asking for confirmation, answering a question, rejecting incorrect slot values (as defined by the form documentation), or rejecting an invalid ask.
- We require the assistant to be **REsponsive** to address a common experience issue with AI assistants: the robotic and opaque behavior when the user goes off the expected conversational path. We wish that AI assistant infers and acknowledges what the user had in mind, while explaining that their request or question is not possible. In particular, we want to cover the following scenarios:
  - if the user tries to fill a slot with an invalid value, the assistant should acknowledge the value and respond that it is invalid;
  - if the user offers information that looks like a value for a nonexistent but plausible slot, the assistant should acknowledge the value and the inferred name of the slot and respond that such slot is not available in this form;
  - if at the form-filling stage the user’s ask looks like a request for another form, the agent should acknowledge their ask and say that it can not fulfill it right now (note that in our setup the user must either finish filling the current form or exit);
  - if the user asks a plausible question that the form documentation does not answer, the agent should acknowledge the question and say that it can not be answered.

The user may ask other requests that have nothing to do with form-filling (i.e. weather requests). In that case, the agent must politely decline the request and keep moving the conversation towards either submitting or aborting the request. To align our definition of responsiveness with the common sense meaning of this word, we also require a **REsponsive** assistant to acknowledge all valid slot values and valid questions. Thus we often deem a response that is not **Transparent**, or not **Helpful** also not **REsponsive**.

The GREADTH criteria above are binary, a conscious choice that we made to simplify the evaluation and the analysis. We acknowledge that this makes them somewhat crude, as e.g. two assistant answers can be both technically correct but can widely differ in readability and in the choice of the information to present. One can complement these criteria with a preference-based experience evaluation that implicitly covers fluency, verbosity, and other aspects of the assistant’s response.

### 5.3 Design of a Form Filling TapeAgent

**Tape Structure** To build an agent that provides a GREADTH experience, we decompose the conversational form-filling task into smaller reasoning steps before each agent message. Having the GREADTH metrics in mind, we define the agent’s thoughts to help it plan its response. The thoughts are used to represent a chain-of-thought of the agent, which includes (a) analyzing the user’s intent (e.g. the form that

is requested, the provided slot values, a question being asked), (b) updating the internal state of the conversation (e.g. slot-filling), and (c) planning the next actions (e.g. requesting a specific slot value or requesting a user confirmation). In particular, after each user message or observation, the agent must return:

1. A list of **thoughts**: these include slot-filling related thoughts such as `UpdateFunctionParametersThought`<sup>3</sup>; and message planning thoughts specifying the next slot to request (`RequestFunctionParametersThought`), the need to ask for confirmation before submitting the request (`RequestFunctionCallConfirmationThought`), planning to answer a question (`AnswerFromFunctionSchemaThought`), planning to inform the user that their question cannot be answered (`NoAnswerInFunctionSchemaThought`), refusing unsupported request/behavior/slot values (`RefuseInexistentFunctionThought` / `RefuseToEngageThought` / `RefuseInvalidParameterValueThought`), etc. The full list of thoughts is described in Appendix C.2.
2. A single **action**: the agent returns a single action, such as searching available forms (`ResolveFunctionAction`), retrieving form documentation (`InspectFunctionAction`), replying to the user (`PromptUserForTextMessageAction`), submitting the request (`CallFunctionAction`), or exiting the conversation (`ExitAction`). Each action results in new observations (available forms, retrieved documentation, or user input), and ends the current agent turn.

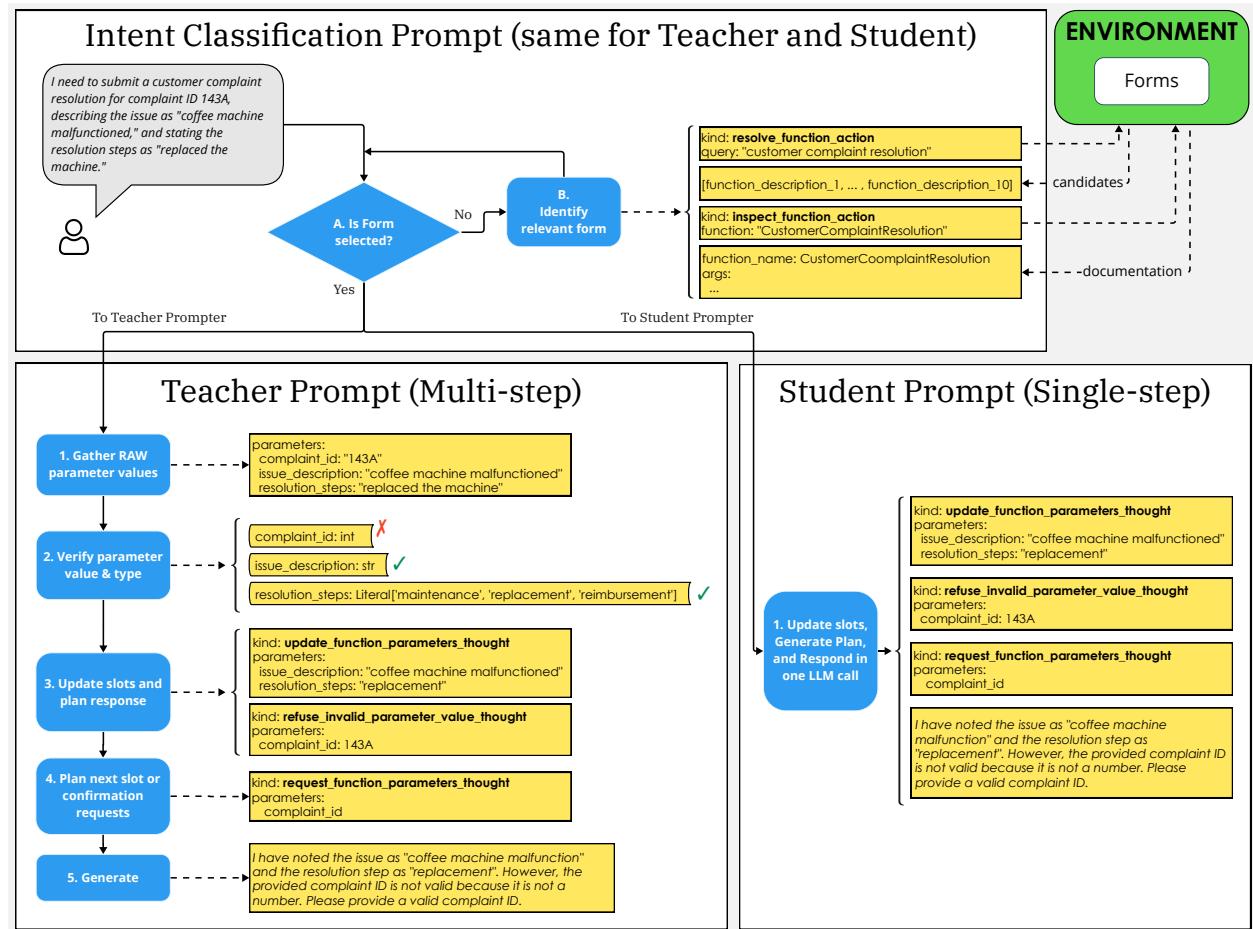
**Multi-node Teacher Agent** We experimented with various combinations of prompting techniques and teacher LLMs to obtain the best performance (see GREADTH metrics in Sections 5.2 and human evaluation procedure in Section 5.4). We found that using Llama-405B in multi-node teacher agent yielded the most promising results based on human evaluation. For additional details about the choice of model and agent structure, see Appendix C.4.

Our teacher TapeAgent is made of 7 nodes: 2 for intent classification and 5 for slot filling (Figure 5):

- **Is Form Selected?** The initial node checks if we identified the user’s intent. This node does not call any LLM, it simply checks if an `InspectFunctionAction` step is already present in the current Tape. If a form has not yet been selected, the model proceeds to the next step (Step (B) in Figure 5). Otherwise, the model moves directly to the form-filling phase (Step 1 (left) in Figure 5).
- **Intent Discovery.** This node (i) lists the available forms (`ResolveFunctionAction` queries the Environment to return a list of available forms) and (ii) finds the relevant one based on the previous user’s message (`InspectFunctionAction` queries the Environment to return the form documentation). If the LLM cannot identify the relevant form, it is prompted to ask again the user for its intent (`RefuseInexistentFunctionThought`; `RequestFunctionThought`; `PromptUserForTextMessageAction`).
- **Gather Raw Parameter Values.** Once the user’s intent is discovered and its form documentation imported, the LLM is prompted to extract all raw slot values present in the user’s message and yield a `GatherValuesThought` step.
- **Verify Parameter Values & Types.** In this node, the LLM is prompted to verify all extracted slot values based on the form documentation imported. The LLM verifies that each value is of the correct type and is valid (in the case of categorical slots) with a `VerifyValuesThought` step.
- **Update Slots and Plan Response.** Once we have identified the correct/incorrect slot values, the LLM is prompted to update the filled slots with a `UpdateFunctionParametersThought` step, refuse invalid slots with `RefuseInvalidParameterValueThought` steps, and start planning its response to the user with `AnswerFromFunctionSchemaThought` if applicable.

---

<sup>3</sup>While working on the form-filling case-study we interchangeably used the terms form / request / function and the terms slot / field / parameter. In this report we use the original technical names for the tape’s steps to maximize the coherence between the text and the code.



**Figure 5:** Node structure of the teacher and the student agents. The teacher agent combines intent classification and teacher prompts, while the student agent combines intent classification and student prompts. We represent nodes in blue and produced steps in yellow. Please note that words *intent/form/function* and *slot/field/parameter* are used interchangeably in this report.

- **Plan Next Slot or Confirmation Requests.** This node prompts the LLM to move the conversation forward by either (i) requesting the next slot (`RequestFunctionParametersThought`), (ii) confirming that the user wants to submit the current request (`RequestFunctionCallConfirmationThought`), or (iii) confirming that the user wants to exit the chat (`RequestExitConfirmationThought`).
- **Generate.** Eventually, the final node of our multi-step Teacher TapeAgent prompts the LLM to generate the next action. Based on the previous steps/thoughts in the Tape, the agent can either (i) write a message (`PromptUserForTextMessageAction`), (ii) submit the request (`CallFunctionAction`), or (iii) exit the conversation (`ExitAction`).

See Figure 12 of Appendix C.5 for a sample Teacher tape.

**Student Prompt (Single-Node)** We design the student agent with the goal to minimize input and output token counts in order to optimize cost. While we could get rid of the instructions entirely, we still provide enough instructions that a strong model (e.g. Llama-40B) can still make sense of the task in a zero-shot setting.

- **Is Form Selected?** Same structure as the teacher but with fewer and more compact instructions.
- **Intent Discovery.** Same structure as the teacher but with fewer and more compact instructions.

- **Update Slots, Generate Plan, and Respond.** Using a single short prompt, we ask the agent to generate all its thoughts (e.g. `RefuseInvalidFunctionParameterValueThought`; `UpdateFunctionParametersThought`; `RequestFunctionParametersThought`) and end with an action (e.g. `PromptUserForTextMessageAction`).

We show an example of the student tape in Figure 13 of Appendix C.5.

## 5.4 Experiments

We consider the task of training a cost-effective conversational agent to help the user fill and submit a form. We seek the conversational experience to score high in GREADTH metrics (Section 5.2), while requiring less tokens per interaction and less cost per processed token. We attempt to distill a multi-node teacher that uses a large model into a single-node teacher that uses a small model. For both finetuning and evaluation purposes, we simulate the environment (synthetic forms and user interactions).

**Synthetic Companies** To simulate an enterprise environment with multiple forms, we prompt a Llama-3-70B-Instruct model to generate the name and descriptions of 6 fictitious companies, which we divide into training domains (`FlyCorp`, `BigBankCorp`, `CoffeeCorp`) and testing domains (`DriveCorp`, `LuxuryCorp`, `ShopCorp`). For each company, we then prompt Llama-3-70B-Instruct to generate 10 plausible request forms based on the company name and description. Each form has a name and a description. Eventually, for each generated form description, we prompt Llama-3-70B-Instruct to generate a FunctionSchema for that request form. FunctionSchemas are structured data representations including a name, a description, a Json-Schema describing slots to fill, and a JsonSchema describing the object returned once the form is submitted. Parameter slots have a name, description, type (categorical, date, email, string), optionality, possible and default values. The prompts used to generate synthetic companies are described in Appendix C.1. Each of our simulated environments has 10 available forms and identifying which form the user requests is part of the form-filling task we tackle.

**User Agents** In the multi-turn dialogue setting, distillation is more complex than running the teacher on a set of static contexts because of alternating agent and user turns. The user turns need to either be produced by a human or generated. We define 19 different (single-node) User Agents to generate the next user message by simulating a variety of user behaviors. Some user behaviors are “easy” such as “*answer the agent’s question*”, while others can be quite adversarial such as “*provide a good value for slot X and a bad value for slot Y*” or “*ask for something unrelated*”. All user agents and their respective behaviors are described in Appendix C.3. Each of these User agents serves as a special instance of the Tape Environment that responds to the assistant after each `PromptUserForTextMessageAction` step.

**Synthetic Dialogue Generation** We generate a dataset of dialogues by repeatedly and alternatively prompting the teacher agent and a user agent to generate the next turn from a set of partial conversations. The minimal partial conversation contains only the “*Hi, how can I help you?*” Assistant message. User agents are sampled randomly after each `PromptUserForTextMessageAction` steps from the Assistant, and yield a user message Observation. The Assistant then continues the conversation until its next message. The conversations end when they reach 18 turns (9 agent messages and 9 user messages), whenever the request is submitted or aborted, or if the agent fails to produce a valid continuation (e.g. trying to fill a nonexistent slot). We show an example of a full conversation between the teacher agent and the user agent in Figure 12 (Appendix C.5). We structure the dataset as a tree of dialogues, while controlling for the width of the tree (beam search), the diversity of user behaviors, and the diversity of forms (requests). We use a beam size of width 500, up to 9 user turns and 9 agent turns. This results in roughly **13k train** and **13k test** agent continuations per synthetic company.

**Human Evaluation** We perform human evaluation to score each agent turn on the basis of the GREADTH metrics: groundedness, responsiveness, accuracy, discipline, transparency, and helpfulness (Section 5.2). The labeling services were provided by Toloka<sup>4</sup>, the labelers were paid above the minimum wage in their respective

---

<sup>4</sup><https://toloka.ai/>

**Table 3: GREADTH Form Filler experiment results.** The teacher agent<sup>1</sup> is a Llama 3.1 405B Instruct FP8 model with a multi-node flow. The student agent<sup>2</sup> is a Llama 3.1 8b Instruct with a single node. We also evaluate a teacher-equivalent agent with GPT-4o as a reference comparison. The metrics are computed over 1524 partial dialogues from the test domains. Read full analysis in Section 5.4.

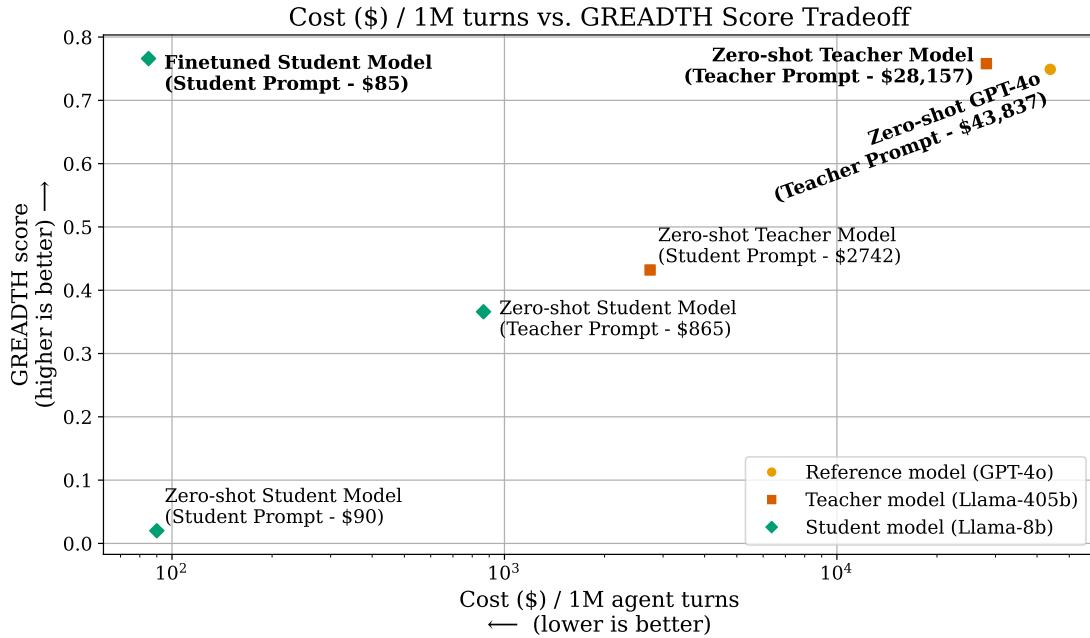
Agent(Model+Prompt)	G	Re	A	D	T	H	GREADTH Score (Human Raters)
<i>Reference Comparison (GPT-4o-2024-08-06)</i>							
Teacher Prompt (0-shot)	91.3%	87.1%	91.4%	92.7%	94.3%	87.2%	74.9%
<i>Teacher Model (Llama-3.1-405B-Instruct)</i>							
<b>Teacher Prompt<sup>1</sup></b> (0-shot)	89.8%	85.0%	87.9%	91.6%	92.5%	86.5%	75.8%
Student Prompt (0-shot)	74.2%	72.0%	76.8%	67.3%	78.9%	61.9%	43.2%
<i>Student Model (Llama-3.1-8B-Instruct)</i>							
Teacher Prompt (0-shot)	75.5%	57.7%	72.4%	74.0%	76.3%	60.3%	36.6%
Student Prompt (0-shot)	18.8%	6.2%	10.9%	11.6%	9.4%	12.7%	2.0%
<b>Student Prompt<sup>2</sup></b> (finetuned)	92.1%	86.4%	90.2%	94.4%	95.1%	87.1%	76.6%

**Table 4: Agent Cost vs. Performance Tradeoff.** On the evaluation set, we report the average number of input and output tokens per agent turn, and the average cost per million agent turns, by multiplying the number of tokens with the retail price of the cheapest provider on Openrouter as of October 3rd 2024 (\$0.055 per million input/output tokens for Llama-3.1-8B-instruct and \$1.79 for Llama-3.1-405B-instruct). The cost per million agent turns is only \$85 for the student vs. \$28,157 for the teacher, which represents a **factor 300** in savings. The cost for the teacher could be mitigated by prefix caching, but most of the teacher prompts are still context-dependent. Read full analysis in Section 5.4.

Agent (Model+Prompt)	Input Tokens /turn	Output Tokens /turn	Cost /1M turns	GREADTH score
<i>Reference Comparison (GPT-4o-2024-08-06)</i>				
Teacher Prompt (0-shot)	15,431	526	\$43,837	74.9%
<i>Teacher Model (Llama-3.1-405B-Instruct)</i>				
<b>Teacher Prompt</b> (0-shot)	15,189	541	\$28,157	75.8%
Student Prompt (0-shot)	1,435	97	\$2,742	43.2%
<i>Student Model (Llama-3.1-8B-Instruct)</i>				
Teacher Prompt (0-shot)	15,189	541	\$865	36.6%
Student Prompt (0-shot)	1,437	208	\$90	2.0%
<b>Student Prompt</b> (finetuned)	1,441	110	\$85	76.6%

work locations. Labelers provide the **6 GREADTH binary labels** per agent turn. We also have expert labelers who audit 10% of the labels. We report scores for each metric as well as the **GREADTH score**, which is *equal to 1 only if all 6 metrics are satisfied, and 0 otherwise* (binary AND). If the agent fails to produce a valid continuation – due to unparsable JSON, invalid schema, attempting to fill an invalid slot value, submitting a form with missing values – then the agent automatically gets 0 for all 6 metrics.

**Agent Distillation** We construct a **teacher agent** by prompting a very large model, Llama-3.1-405B-Instruct-FP8 (runs on 8 H100-80GB GPUs) with sequential and lengthy instructions (**teacher prompt**) in order to obtain the best GREADTH score, with no regards to prompt length optimization or latency. From a qualitative perspective, our interaction with the teacher agent reveals a generally satisfactory user experience and validates its choice as a teacher agent. We construct a **student agent** by combining a much smaller model, Llama-3.1-8B-Instruct model (runs on a single A100-80GB GPU), with a single-shot prompt (**student prompt**) optimized for length. We distill the teacher agent into the student agent by finetuning it over **13k teacher agent turns** (continuations) on the training domains. We perform a single epoch of LoRA (Hu et al., 2021) optimization using AdamW (Loshchilov and Hutter, 2017) with learning rate 1e-5 and batch size 32. More epochs did not seem to help.



**Figure 6: Cost per 1M Agent Turns vs. GREADTH Score Tradeoff.** The finetuned student agent (top-left) performs on par with the teacher and reference agents (top-right) for a fraction of the cost. We also provide ablations for various combinations of models and prompters, which show that finetuning is instrumental in getting the desired performance when using a short (student) prompt. See Section 5.4 for the full discussion.

**Results** We evaluate the teacher and student agents by generating a single agent turn (continuation) over a set of **1524 partial dialogues** subsampled from the testing domains generated previously. We control for the diversity of user behaviors (19 mostly uniform) and for the diversity of requested forms (30 forms split over 3 domains + dialogues where no form has been selected yet). Here is what we observed:

1. The teacher agent achieves a GREADTH Score of **75.8%** (Table 3) at the expensive cost of **28,157\$/1M** agent turns (Table 4), due to using a large model and multiple prompts with lengthy instructions. In comparison, the student agent achieves an honorable **76.6%** GREADTH score comparable to the teacher model but costing only **85\$/1M** agent turns, a **factor 300x cheaper**, thanks to the shorter prompt and smaller model.
2. When we prompt the teacher model and student model with the student prompt (0-shot, with no finetuning), we achieve GREADTH scores of only 43.2% and 2.0% respectively. This is not particularly surprising given that the shorter prompt may be ambiguous, though it is interesting to observe the big gap between student and teacher model (43.2%-2.0%=41.2% points) is entirely closed by the finetuning process (the student is 0.8% points above the teacher).
3. Similarly, when prompting the teacher and student models with the teacher prompt, we observe a large difference of 75.8%-36.6%=39.2% points, which mostly confirms that while using the largest models is most crucial in the zero-shot regime, finetuning the smaller model can close the gap and generalize across domains.

## 6 Related Work

Among the developer-oriented frameworks most relevant to TapeAgents are LangGraph (Chase, 2023) and AutoGen (Wu et al., 2024a). In LangGraph, one builds the agent at a very low-level as a concurrent graph-based state machine. AutoGen offers a high-level paradigm to build multi-agent teams. TapeAgents

	Build	Debug	Serve	Optimize
DSPy	✓	⚠	🛑	✓
LangGraph	⚠	✓	✓	🛑
AutoGen	⚠	⚠	⚠	🛑
TapeAgents	✓	✓	⚠ <sup>z</sup>	✓

**Figure 7: TapeAgents vs Other Frameworks.** TapeAgents stands out in how it supports the practitioner in tasks throughout the development cycle. In this figure we use the stop sign to indicate that major core changes would be required for the framework to be helpful for a task. Warning sign indicates partial support. We indicate limitations of serving support in TapeAgents that we will address in future work, see Section 7 for a longer discussion.

combines the best of these two worlds, as it allows both the low-level control and the implementation of higher-level low-code paradigms like AutoGen. Neither LangGraph nor AutoGen are designed with agent optimization in mind.

On the other side of the spectrum are AI frameworks that have recently demonstrated techniques to automatically optimize prompts and tweak other aspects of agent configuration, such as the flow or the assignment of tools to the agents. Solutions using DSPy (Khattab et al., 2023b) and TextGrad (Yuksekgonul et al., 2024) attain higher performance compared to human prompt engineering. AgentOptimizer (Zhang et al., 2024) and agent symbolic learning (Zhou et al., 2024) enable improving agent tools and pipelines. Meta-Agent Search (Hu et al., 2024) aims to create the best multi-agent architecture. The rich metadata that links the tape with the agent configuration in TapeAgents provides a perfect medium for implementing agent optimization algorithms like the ones from the above works. Notably, DSPy and TextGrad implement control flow in pure Python, which created challenges for resuming the agent from a persistent session state. In TapeAgents, developers can freely use Python within one node, but between-node control flow is handled by adding steps to the tape, which makes the tape perfect for session persistence.

We refer the reader to Figure 7 for a tabular summary of the above analysis. One can see that TapeAgents uniquely helps practitioners to both develop the agent and optimize it in a data-driven way.

Agent observability software such as LangSmith<sup>5</sup> and Langfuse<sup>6</sup> adds visibility to agent execution. They allow one to incrementally instrument agent code to track specific components. In TapeAgents, the tape offers complete observability by design, but beyond that it can also be used for point-in-time resumption and agent optimization.

## 7 Discussion and Future Work

We have presented TapeAgents, a holistic framework that targets all stages of the LLM Agent lifecycle. We believe the tape-centered approach of our framework can facilitate responsible deployment and continual improvement of LLM agents. Initial tapes will help debugging and testing at the development stages, historical tapes will serve as a machine-readable source of evaluation and training data. Red-teaming algorithms can use historical tapes as seed data for testing the agent on potential attacks or business-critical dangerous situations. The practitioner can also use historical tapes to seed the simulation that they use for testing the agent. These are but a few benefits that TapeAgents can bring to practitioners.

<sup>5</sup><https://www.langchain.com/langsmith>

<sup>6</sup><https://github.com/langfuse/langfuse>

## 7.1 Immediate Next Steps

TapeAgents is still in early development stages. A key next step for TapeAgents is adding coroutine implementations for the agent loop and for agent-environment orchestration. This will enable both running many agent-environment loops in parallel on their respective tapes and running members of the same agent team on their shared tape. The latter will require changes in the tape view computation to ignore steps of the agents running in parallel, but we believe that the main framework concepts that we introduced in this paper will stay the same.

On the optimization front, we will soon release an online Reinforcement Learning (RL) trainer for TapeAgents, which will improve the assistant agent using the rewards that the annotator agent computes. Another welcome optimizer addition would be implementing in TapeAgents a text-based feedback-propagation algorithm like the one in TextGrad. Tapes steps are a perfect medium to attach feedback to.

## 7.2 Agent as an Optimizable Workflow

Stepping back from the immediate future plans, we believe it is worth reflecting on what should be called an LLM Agent and what should be called “just” a program, a workflow or software. In the TapeAgents context, this philosophical question is what the developer asks themselves when they build an agent—whether they should implement as agent nodes, and what should go in the application that uses the agent. Our current recommendation is that one should treat and implement the parts of the system that they intend to optimize with data-driven algorithms as LLM Agents. Frameworks provide the structure that the algorithms require to identify the issue, propose a change, and test the change’s outcome. In TapeAgents, this process is particularly clear: the algorithm will identify an issue in the tape, attribute it to a root cause step, propose a change to the agent configuration and test this change by resuming the agent from intermediate points in the tape. Thus the gains from algorithmic improvement will compensate for the overhead of respecting the TapeAgents engineering constraints. To sum up, our recommendation is to implement *optimizable workflows* as LLM Agents and use other appropriate tools for the software that will not be subject to data-driven improvement.

## 7.3 Synthetic Data Generation with Worlds of TapeAgents

In addition to helping practitioners with their solution-specific challenges, we envision synthetic data generation as another application area where TapeAgents can make an impact. A key trend in the data-making trade is building modular pipelines with many agent-like modules, such as judges (Bai et al., 2022), meta-judges (Wu et al., 2024b), process supervisors (Lightman et al., 2023; Uesato et al., 2022), annotator augmented with tools (Wei et al., 2024) among other examples. We believe TapeAgents is a great foundation for the continual improvement of such multi-agent pipelines with human feedback, as implementing all pipeline modules as TapeAgents immediately makes them optimizable.

## References

- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback.
- Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Stefan, U., Osman, R., and Gašić, M. (2018). Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Chase, H. (2022). LangChain.

- Chase, H. (2023). LangGraph.
- Drouin, A., Gasse, M., Caccia, M., Laradji, I. H., Verme, M. D., Marty, T., Vazquez, D., Chapados, N., and Lacoste, A. (2024). Workarena: How capable are web agents at solving common knowledge work tasks? In *Forty-first International Conference on Machine Learning*.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv: 2106.09685*.
- Hu, S., Lu, C., and Clune, J. (2024). Automated design of agentic systems.
- Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., and Potts, C. (2023a). Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.
- Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., and Potts, C. (2023b). Dspy: Compiling declarative language model calls into self-improving pipelines.
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. (2023). Let's verify step by step. *arXiv preprint arXiv:2305.20050*.
- Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *International Conference on Learning Representations*.
- Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. (2024). GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*.
- Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. (2022). Solving math word problems with process- and outcome-based feedback.
- Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., and Le, Q. V. (2024). Long-form factuality in large language models.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. (2024a). Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*.
- Wu, T., Yuan, W., Golovneva, O., Xu, J., Tian, Y., Jiao, J., Weston, J., and Sukhbaatar, S. (2024b). Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.
- Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., and Zou, J. (2024). Textgrad: Automatic "differentiation" via text.
- Zhang, S., Zhang, J., Liu, J., Song, L., Wang, C., Krishna, R., and Wu, Q. (2024). Training language model agents without modifying language models. *ICML'24*.
- Zhou, W., Jiang, Y. E., Li, L., Wu, J., Wang, T., Qiu, S., Zhang, J., Chen, J., Wu, R., Wang, S., Zhu, S., Chen, J., Zhang, W., Tang, X., Zhang, N., Chen, H., Cui, P., and Sachan, M. (2023). Agents: An open-source framework for autonomous language agents.
- Zhou, W., Ou, Y., Ding, S., Li, L., Wu, J., Wang, T., Chen, J., Wang, S., Xu, X., Zhang, N., Chen, H., and Jiang, Y. E. (2024). Symbolic learning enables self-evolving agents.

## A Agent Tree and Tape



**Figure 8:** A multi-agent tree configuration showing nodes (left) and a tape resulting from their collaboration (right) with color-coded steps: yellow for external agent thoughts (enabling collaboration), purple for internal agent thoughts, blue for actions, and green for observations. The step’s author is indicated in grey using the “Agent.node” format.

## B Tape Tools

**Raw Tape content**

Press Enter to rerender the tape

```
metadata:
id: 67103d7e-a354-4c3b-9c72-080019bf07fb
parent_id: c0e7a438-1a40-47f6-b7cb-1dbe00fd96e
author: ""
author_tape_id: null
n_added_steps: 7
error: null
result: null
context: null
steps:
- prompt_id: 8dca5402-bee0-4f1d-8b30-8e35016107f1
task: call
by: UserProxy
role: call
content: Make a plot comparing the stocks of ServiceNow and Salesforce since beginning of 2024. Save it to a PNG file.
agent_name: GroupChatManager
- prompt_id: 5bb44b5d-3bee-4a18-86dd-73c2dd17779
task: broadcast_last_message
by: UserProxy/GroupChatManager
content: Make a plot comparing the stocks of ServiceNow and Salesforce since beginning of 2024. Save it to a PNG file.
```

Pop N last steps

Press Enter to proceed

Keep N first steps

Press Enter to proceed

Load tape by id

Press Enter to load tape

Choose the tape renderer

full

**Metadata**

▶ Show / Hide

**Steps**

[1] **UserProxy calls GroupChatManager**  
Make a plot comparing the stocks of ServiceNow and Salesforce since beginning of 2024. Save it to a PNG file.

[2] **Thought: Broadcast**  
by: UserProxy/GroupChatManager  
content: Make a plot comparing the stocks of ServiceNow and Salesforce since beginning of 2024. Save it to a PNG file.  
from\_: UserProxy  
to:  
- CodeExecutor  
- SoftwareEngineer

▶ Prompt 494 characters

▶ Completion

[3] **GroupChatManager calls SoftwareEngineer**

▶ Prompt 2403 characters

▶ Completion

[4] **SoftwareEngineer responds to GroupChatManager**  
To help you plot and compare the stock prices of ServiceNow (ticker: NOW) and Salesforce (ticker: CRM) since the beginning of 2024, we need to:  
1. Gather the stock price data for both companies starting from January 1, 2024.  
2. Use Python (with pandas, yfinance, and matplotlib libraries) to process the data and create the plot.  
3. Save the plot to a PNG file.  
  
We will start by writing a Python script to accomplish these steps.

```
'''python
# filename: stock_comparison.py

import yfinance as yf
import pandas as pd
import matplotlib.pyplot as plt

# Define the stock tickers and date range
tickers = ['NOW', 'CRM']
start_date = '2024-01-01'
end_date = pd.Timestamp.today().strftime("%Y-%m-%d")
```

**Agent Org Chart**

Press Enter to rerender the org chart

```
UserProxy
- GroupChatManager
- SoftwareEngineer
- CodeExecutor
```

**Agent configuration**

Press Enter to update the agent

```
name: UserProxy
llms: []
subagents:
- name: GroupChatManager
llms:
default:
model_name: gpt-4o
parameters:
timeout: 15.0
context_size: 32000
tokenizer_name: ""
use_cache: false
subagents:
- name: SoftwareEngineer
llms:
default:
model_name: gpt-4o
parameters:
timeout: 15.0
context_size: 32000
tokenizer_name: ""
use_cache: false
subagents: []
templates:
system: "You are a helpful AI assistant.\nSolve tasks using your coding and"
and[
language skills.\nIn the following cases, suggest python code (in a python)
coding block) or shell script (in a sh coding block) for the user execute.\n\
1. When you need to collect info, use the code to output the info you\
need, for example, browse or search the web, download/read a file,
print\
```

Figure 9: **TapeAgents Studio:** Application to help AI Admin to edit Tape, resume and debug Agentic Systems

**Context**

**Steps**

**Question**  
If Eliud Kipchoge could maintain his record-making marathon pace indefinitely, how many thousand hours would it take him to run the distance between the Earth and the Moon at its closest approach? Please use the minimum perigee value on the Wikipedia page for the Moon when carrying out your calculation. Round your result to the nearest 1000 hours and do not use any comma separators if necessary.

► Prompt 2064 tokens

Agent Thought: Plan  
plan:  
- Look up Eliud Kipchoge's record marathon pace in minutes per mile or kilometers per hour.  
- Convert Eliud Kipchoge's marathon pace to a consistent unit (e.g., kilometers per hour).  
- Calculate the time it would take to run the distance between the Earth and the Moon at the closest approach using the converted pace.  
- Convert the time from hours to thousand hours.  
- Round the result to the nearest 1000 hours.  
- Provide the final answer.

► Prompt 2167 tokens

► Agent Thought: List Of Facts

► Prompt 2310 tokens

Agent Action: Search  
source: web\_search  
query: Eliud Kipchoge record marathon pace

► Observation: Search Results

► Prompt 2901 tokens

Agent Thought: New Fact  
fact\_name: eliud\_kipchoge\_marathon\_pace  
unit: min/km  
value: 2.5

► Prompt 2951 tokens

Agent Action: Search  
source: wikipedia  
query: minimum perigee value of the Moon

► Observation: Search Results

► Prompt 3466 tokens

Agent Thought: New Fact  
fact\_name: minimum\_perigee\_moon  
unit: km  
value: 362600

► Prompt 3510 tokens

► Agent Action: Convert Fact

Observation: Calculation Result  
name: eliud\_kipchoge\_speed  
result: 24.0

► Prompt 3689 tokens

► Agent Action: Use Calculator

Observation: Calculation Result  
name: time\_to\_run\_to\_moon  
result: 15180.333333333334

► Prompt 3841 tokens

► Agent Action: Use Calculator

Observation: Calculation Result  
name: time\_to\_run\_to\_moon\_thousand\_hours  
result: 15000.0

► Prompt 4018 tokens

Agent Action: Plan Finished  
success: True  
overview: Calculated the time it would take Eliud Kipchoge to run the distance between the Earth and the Moon at its closest approach.  
answer\_unit: thousand hours  
answer: 15000

Answer:  
answer: 15000  
success: True  
overview: Calculated the time it would take Eliud Kipchoge to run the distance between the Earth and the Moon at its closest approach.

**Reload Tapes**

**File**  
l1\_gpt4o-2024-05-13\_2024-07-02\_shortpron

**Accuracy 50.94%, 27 out of 53**

**Tape**  
L11: If Eliud Kipchoge could maintain

**Tape Result**  
Golden Answer: 17  
Agent Answer: 15000  
Steps: 17  
Failures: 0  
Finalized successfully: True  
LLM Call: 10  
Overview:  
Calculated the time it would take Eliud Kipchoge to run the distance between the Earth and the Moon at its closest approach.

**Figure 10: Tape Browser:** Application to inspect a batch of tapes result. This tape is a GAIA task where the Agent did not provide the right answer. No step failed during the session.

Accuracy: 50.94%	Accuracy: 41.51%
<p><b>Question:</b> If Eliud Kipchoge could maintain his record-making marathon pace indefinitely, how many thousand hours would it take him to run the distance between the Earth and the Moon its closest approach? Please use the minimum perigee value on the Wikipedia page for the Moon when carrying out your calculation. Round your result to the nearest 1000 hours and do not use any comma separators if necessary.</p>	<p><b>Question:</b> If Eliud Kipchoge could maintain his record-making marathon pace indefinitely, how many thousand hours would it take him to run the distance between the Earth and the Moon its closest approach? Please use the minimum perigee value on the Wikipedia page for the Moon when carrying out your calculation. Round your result to the nearest 1000 hours and do not use any comma separators if necessary.</p>
<p><b>Agent Thought: Plan</b> plan: - Lookup Eliud Kipchoge's record marathon pace in terms of speed (e.g., km/h). - Lookup the minimum perigee value (closest approach) of the Moon from the Wikipedia page for the Moon. - Convert the minimum perigee distance from kilometers to meters if necessary. - Calculate the time it would take for Eliud Kipchoge to run the distance between the Earth and the Moon at his marathon pace. - Convert the calculated time from hours to thousand hours. - Round the result to the nearest 1000 hours.</p>	<p><b>Agent Thought: Plan</b> plan: - Look up Eliud Kipchoge's record marathon pace in terms of speed (e.g., km/h). - <b>FindLookup</b> the minimum perigee value (closest approach) of the Moon to Earth from the Wikipedia page for the Moon. - Convert the minimum perigee distance from kilometers to hours using Kipchoge's marathon pacemeters if necessary. - Calculate the total time in hours <b>time</b> it would take Kipchoge to run the distance between the <b>for Eliud Kipchoge to run the distance between the Earth and the Moon at its closest approach his marathon pace</b>. - Convert the <b>totalcalculated</b> time from hours to thousand hours. - Round the result to the nearest 1000 hours.</p>
<p>► Agent Thought: List Of Facts</p>	<p>► Agent Thought: List Of Facts</p>
<p><b>Agent Action: Search</b> source: web_search query: Eliud Kipchoge marathon record pace km/h</p>	<p><b>Agent Action: Search</b> source: web_search query: Eliud Kipchoge marathon record pace km/h</p>
<p>► Observation: Search Results</p>	<p>► Observation: Search Results</p>

Figure 11: Tape Diff: Application to compares two batches of tape.

## C GREADTH Form Filler

### C.1 Virtual Companies Prompts

To generate synthetic environments/companies with 10 request forms available in each we used a three-step prompting method with Llama-3-70B.

1. The first step is to generate a description of a fake company. For this we use the following prompt:

```
messages = [
    {
        "role": "system",
        "content": "You are a helpful assistant."
    },
    {
        "role": "user",
        "content": f"Give me a description of {real_name} but replace all occurrences of
                    `'{real_name}` by `'{fake_name}`."
    }
]
```

with `real_name` and `fake_name` being set to different company names (e.g. “Starbucks” and “CofeeCorp” respectively). This step produces a `DESCRIPTION` variable that will be used in the next step.

2. The second step is to generate a list of 10 request forms for each fake company. For now, we only ask the model to generate a name and a description for each form with the following system and user prompts:

```
SYSTEM_MESSAGE = """
You are a helpful enterprise assistant who is very well integrated into the
internal system of [ENTERPRISE_NAME]. You have access to the database of
```

```
[ENTERPRISE_NAME], which contains REQUEST_FORMS. REQUEST_FORMS can be used  
by employees and clients to submit requests and trigger automations.  
REQUEST_FORMS have the following data format:  
```python  
class FunctionSchema(BaseModel):  
    name: FunctionName  
    description: str = Field(description="The description of the function.")  
    parameters: JsonSchema = Field(default=None, description="The JSON schema  
        of the function's parameters.")  
    return_value: JsonSchema = Field(default=None, description="The JSON  
        schema of the function's return value.")  
    ...  
    """  
    messages = [  
        {"role": "system", "content": SYSTEM_MESSAGE},  
        {"role": "user", "content": f"""  
            [ENTERPRISE_NAME] is '{fake_name}'.  
            {DESCRIPTION}  
            Give me the name and description of 10 REQUEST_FORMS that are often used  
            at {fake_name}. Make sure the REQUEST_FORMS are specific to {fake_name}  
            use cases. Format the output as a JSON list of JSON dictionaries with  
            'name' and 'description' keys. Make sure the output is JSON parsable.  
        """},  
    ]
```

with `fake_name` and `DESCRIPTION` defined in the previous step. The output is a list of 10 request form names and descriptions. We programmatically verify that the output is JSON parsable.

3. Eventually, for each request form dictionary (`FORM`) generated previously, we prompt llama3-70B to generate the `FunctionSchema` like this:

```
messages = [  
    {"role": "system", "content": SYSTEM_MESSAGE},  
    {"role": "user", "content": f"""  
        [ENTERPRISE_NAME] is '{fake_name}'.  
        {DESCRIPTION}  
        Show me the FunctionSchema of the following REQUEST_FORM:  
        ```json  
        {FORM}  
        ```  
  
        First, rephrase the description to be more specific. The description must include  
        rules, facts, and policies at {fake_name} about {FORM['name']}, such as when  
        to fill this request, who will process the request, etc... The description must  
        also include information about the parameters to help the user fill the form.  
        For multiple choice parameters (enum), the description must explain the  
        difference between each possible value.
```

Once the description is written, remove all colon ( : ) characters from it.

Second, write the yaml format of this `FunctionSchema` by replacing the description with your detailed version. Parameters cannot be nested objects, but return values can. The `FunctionSchema` must contain both required and optional parameters.  
Write the output in yaml format. The output must be parsable as a `FunctionSchema`.

```
"""},  
]
```

with `SYSTEM_MESSAGE`, `fake_name`, `DESCRIPTION`, and `FORM` defined in the previous steps. We verify programmatically that the output is parsable as a `FunctionSchema`. An example of generated `FunctionSchema` can be seen in Step [5] of the Tapes in Figures 12 & 13 (Appendix C.5).

## C.2 List of Form-Filler Agent Thoughts and Actions

**Form-filler thoughts.** Below, we list the available thoughts for our Form-Filler TapeAgent and explain when they should be used:

- `AnswerFromFunctionSchemaThought`: This thought premeditates answering the user's question that the agent knows how to answer based on the form description. Attributes:
  - `function`: The name of the function to answer a question about.
- `NoAnswerInFunctionSchemaThought`: This thought premeditates informing the user that the agent does not know the answer to their question. Attributes:
  - `function`: The name of the function.
- `RefuseInexistentFunctionThought`: A thought that indicates that the query could not be resolved to a function.
- `RefuseInvalidFunctionParameterValueThought`: This thought premeditates informing the user that a function parameter value is invalid. Attributes:
  - `function`: The name of the function.
  - `parameter`: The name of the parameter the user tried to set a value for.
  - `parameter_value`: The value the user tried to set for the parameter.
- `RefuseSkippingParameterThought`: This thought premeditates informing the user that a function parameter cannot be skipped because it is required. Attributes:
  - `function`: The name of the function.
  - `parameter`: The name of the required parameter that the user tried to skip.
- `GatherValuesThought`: This thought records extracted parameters from the user's message and is mainly used in the multi-step teacher prompter. Attributes:
  - `function`: The name of the function.
  - `parameters`: Dictionary mapping parameter names to their extracted values.
- `VerifyValuesThought`: This thought records the validity status of parameter values extracted from the user's message and is mainly used in the multi-step teacher prompter. Attributes:
  - `function`: The name of the function.
  - `parameters`: Dictionary mapping parameter names to their value, their validity, and the explanation when invalid.
- `UpdateFunctionParametersThought`: This thought updates the values of the parameters of a function based only on the user's last message. It is used when the user provides NEW information about the parameters of a function or when the user wants to skip an optional parameter. Attributes:
  - `function`: The name of the function.
  - `assign`: The dictionary assignment of parameter names to their NEW values (optional, default: `{}`). If no NEW values are provided, set 'assign' to `{}`.
  - `skip`: The list of NEW optional parameter names to skip (optional, default: `[]`). If no NEW parameters are skipped, set 'skip' to `[]`.

The ‘assign’ attribute sets or updates parameter values based on the last user message. It is only applied to parameters that are newly filled or have updated values. Parameters that remain unchanged will not have the ‘assign’ attribute. The ‘skip’ attribute is set for optional parameters when the user wants to ignore or skip an optional parameter. It is only applied to new optional parameters that the user wants to skip and will not be used for parameters that were already skipped.

- **RequestFunctionThought:** This thought premeditates requesting the user to select a function. Attributes:
  - functions: The list of available functions the user can select.
- **RequestFunctionParametersThought:** This thought premeditates requesting the user to provide a value for one parameter of a given function. Attributes:
  - function: The name of the function to request parameters for.
  - parameters: A list containing only one value which is the parameter to request.
- **RequestFunctionCallConfirmationThought:** This thought is a preliminary step before calling a function. This step is ONLY used once ALL parameters are filled or skipped. Attributes:
  - function: The name of the function.
- **RequestExitConfirmationThought:** This thought is a preliminary step before exiting the dialogue. The message MUST be short and concise. The message should contain clear and explicit confirmation of the acceptance of the last value provided by the user.

**Form-filler actions.** Here are the available actions our Form-Filler TapeAgent can take.

- **ResolveFunctionAction:** An action that resolves which function candidates are compatible with a given user query. Attributes:
  - query: The query to resolve the function with, by default the user’s last message.
  - result: A list of candidate functions that are compatible with the query.
- **InspectFunctionAction** An action that inspects the schema of a function given its name. This action inspects and returns the schema of the function, which contains its description and the names and types of the function’s parameters. Attributes:
  - function: The name of the function.
  - result: The schema of the function.
- **PromptUserForTextMessageAction:** An action that prompts the user to enter a message in a text input field. This action is used to capture free-form text input from the user through a standard text input field. This action must be used in every single agent’s response. Attributes:
  - prompt: The message from the agent to prompt the user to enter text. The message MUST be short and concise. The message should contain clear and explicit confirmation of the acceptance of the last value provided by the user.
  - result: User message
- **CallFunctionAction:** This action calls a function with all available parameter values. Attributes:
  - function: The name of the function to call.
  - result: Result of the function call
- **ExitAction:** With this action the agent indicates that the dialogue has ended. No further steps will be executed after this step. Attributes:
  - text: The message from the agent to inform the user that the dialogue has ended.

### C.3 User Agents

**Table 5:** The list of user agents that we implemented for all domains. The User LLM is a prompted Llama-3.1-405B-Instruct model.

Agent Name	Behavior
UserInitMessageAmazing	At the beginning of the conversation, the user should request a specific intent and provide values for some of the parameters in that request.
UserInitMessageShort	At the beginning of the conversation, the user should request a specific intent in a short message.
UserInitMessageAsk	At the beginning of the conversation, the user should ask the agent what it can do.
UserBadInitMessage	At the beginning of the conversation, the user should ask to do something impossible.
UserHappyPath	The user correctly answers the agent question.
UserMultislotInstruct	The user replies to the agent question and provides values for additional parameters in the request.
UserMultislotFuture	The user provides values for additional parameters in the request.
UserChangesSlot	The user changes the value to a previously set request parameter.
UserAdjectiveOrPosition	For categorical parameters, the user answers the agent but instead of giving the choice name, uses an adjective or a positional number that defines its choice.
UserSkipsOptional	For optional parameters, the user asks to skip it.
UserAsksAboutDocs	The user does not answer the agent, and instead asks a question that can be answered based on the request documentation.
UserAsksAboutFilledParameters	The user does not answer the agent, and instead asks a question about the parameters filled, or still missing.
UserAsksForHelp	For categorical parameters, the user ask the agent for help in choosing a value based on its problem.
UserCorrectValueInvalidValue	The user answers the agent correctly, but also gives an invalid value for a categorical parameter not yet filled.
UserInvalidValue	For categorical parameters, the user gives an impossible choice.
UserChangesTopic	The user does not answer the agent's question and change the topic of conversation instead.
UserSkipsUnskipable	For required parameters, the user asks to skip it.
UserIncorrectDate	For date/time parameters, the user gives an invalid date.
UserIncorrectWeb	For technical parameters such as url/email, the user gives a invalid value.

#### C.4 Choosing the teacher model

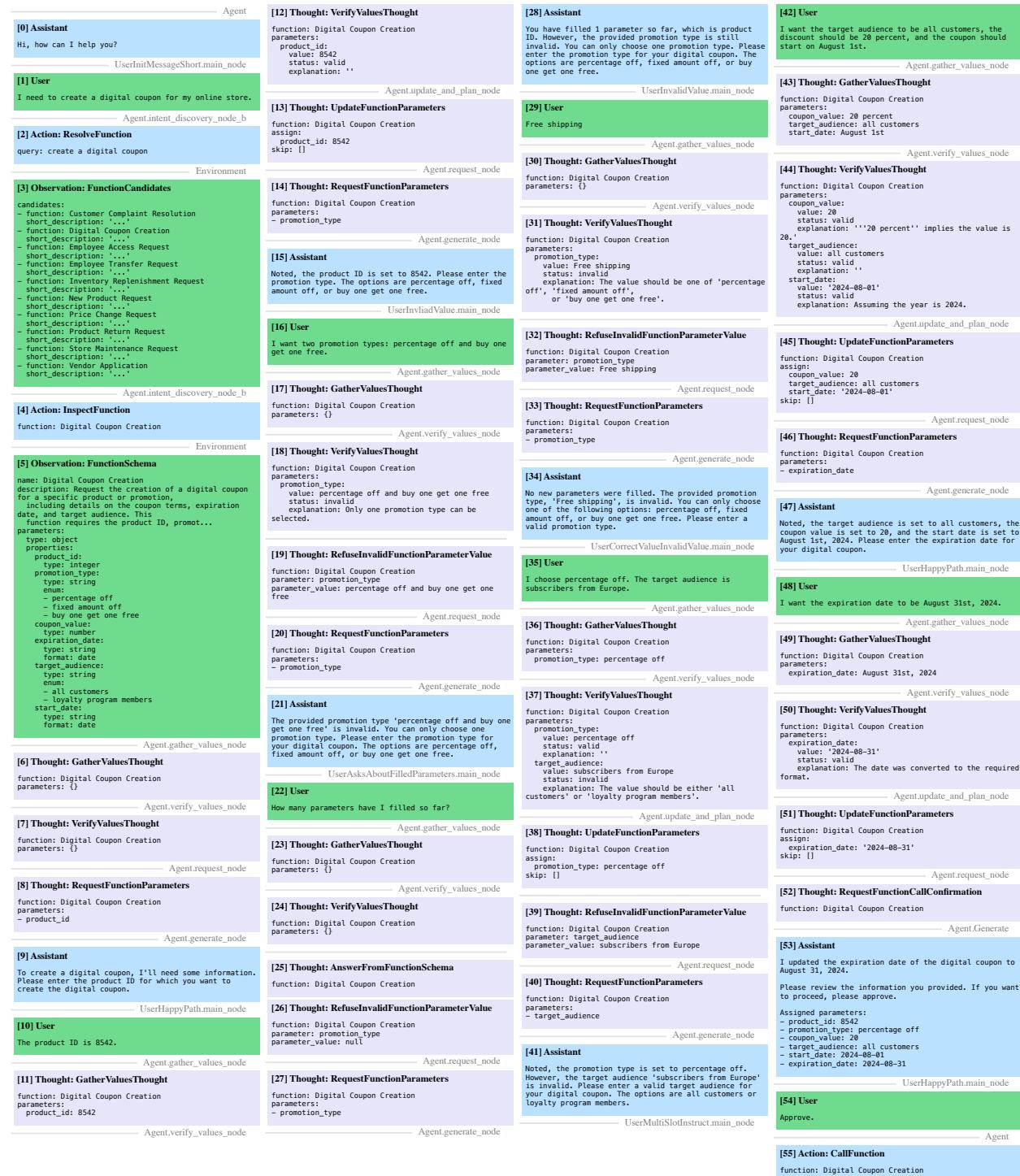
The teacher model is a Llama-405B-Instruct model<sup>7</sup> that is carefully prompted to perform the form-filling task according to all our metrics (Section 5.2). We experimented with various language models such as GPT4o, and Mistral-Large, but according to human evaluation, Llama 405B was on par or better than other models. We explored two different prompting techniques. The first (referred to as ‘simple’ in Table 6) consists of crafting one detailed paragraph explaining how the agent should make a thought plan and answer the user. The other prompting technique explored (referred to as ‘multistep’ in Table 6) consists in decomposing the task into multiple smaller subtasks. The agent is first tasked to identify potential parameter values provided by the last user message, then it is prompted to verify whether the provided values are correct or not, eventually, the agent is prompted to note valid information, refuse invalid information, and ask the user for the next parameter to fill. We generated agent messages as a continuation to 1016 unfinished conversations where the last user message behaved according to specific behaviors and sent these messages to human labelers. Results show that the multistep approach yields better performance overall, so we decided to keep this prompting strategy as our final Teacher model in Section 5.

**Table 6:** Evaluation of the teacher model with two prompting methods (simple & multistep) across all user behaviors on 1016 test domain conversations.

Author of last user message	Count	GREADTH metric	
		simple	multistep
UserInitMessageAmazing	46	60.87%	86.96%
UserInitMessageShort	48	87.50%	89.58%
UserInitMessageAsk	46	91.30%	84.78%
UserBadInitMessage	46	97.83%	100.00%
UserHappyPath	186	72.58%	89.25%
UserMultislotInstruct	48	68.75%	83.33%
UserMultislotFuture	46	58.70%	76.09%
UserChangesSlot	46	71.74%	82.61%
UserAdjectiveOrPosition	46	58.70%	76.09%
UserSkipsOptional	46	71.74%	71.74%
UserAsksAboutDocs	46	84.78%	93.48%
UserAsksAboutFilledParameters	48	83.33%	81.25%
UserAsksForHelp	45	40.00%	42.22%
UserCorrectValueInvalidValue	45	62.22%	77.78%
UserInvalidValue	46	63.04%	91.30%
UserChangesTopic	46	76.09%	86.96%
UserSkipsUnskipable	48	87.50%	95.83%
UserIncorrectDate	46	47.83%	19.57%
UserIncorrectWeb	42	16.67%	23.81%
<b>ALL_BEHAVIORS</b>	<b>1016</b>	<b>69.39%</b>	<b>78.54%</b>

<sup>7</sup><https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct-FP8>

## C.5 Teacher and Student Tapes



**Figure 12: Sample Tape between Teacher agent and User agent.** The user is requesting the creation of a digital coupon for their online store. Steps are color-coded: purple for internal agent thoughts, blue for actions, and green for observations. The step's author is indicated in grey using the “Agent.node” format. User agent names are described in Appendix C.3.

<p><b>[0] Assistant</b></p> <pre>Hi, how can I help you?</pre>	<p><b>[1] User</b></p> <pre>I need to create a digital coupon for my online store.</pre>	<p><b>[2] Action: ResolveFunction</b></p> <pre>query: create a digital coupon</pre>	<p><b>[3] Observation: FunctionCandidates</b></p> <pre>candidates: - function: Customer.ComplaintResolution short_description: '...' - function: Digital.CouponCreation short_description: '...' - function: EmployeeAccessRequest short_description: '...' - function: EmployeeTransferRequest short_description: '...' - function: ProductReplenishmentRequest short_description: '...' - function: NewProductRequest short_description: '...' - function: PriceChangeRequest short_description: '...' - function: ProductReturnRequest short_description: '...' - function: SystemMaintenanceRequest short_description: '...' - function: VendorApplication short_description: '...'</pre>	<p><b>[4] Action: InspectFunction</b></p> <pre>function: Digital.CouponCreation</pre>	<p><b>[5] Observation: FunctionSchema</b></p> <pre>name: Digital.CouponCreation description: 'Creates a digital coupon for a specific product or promotion, including details on the coupon terms, expiration date, and target audience. This function requires the product ID, promotion type, and target audience.'</pre> <pre>parameters:   - type: object     properties:       product_id: string       promotion_type: integer       target_audience: string       expiration_date: string       format: date</pre>
<p><b>[6] Thought: RequestFunctionParameters</b></p> <pre>function: Digital.CouponCreation parameters: - product_id</pre>	<p><b>[7] Assistant</b></p> <pre>To create a digital coupon, I'll need some information. Please enter the product ID for which you want to create the digital coupon.</pre>	<p><b>[8] User</b></p> <pre>The product ID is 8542.</pre>	<p><b>[9] Thought: UpdateFunctionParameters</b></p> <pre>function: Digital.CouponCreation assign:   product_id: 8542 skip: []</pre>	<p><b>[10] Thought: RequestFunctionParameters</b></p> <pre>function: Digital.CouponCreation parameters: - promotion_type</pre>	<p><b>[11] Assistant</b></p> <pre>Noted, the product ID is set to 8542. Please enter the promotion type. The options are percentage off, fixed amount off, or buy one get one free.</pre>
<p><b>[12] User</b></p> <pre>I want two promotion types: percentage off and buy one get one free.</pre>	<p><b>[13] Thought: RefuseInvalidFunctionParameterValue</b></p> <pre>function: Digital.CouponCreation parameter: promotion_type parameter_value: percentage off and buy one get one free</pre>	<p><b>[14] Thought: RequestFunctionParameters</b></p> <pre>function: Digital.CouponCreation parameters: - promotion_type</pre>	<p><b>[15] Assistant</b></p> <pre>The provided promotion type 'percentage off and buy one get one free' is invalid. You can only choose one promotion type. Please enter the promotion type for your digital coupon. The options are percentage off, fixed amount off, or buy one get one free.</pre>	<p><b>[16] User</b></p> <pre>How many parameters have I filled so far?</pre>	<p><b>[17] Thought: AnswerFromFunctionSchema</b></p> <pre>function: Digital.CouponCreation</pre>
<p><b>[18] Thought: RefuseInvalidFunctionParameterValue</b></p> <pre>function: Digital.CouponCreation parameter: promotion_type parameter_value: null</pre>	<p><b>[19] Thought: RequestFunctionParameters</b></p> <pre>function: Digital.CouponCreation parameters: - promotion_type</pre>	<p><b>[20] Assistant</b></p> <pre>You have filled 1 parameter so far, which is product ID. However, the provided promotion type is still invalid. You can only choose one promotion type. Please enter the promotion type for your digital coupon. The options are percentage off, fixed amount off, or buy one get one free.</pre>	<p><b>[21] User</b></p> <pre>Free shipping</pre>	<p><b>[22] Thought: RefuseInvalidFunctionParameterValue</b></p> <pre>function: Digital.CouponCreation parameter: promotion_type parameter_value: Free shipping</pre>	<p><b>[23] Thought: RequestFunctionParameters</b></p> <pre>function: Digital.CouponCreation parameters: - promotion_type</pre>
<p><b>[24] Assistant</b></p> <pre>No new parameters were filled. The provided promotion type, 'Free shipping', is invalid. You can only choose one of the following options: percentage off, fixed amount off, or buy one get one free. Please enter a valid promotion type.</pre>	<p><b>[25] User</b></p> <pre>I choose percentage off. The target audience is subscribers from Europe.</pre>	<p><b>[26] Thought: UpdateFunctionParameters</b></p> <pre>function: Digital.CouponCreation assign:   promotion_type: percentage off skip: []</pre>	<p><b>[27] Thought: RefuseInvalidFunctionParameterValue</b></p> <pre>function: Digital.CouponCreation parameter: target_audience parameter_value: subscribers from Europe</pre>	<p><b>[28] User</b></p> <pre>Approve.</pre>	<p><b>[29] Assistant</b></p> <pre>Assigned parameters: - product_id: 8542 - promotion_type: percentage off - coupon_value: 20 - target_audience: all customers - start_date: 2024-08-01 - expiration_date: 2024-08-31</pre>
<p><b>[30] User</b></p> <pre>I want the target audience to be all customers, the discount should be 20 percent, and the coupon should start on August 1st.</pre>	<p><b>[31] Thought: UpdateFunctionParameters</b></p> <pre>function: Digital.CouponCreation assign:   coupon_value: 20   target_audience: all customers   start_date: '2024-08-01' skip: []</pre>	<p><b>[32] Thought: RequestFunctionParameters</b></p> <pre>function: Digital.CouponCreation parameters: - expiration_date</pre>	<p><b>[33] Assistant</b></p> <pre>Noted, the target audience is set to all customers, the coupon value is set to 20, and the start date is set to August 1st, 2024. Please enter the expiration date for your digital coupon.</pre>	<p><b>[34] User</b></p> <pre>I want the expiration date to be August 31st, 2024.</pre>	<p><b>[35] Thought: UpdateFunctionParameters</b></p> <pre>function: Digital.CouponCreation assign:   expiration_date: '2024-08-31' skip: []</pre>
<p><b>[36] Thought: RequestFunctionCallConfirmation</b></p> <pre>function: Digital.CouponCreation</pre>	<p><b>[37] Assistant</b></p> <pre>I updated the expiration date of the digital coupon to August 31, 2024.</pre>	<p><b>[38] User</b></p> <pre>Approve.</pre>	<p><b>[39] Action: CallFunction</b></p> <pre>function: Digital.CouponCreation</pre>	<p>Agent</p>	

**Figure 13: Sample tape between Student agent and User agent.** The user is requesting the creation of a digital coupon for their online store. Steps are color-coded: purple for internal agent thoughts, blue for actions, and green for observations. The step's author is indicated in grey using the “Agent.node” format. User agent names are described in Appendix C.3.