CHAPTER 8

RESULTS

8.1 Automatic Speech Recognition (ASR)

Word Error Rate (WER) is a metric that compares the accuracy of transcripts produced by speech recognition APIs. It's the ratio of errors in a transcript to the total words spoken, with a lower WER representing better accuracy in recognizing speech. The formula for calculating WER is:

$$\text{WER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Number of Words Spoken}}$$

S: stands for substitutions
I: stands for insertions
D: stands for deletions
N: is the number of words in the reference

$$\text{WER} = \frac{\text{S} + \text{I} + \text{D}}{\text{N}}$$

$$\text{WER} = \frac{5 + 2 + 1}{146}$$

$$\text{WER} = 0.0547$$

$$\text{Accuracy} = (1 - 0.0547) * 100$$

$$\text{Accuracy} = 94.53\%$$

Serial No.	Spoken Speech	Transcription Result	Error Type
1	I want to book a flight from New York to Los Angeles	I want to book a flight from New York Los Angeles	Deletion
2	Show me hotels near the Eiffel Tower in Paris	Show me hotels near the Eiffel Tower in Paris	Correct
3	What are the top attractions in Rome?	What are the top attractions in room?	Substitution
4	Find me a restaurant with good ratings in London	Find me a restaurant with good ratings in London	Correct
5	How far is it from Chicago to San Francisco?	How far is it from Chicago to San Francisco	Deletion
6	Reserve a rental car for pickup at JFK Airport	Reserve a rental car for pick up at a JFK Airport	Insertion
7	Book a hotel room for two nights in Tokyo	Book a hotel room for two nights in Tokyo	Correct

Table 8.1: Speech Transcription Results for Travel Buddy Use Case (Part 1)

Serial No.	Spoken Speech	Transcription	Error Type
8	Is there a direct train from Berlin to Munich?	Is there a direct train from Berlin to Munich?	Correct
9	Can you suggest some budget-friendly accommodations in Amsterdam?	Can you suggest some budget friendly accommodations in Amsterdam	Substitution
10	What's the weather like in Sydney next week?	What's the weather like in Sydney next week?	Correct
11	Show me the best travel deals to Bali	Show me the best travel deals to Bali	Correct
12	I need to change my flight reservation to next Friday	I need to change my flight reservation to next Friday	Correct
13	What are the must-visit places in Barcelona?	What are the must visit places in Barcelona	Substitution
14	How can I get from Heathrow Airport to central London?	How can I get from Heathrow Airport to central London	Correct
15	Can you recommend a reliable taxi service in Rome?	Can you recommend a reliable taxi service in Rome	Correct

Table 8.2: Speech Transcription Results for Travel Buddy Use Case (Part 2)

8.2 Travel Planning

8.2.1 Dataset Used:

TravelPlanner by osunlp
Link: https://huggingface.co/datasets/osunlp/TravelPlanner

8.2.2 Dataset Information:

In TravelPlanner, for a given query, language agents are expected to formulate a comprehensive plan that includes transportation, daily meals, attractions, and accommodation for each day. TravelPlanner comprises 1,225 queries in total. The number of days and hard constraints are designed to test agents' abilities across both the breadth and depth of complex planning.

8.2.3 Types of Constraints in TravelPlanner:

1. Environment Constraint
2. Commonsense Constraint
3. Hard Constraint

Tool	**Data Entries (#)**
CitySearch	312
FlightSearch	3,827,361
DistanceMatrix	17,603
RestaurantSearch	9,552
AttractionSearch	5,303
AccommodationSearch	5,064

Table 8.3: Number of Data Entries for Different Tools in the Dataset

8.2.4 Evaluation Modes

1. **Two-stage Mode**
   In the two-stage mode, language agents are tasked to with employing various search tools to gather information. Based on the collected information, language agents are expected to deliver a plan that not only meet the user's needs specified in the query but also adheres to commonsense constraints.

2. **Sole-Planning Mode**
   TravelPlanner also provides an easier mode solely focused on testing their planning ability. The sole-planning mode ensures that no crucial information is missed, thereby enabling agents to focus on planning itself.

8.2.5 Evaluation Metrics

1. **Delivery Rate:**
   This metric assesses whether agents can successfully deliver a final plan within a limited number of steps. Falling into dead loops, experiencing numerous failed attempts, or reaching the maximum number of steps (30 steps in our experimental setting) will result in failure.

2. **Commonsense Constraint Pass Rate:**
   Comprising eight commonsense dimensions, this metric evaluates whether a language agent can incorporate commonsense into their plan without explicit instructions.

3. **Hard Constraint Pass Rate:**
   This metric measures whether a plan satisfies all explicitly given hard constraints in the query, which aims to test the agents' ability to adapt their plans to diverse user needs.

4. **Final Pass Rate:**
   This metric represents the proportion of feasible plans that meet all aforementioned constraints among all tested plans. It serves as an indicator of agents' proficiency in producing plans that meet a practical standard.

8.2.6 Evaluation Strategies

1. **Micro Pass Rate**
   The micro strategy calculates the ratio of passed constraints to the total number of constraints.

2. **Macro Pass Rate**
   The macro strategy calculates the ratio of plans that pass all commonsense or hard constraints among all tested plans.

8.2.7 Validation Results: Two-Stage

Model	Tool-use Strategy	Planning Strategy	Organization	Delivery Rate
Human	Human	Human	TravelBuddy Team	100
gpt-3.5-turbo	ReAct	Direct	TravelBuddy Team	86.67
Mistral-7B	ReAct	Direct	TravelBuddy Team	8.89
Gemini Pro	Tips	Direct	TravelBuddy Team	85
gpt-4	ReAct	Direct	TravelBuddy Team	89.44
Gemini Pro	ReAct	Direct	TravelBuddy Team	28.89

Table 8.4: Validation Results: Two-Stage (Part 1)

Commonsense Constraint Micro Pass Rate	Commonsense Constraint Macro Pass Rate	Hard Constraint Micro Pass Rate	Hard Constraint Macro Pass Rate	Final Pass Rate
100	100	100	100	100
53.96	0	0	0	0
5.9	0	0	0	0
56.6	3.89	8.57	3.89	1.11
61.11	2.78	15.24	10.56	0.56
18.82	0	0.48	0.56	0

Table 8.5: Validation Results: Two-Stage (Part 2)

8.2.8 Validation Results: Sole-Planning

Model	Tool-use Strategy	Planning Strategy	Organization	Delivery Rate
Human	Human	Human	TravelBuddy Team	100
gpt-3.5	-	Reflexion	TravelBuddy Team	93.89
gpt-3.5	-	CoT	TravelBuddy Team	100
gpt-3.5	-	Direct	TravelBuddy Team	100
gpt-3.5	-	ReAct	TravelBuddy Team	82.22
gpt-4	-	Direct	TravelBuddy Team	100

Table 8.6: Validation Results: Sole-Planning (Part 1)

Commonsense Constraint Micro Pass Rate	Commonsense Constraint Macro Pass Rate	Hard Constraint Micro Pass Rate	Hard Constraint Macro Pass Rate	Final Pass Rate
100	100	100	100	100
53.75	2.78	10.95	2.78	0
66.32	3.33	11.9	5	0
60.21	4.44	10.95	2.78	0
47.64	3.89	11.43	6.67	0.56
80.42	17.22	47.14	22.22	4.44

Table 8.7: Validation Results: Sole-Planning (Part 2)

8.2.9 Test Results: Two-Stage

Model	Tool-use Strategy	Planning Strategy	Organization	Delivery Rate
Human	Human	Human	TravelBuddy Team	100
gpt-4	ReAct	Direct	TravelBuddy Team	93.1
gpt-3.5	ReAct	Direct	TravelBuddy Team	91.8
Mistral-7B	ReAct	Direct	TravelBuddy Team	7
Gemini Pro	ReAct	Direct	TravelBuddy Team	39.1

Table 8.8: Test Results: Two-Stage (Part 1)

Commonsense Constraint Micro Pass Rate	Commonsense Constraint Macro Pass Rate	Hard Constraint Micro Pass Rate	Hard Constraint Macro Pass Rate	Final Pass Rate
100	100	100	100	100
63.25	2	10.52	5.5	0.6
57.86	0	0.52	0.6	0
4.81	0	0	0	0
24.88	0	0.57	0.1	0

Table 8.9: Test Results: Two-Stage (Part 2)

8.2.10 Test Results: Sole-Planning

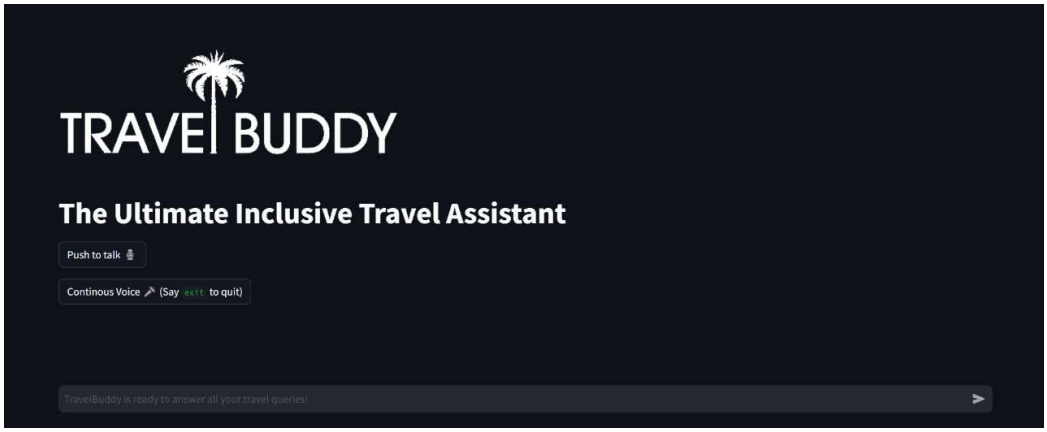
Model	Tool-use Strategy	Planning Strategy	Organization	Delivery Rate
Human	Human	Human	TravelBuddy Team	100
gpt-3.5	-	Reflexion	TravelBuddy Team	92.1
gpt-3.5	-	CoT	TravelBuddy Team	100
gpt-3.5	-	Direct	TravelBuddy Team	100
gpt-3.5	-	ReAct	TravelBuddy Team	81.6
gpt-4	-	Direct	TravelBuddy Team	100

Table 8.10: Test Results: Sole-Planning (Part 1)

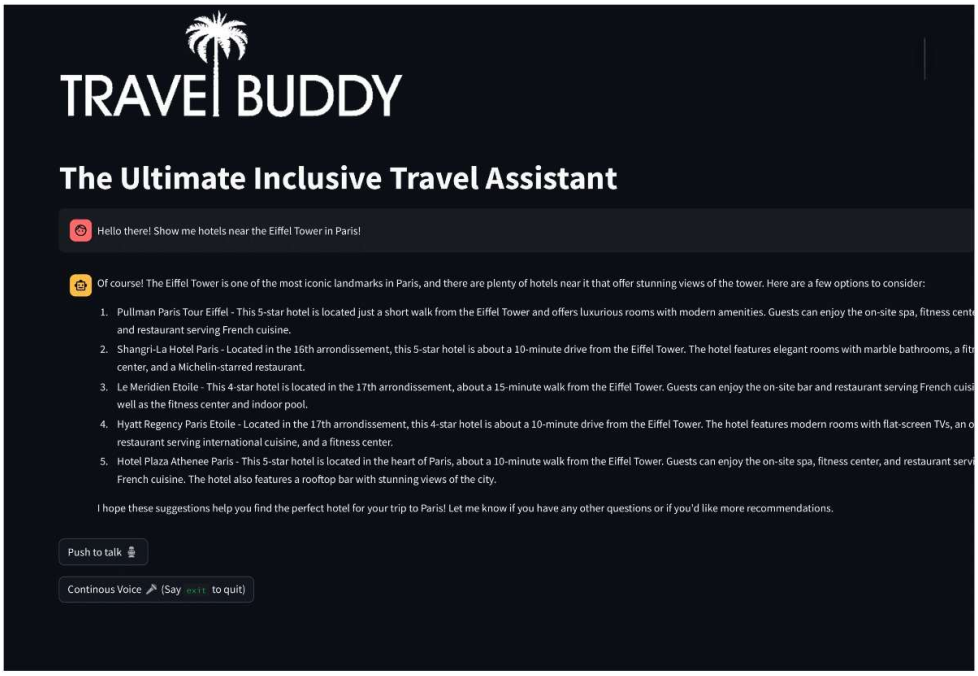
Commonsense Constraint Micro Pass Rate	Commonsense Constraint Macro Pass Rate	Hard Constraint Micro Pass Rate	Hard Constraint Macro Pass Rate	Final Pass Rate
100	100	100	100	100
52.05	2.2	9.91	3.8	0.4
64.38	2.3	9.83	3.8	0.4
59.49	2.7	9.48	4.4	0.6
45.85	2.5	10.66	3.1	0.7
80.55	15.2	44.28	23.1	4.4

Table 8.11: Test Results: Sole-Planning (Part 2)

CHAPTER 9
SCREENSHOTS



(a)



(b)

Figure 9.1: Front End Interface