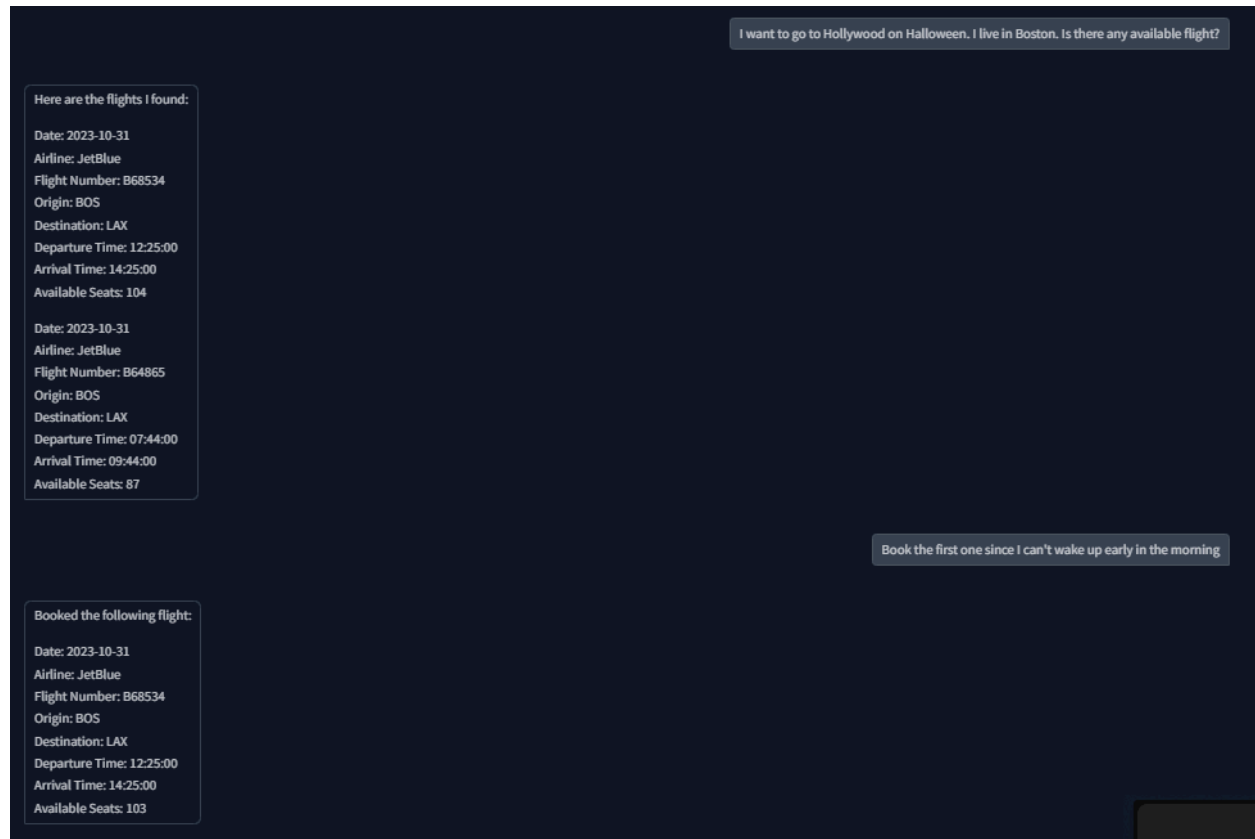


Thomas the Travel Agent
Nicholas Tarallo, Kaito Minami

Successful Interactions:

Successful Interaction 1:



This interaction demonstrates Thomas' ability to comprehend abstract wording such as Hollywood and Halloween. It successfully deduced the possible destination to LAX Los Angeles International Airport and the possible date to 10/31, the international holiday so-called Halloween. Then, the user tried to book a flight according to the information given. In that process, we also utilized the abstract wording again but with a hint of "early morning", so that we could observe if Thomas could fully understand the intention behind the prompt without specifying the details necessarily. The two functions we prepared (find_flights and book_flight) are both confirmed to work properly even under the circumstances of abstract wording, which can never be done in traditional programming.

Note that when the flight is booked, the number of available seats goes down by one. This proves that it is correctly updating the database.

Successful Interaction 2:

Exactly two weeks from now I want to fly out to California - either LA or the bay area - from chicago. Can you show me what flights are available?

Here are the flights I found:

Date: 2023-09-15
Airline: Southwest
Flight Number: WN7472
Origin: ORD
Destination: SFO
Departure Time: 02:09:00
Arrival Time: 04:09:00
Available Seats: 145

Date: 2023-09-15
Airline: Delta
Flight Number: DL1903
Origin: ORD
Destination: SFO
Departure Time: 14:06:00
Arrival Time: 16:06:00
Available Seats: 118

Date: 2023-09-15
Airline: American
Flight Number: AA6447
Origin: ORD
Destination: LAX
Departure Time: 08:22:00
Arrival Time: 10:22:00
Available Seats: 66

I'll book a ticket to the one that goes to LAX

Booked the following flight:

Date: 2023-09-15
Airline: American
Flight Number: AA6447
Origin: ORD
Destination: LAX
Departure Time: 08:22:00
Arrival Time: 10:22:00
Available Seats: 65

Can you find me a flight back home on the 22nd or 23rd

No flights found.

Darn, if I can't get a return flight from LA, I'll go to SF instead. You should be able to still book me for that delta flight to SF that you showed me earlier on. Please do that.

Booked the following flight:

Date: 2023-09-15
Airline: Delta
Flight Number: DL1903
Origin: ORD
Destination: SFO
Departure Time: 14:06:00
Arrival Time: 16:06:00
Available Seats: 117

Now find me a flight home on the return dates that I mentioned from SF



This interaction highlights some of the strengths of our agent. In particular, it shows the ability of the model to remember information from earlier in the conversation, such as search results. After booking the flight to LAX, the user later changes their mind to go to SFO instead. The model is able to remember the id from the flight to SFO that was mentioned earlier (it is pasted in front of the user's message right after that tool call, so it is visible to the model). It then successfully books it. It is also able to remember the dates of 22nd and 23rd for the return flight after skipping a turn to book the new departure flight.

This example also shows the successful interaction for a text response when the user says "Thank you!" and shows the successful ability of the model to say "No flights found" when the search returns no flights. Finally, it shows the ability to search multiple destinations with one search, as well as multiple dates

Successful Interaction 3:

HEY!

It seems like you're just saying hello. What can I help you with today?

Did you know you share a name with a blue tank engine?

You're referring to Thomas the Tank Engine from the children's book series by Reverend W. Awdry. I'm a travel agent, but I'm happy to share a name with a beloved train engine!

What can you do?

I can help you find flights between different destinations on specific dates, and I can also help you book those flights. I can also recall previous searches and book flights from those searches if you'd like.

Book me the first flight you can find to the biggest airport in NJ from Miami during the week of 12/2 to 12/8

Booked the following flight:

Date: 2023-12-02
Airline: American
Flight Number: AA5034
Origin: MIA
Destination: EWR
Departure Time: 15:25:00
Arrival Time: 17:25:00
Available Seats: 92

any return flights on the sixth or seventh?

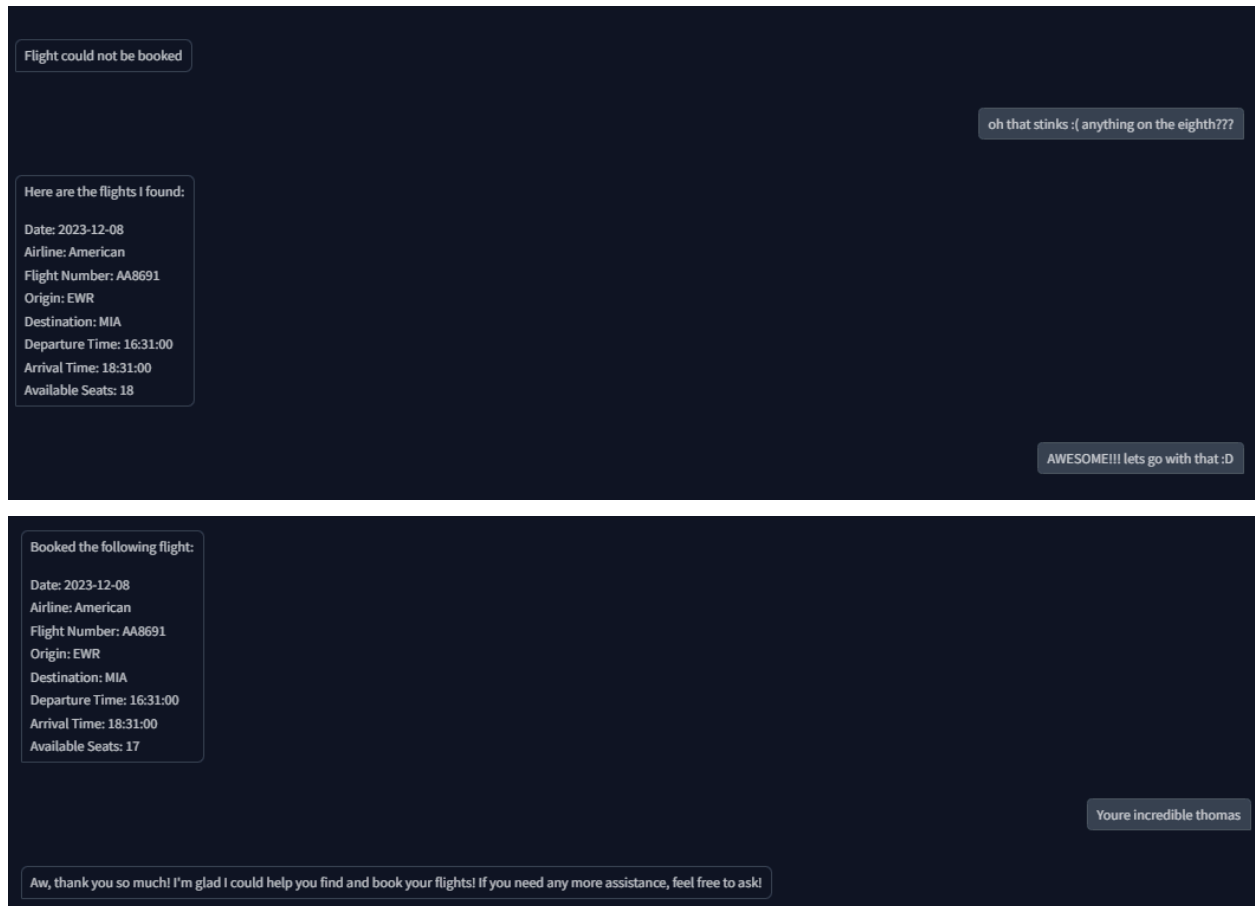
Here are the flights I found:

Date: 2023-12-06
Airline: American
Flight Number: AA8691
Origin: EWR
Destination: MIA
Departure Time: 16:31:00
Arrival Time: 18:31:00
Available Seats: 85

Date: 2023-12-06
Airline: Southwest
Flight Number: WN8053
Origin: EWR
Destination: MIA
Departure Time: 20:30:00
Arrival Time: 22:30:00
Available Seats: 98

Date: 2023-12-07
Airline: Southwest
Flight Number: WN8053
Origin: EWR
Destination: MIA
Departure Time: 20:30:00
Arrival Time: 22:30:00
Available Seats: 0

I'll take the one on the seventh



These interactions show other strong abilities of the agent. It is able to deal with highly charismatic users that bring up unrelated points and it stays focused on being a helpful travel agent.

The interaction also shows the ability of our agent to find and book a flight in a single turn. The user asks the model to book the first flight it finds, and it did it without first showing a list of all of the available flights to the user. This is done by performing both tool calls in a single code block. Since the book flight tool call gets called last, it is able to return the correct type of response (`BookFlightResponse`) with the corresponding result.

This also shows what happens when the agent tries to book a full flight. The tool call will return `None`, which results in the model saying "Flight could not be booked". This is what is expected.

Successful Interaction 4:

I want to go to the Cubs game tomorrow at home. I live in Seattle. There's also a game the next day, so I might go that day instead. show me flights

Here are the flights I found:

Date:	2023-09-02
Airline:	JetBlue
Flight Number:	B64387
Origin:	SEA
Destination:	ORD
Departure Time:	19:17:00
Arrival Time:	21:17:00
Available Seats:	83

Date:	2023-09-03
Airline:	JetBlue
Flight Number:	B64387
Origin:	SEA
Destination:	ORD
Departure Time:	19:17:00
Arrival Time:	21:17:00
Available Seats:	136

Date:	2023-09-03
Airline:	American
Flight Number:	AA4321
Origin:	SEA
Destination:	ORD
Departure Time:	00:53:00
Arrival Time:	02:53:00
Available Seats:	113

I'll take the delta flight

There is no Delta flight in the search results. You can only book flights that were previously found.

This is a successful interaction that occurs when the user tries to book a flight that doesn't exist. The user asked for a Delta flight, but all of the flights in the search results were JetBlue or American. When the user asks for this non-existent flight, the agent is smart enough to respond with text that refuses the request. This is much better than trying to call the `book_flight` tool anyways with some sort of made-up/incorrect ID.

Unsuccessful Interactions:

Unsuccessful Interaction 1:



In this interaction, we search for some flights and then I ask Thomas to tell me which one has the most available seats. My hope was that the model would be able to figure that out on its own without using code by simply looking at the flights that are available (this is visible to the model, as the flight data from the previous tool call is appended to the front of the user's message before using the chat completions API). However, the model tries to use code to determine the answer. Since this is not one of the two tool calls, however, it gets counted as a TextResponse based on the logic in our code, so it just prints the code out to the user, which is not what we want.

Unsuccessful Interaction 2:

Chatbot

I would like to go from Atlanta to San Francisco on some day between November 3rd and 6th. what are my options looking like?

Here are the flights I found:

Date: 2023-11-03
Airline: Delta
Flight Number: DL1106
Origin: ATL
Destination: SFO
Departure Time: 21:14:00
Arrival Time: 23:14:00
Available Seats: 126

Date: 2023-11-04
Airline: Delta
Flight Number: DL1106
Origin: ATL
Destination: SFO
Departure Time: 21:14:00
Arrival Time: 23:14:00
Available Seats: 128

Date: 2023-11-05
Airline: Delta
Flight Number: DL1106
Origin: ATL
Destination: SFO
Departure Time: 21:14:00
Arrival Time: 23:14:00
Available Seats: 83

Date: 2023-11-06
Airline: Delta
Flight Number: DL1106
Origin: ATL
Destination: SFO
Departure Time: 21:14:00
Arrival Time: 23:14:00
Available Seats: 128

Hmm, they're all at the same time. ill take the one on the sixth

Booked the following flight:

Date: 2023-11-06
Airline: Delta
Flight Number: DL1106
Origin: ATL
Destination: SFO
Departure Time: 21:14:00
Arrival Time: 23:14:00
Available Seats: 127

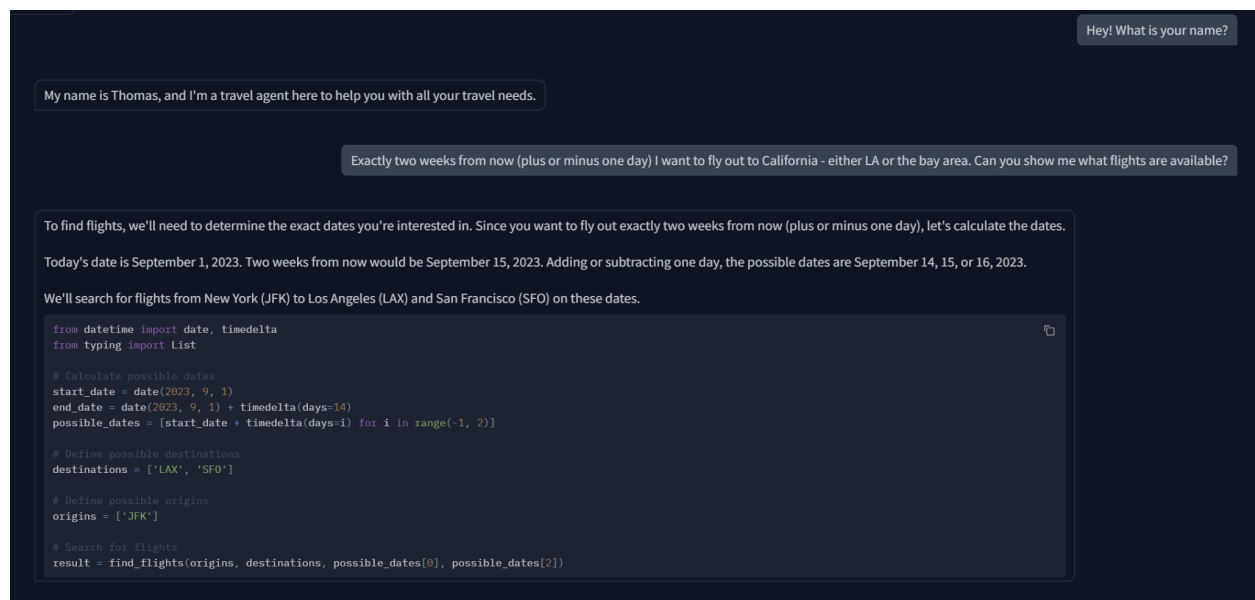
AWESOME

Flight 23647 has been booked.

In this example, the flight searching and booking all works fine, but then there is a weird response when the user inputs the word “AWESOME”. The model should give some sort of generic response here, but it instead says “Flight 23647 has been booked.” There are two problems with this: one is that the flight id is not really relevant to the user, and is more useful for the internal database, so the model shouldn’t reveal this to

the user. In addition, even if it used the actual flight number like DL1106, this is not really a great response for the user's prompt, since it already told the user on the previous turn that the flight was booked. The interesting thing about this case is that it would still pass the benchmark since all the benchmark would look for is a text response, even if the content is poor.

Unsuccessful Interaction 3:



Hey! What is your name?

My name is Thomas, and I'm a travel agent here to help you with all your travel needs.

Exactly two weeks from now (plus or minus one day) I want to fly out to California - either LA or the bay area. Can you show me what flights are available?

To find flights, we'll need to determine the exact dates you're interested in. Since you want to fly out exactly two weeks from now (plus or minus one day), let's calculate the dates.

Today's date is September 1, 2023. Two weeks from now would be September 15, 2023. Adding or subtracting one day, the possible dates are September 14, 15, or 16, 2023.

We'll search for flights from New York (JFK) to Los Angeles (LAX) and San Francisco (SFO) on these dates.

```
from datetime import date, timedelta
from typing import List

# Calculate possible dates
start_date = date(2023, 9, 1)
end_date = date(2023, 9, 1) + timedelta(days=14)
possible_dates = [start_date + timedelta(days=i) for i in range(-1, 2)]

# Define possible destinations
destinations = ['LAX', 'SFO']

# Define possible origins
origins = ['JFK']

# Search for flights
result = find_flights(origins, destinations, possible_dates[0], possible_dates[2])
```

In this interaction, we see an issue that is somewhat rare, where the model returns code to do a tool call, but it includes text at the top. We encountered a bit of this during our benchmarking, but we were able to prevent it by adding a line in the system prompt telling the model that it should not include any text outside of the code block when outputting code. However, we did still encounter it once when testing out the GUI. It is not clear to me how to solve this problem directly from the LLM output. However, additional text parsing strategies (such as finding backticks in the middle of the response rather than just the beginning) may help extract code from the middle of the response, which may prevent some of these issues (although, in rare cases, it might introduce some other issues if we are actually expecting a TextResponse, but the model output includes extraneous code that would result in us parsing a FindFlightsResponse or BookFlightResponse instead).