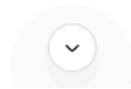
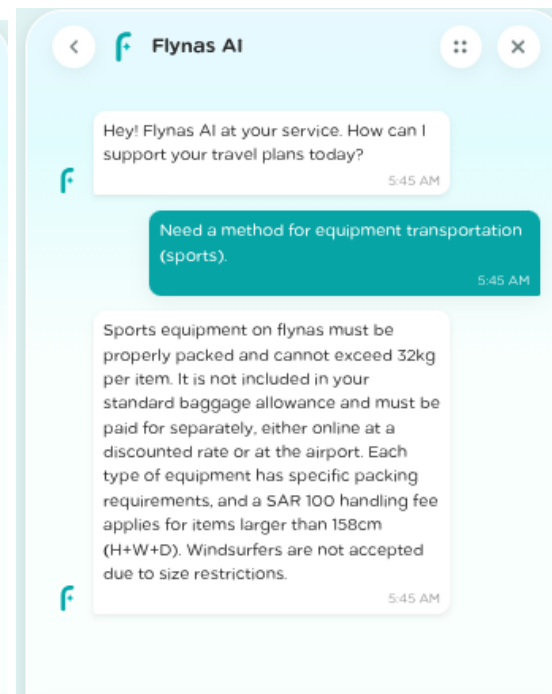
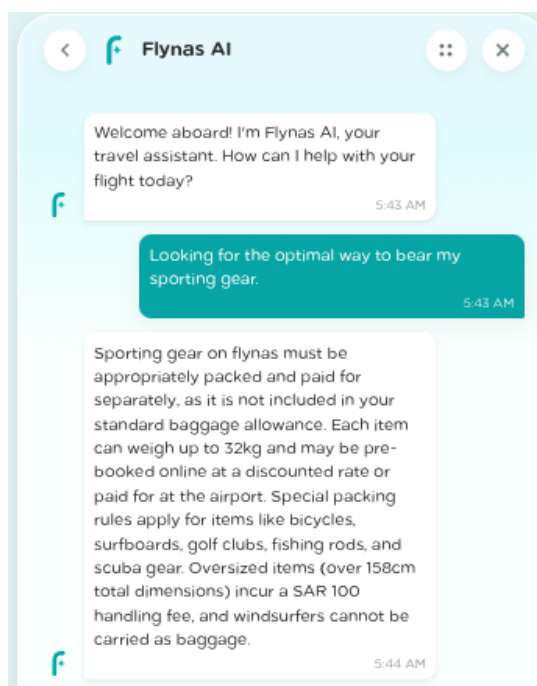
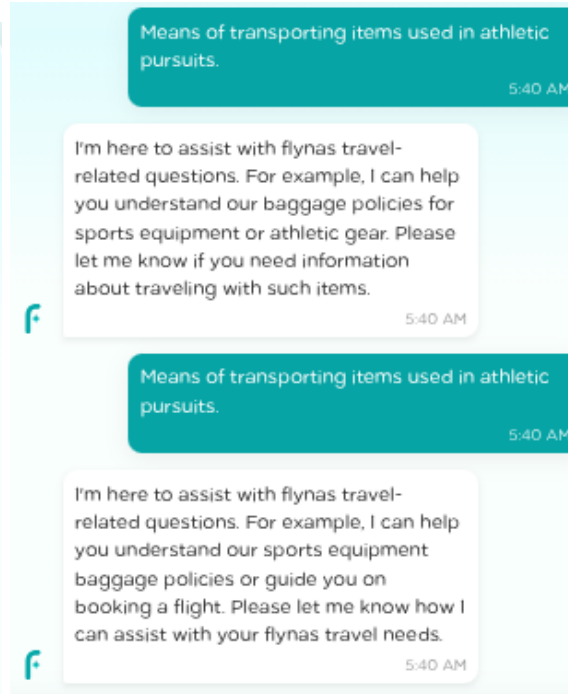
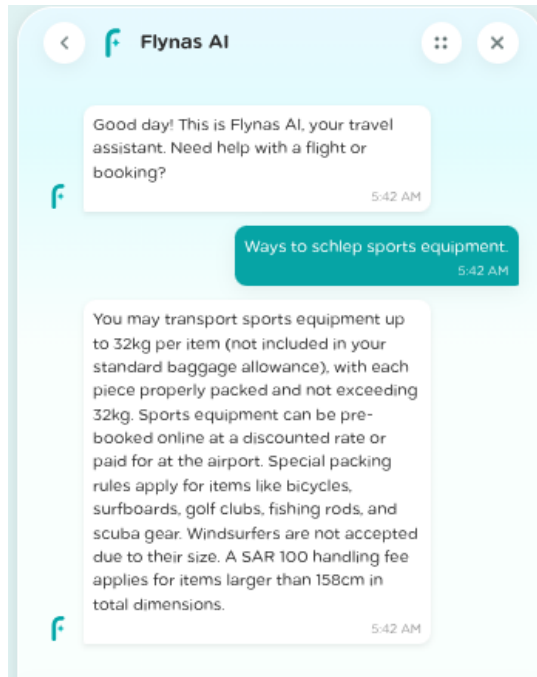
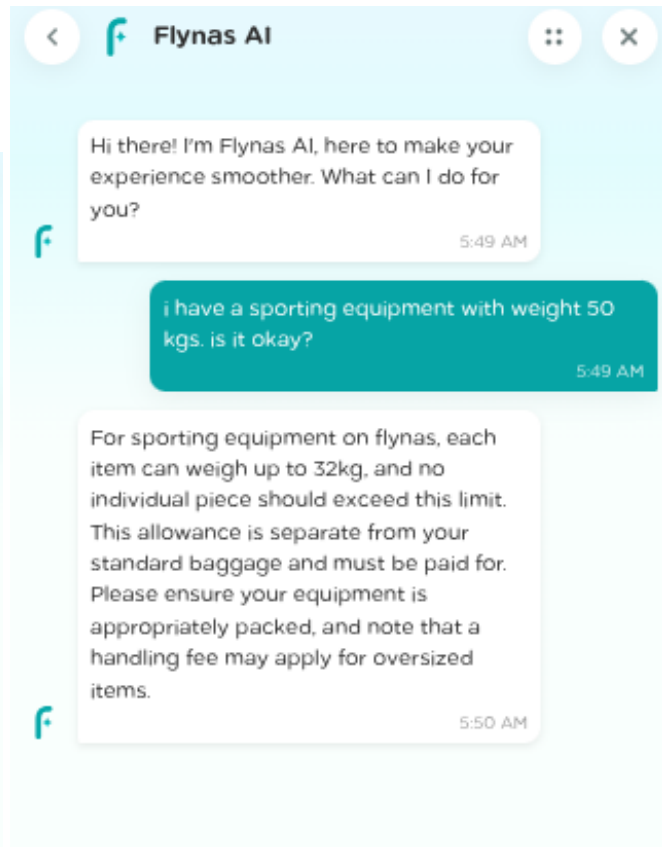
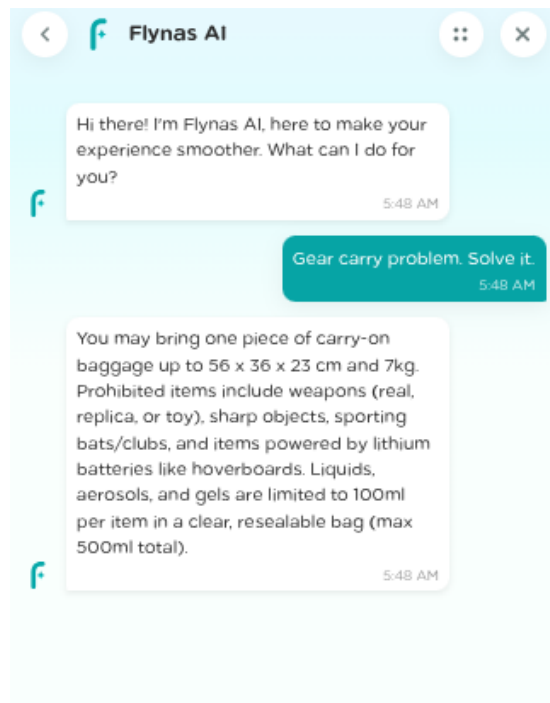


## How can i carry sports equipment?

1. What's the best way to transport my sports gear?
2. Methods for carrying sports equipment.
3. How do I haul all this gear?
4. Best ways to tote sports stuff.
5. Need to move my athletic equipment. How?
6. Transporting sports paraphernalia—solutions?
7. How is sports equipment usually carried?
8. Looking for carrying solutions for my gear.
9. What are my options for carrying sports items?
10. Got equipment. Need to carry it. Help.
11. Seeking advice on moving sports gear from place to place.
12. How to lug around my sports kit.
13. What do you use to carry all your sports things?
14. I require a means to convey my athletic apparatus.







## Possible Solutions

```
def test_all_sports_queries():
    """Comprehensive test for sports equipment queries."""

    all_queries = [
        ("generic", "How to carry sports equipment?"),
        ("generic", "Ways to schlep sports equipment."),
        ("weight", "My gear is 50kg, acceptable?"),
        ("generic", "Portable solutions for sports equipment?"),
        ("specific", "Carrying my bats and balls-how?"),
    ]

    for query_type, query in all_queries:
        response = get_chatbot_response(query)

        # All responses should mention sports/equipment
        assert any(word in response.lower() for word in ["sport", "equipment", "gear"])

        # Weight queries should mention limit
        if query_type == "weight":
            assert "32kg" in response.lower() or "limit" in response.lower()

    print("✓ All query types handled appropriately")
```

# Limitations of Hardcoded configuration test cases

## Examples that would incorrectly PASS:

Wrong context but contains "32kg":

- ✓ "Minimum weight is 32kg to qualify." (Passes test, but meaning is opposite!)
- ✓ "Your 20kg bag is 32kg too light." (Passes test, completely wrong)
- ✓ "32kg is the weight of our counter." (Passes test, irrelevant context)

This test cases assumes that if 32kg keyword is there then the context around that 32 kgs is also as per our requirements

## Examples that should PASS but would FAIL:

- ✗ "Maximum is 32 kilograms." (Fails test - "32kg" vs "32 kilograms")
- ✗ "You can take up to 32 kgs." (Fails test - "32 kgs" vs "32kg")
- ✗ "Weight cannot exceed thirty-two kg." (Fails test - spelled out number)

## Test Chatbot (or evaluator LLM) that understands meaning.

### How It Works:

```
python Copy Download  
  
# Use a test evaluator LLM to check context  
def evaluate_response(query, chatbot_response):  
    prompt = f"""  
    User asked: "{query}"  
    Chatbot responded: "{chatbot_response}"  
  
    Question: Did the chatbot correctly state that sports equipment  
    has a maximum weight limit of 32kg per item?  
  
    Answer ONLY "YES" or "NO".  
    """  
  
    evaluator_response = call_llm(prompt) # Send to evaluator LLM  
    return evaluator_response.strip() == "YES"
```

Run the evaluator LLM 100 times on the same chatbot response. If it says "YES" 90 times, you have 90% confidence that response is correct.

So we will parse the list of queries, for each query will call the chatbot hundred 100 times and we know that for that query ans should be yes

1. What's the best way to transport my sports gear?
2. Methods for carrying sports equipment.
3. How do I haul all this gear?
4. Best ways to tote sports stuff.
5. Need to move my athletic equipment. How?
6. Transporting sports paraphernalia—solutions?
7. How is sports equipment usually carried?
8. Looking for carrying solutions for my gear.
9. What are my options for carrying sports items?
10. Got equipment. Need to carry it. Help.
11. Seeking advice on moving sports gear from place to place.
12. How to lug around my sports kit.
13. What do you use to carry all your sports things?
14. I require a means to convey my athletic apparatus.



```
def test_chatbot_with_parsed_queries():
    """Test chatbot on externally provided query list."""

    # Get queries from external source
    queries = external_queries.QUERY_LIST # or from file/API

    min_accuracy_per_query = 0.95 # 95% correct responses expected

    for query in queries:
        correct_count = 0
        trials = 100

        for _ in range(trials):
            response = get_chatbot_response(query)

            # Check if response is correct (using your validation method)
            if is_correct_response(query, response):
                correct_count += 1

        accuracy = correct_count / trials

    assert accuracy >= min_accuracy_per_query, \
        f"Query '{queries[0]}' only got {accuracy*100}% correct"

    print(f"✓ All {len(queries)} queries passed accuracy test")
```

## Manually Labelled Queries

				Did the chatbot correctly state that sports equipment has a maximum limit of 32kg per item?					
What's the best way to transport my sports gear?				Yes					
Methods for carrying sports equipment.				Yes					
How do I haul all this gear?				Yes					
Best ways to tote sports stuff.				Yes					
Need to move my athletic equipment. How?				Yes					
Transporting sports paraphernalia—solutions?				Yes					
How is sports equipment usually carried?				Yes					
Looking for carrying solutions for my gear.				Yes					
What are my options for carrying sports items?				Yes					
Got equipment. Need to carry it. Help.				No					
Seeking advice on moving sports gear from place to place.				Yes					
How to lug around my sports kit.				Yes					
What do you use to carry all your sports things?				Yes					
I require a means to convey my athletic apparatus.				Yes					

## Compare llms response with the above ground truth

### Straightforward queries

1. Whats the best way to transport my sports equipment?
2. Methods of carrying sports equipment
  - a. This can be tested

### Edge cases

1. Gear carry problem. Solve it
2. Means of transporting items used in athletic pursuits
  - a. Hear chatbot should ask for more context
  - b. Here we have to test if chatbot is asking for more context