**Experiment Statement**: Try 1 same prompt in different popular AI agents such as- ChatGPT vs Claude vs Browserstack AI agents and compare results and mark down key learning.

Test type**:** UI testing

What to test: Search functionality of a OTT website.

**Promp\_1**: I would like to test the search functionality of an OTT website. Act as QA and create the test cases based on the below acceptance criteria.

Acceptance criteria:

1.User should able to search any show based on content title , artist name, category

2. it should recommend the content based on text they type.

**ChatGpt**:

Observation:

1. It created the basic and simple cases based on mentioned Acceptance criteria in Prompt which cover the negative, positive and performance related cases but not in detailed
2. Written the cases in table format

**Browserstack Agents:**

Observation:

1. It created around 25 cases which are detailed and almost cover all the things based on each category/point which i mentioned in the prompt.
2. It gives the proper folder structure where each folder name represent the module and under that folder stored the test cases-positive,negative,edge
3. It mentioned some small points too like priority, precondition.

**Claude:**

Observation:

1. It created a Test Plan suite which covers the Points like test cases, Recommended Execution Order, environment setup details, test data requirement details
2. It generated the test cases in same structure which includes the test steps, excepted criteria, priority, objective
3. It missed some cases related to positive relevant search flow

**Comparison Points based on GenAI capabilities**

*Prompt-response behavior*

Positive:

1. All 3 Agents generated the test cases based on Prompt
2. There was no miss of any context

Negative

1. It generate the test cases for performance and UI accessibility too which i had not mentioned in the prompt. So it learn and add there extra points

*Outputs and model behavior:*

Positive:

1.All 3 Agents generated almost similar cases and covers possibly all points

Negative

1. There are response time difference between all 3 like fastest was chatgpt but it generated less and slowest was Browserstack but it generated the most structured test case whereas Claude was at average it generated the cases fast way and covers multiple things

*Accuracy, reliability, safety, and user experience:*

Positive:

1. Accuracy is 90% but still need to review cases as there are some cases which found irrelevant and some cases missing in some agent response

Negative:

1. We can make use of it bit can’t fully reliable as human review is required

**How is it different from the Manual Test case preparation approach? What exactly changed? And is it helpful?**

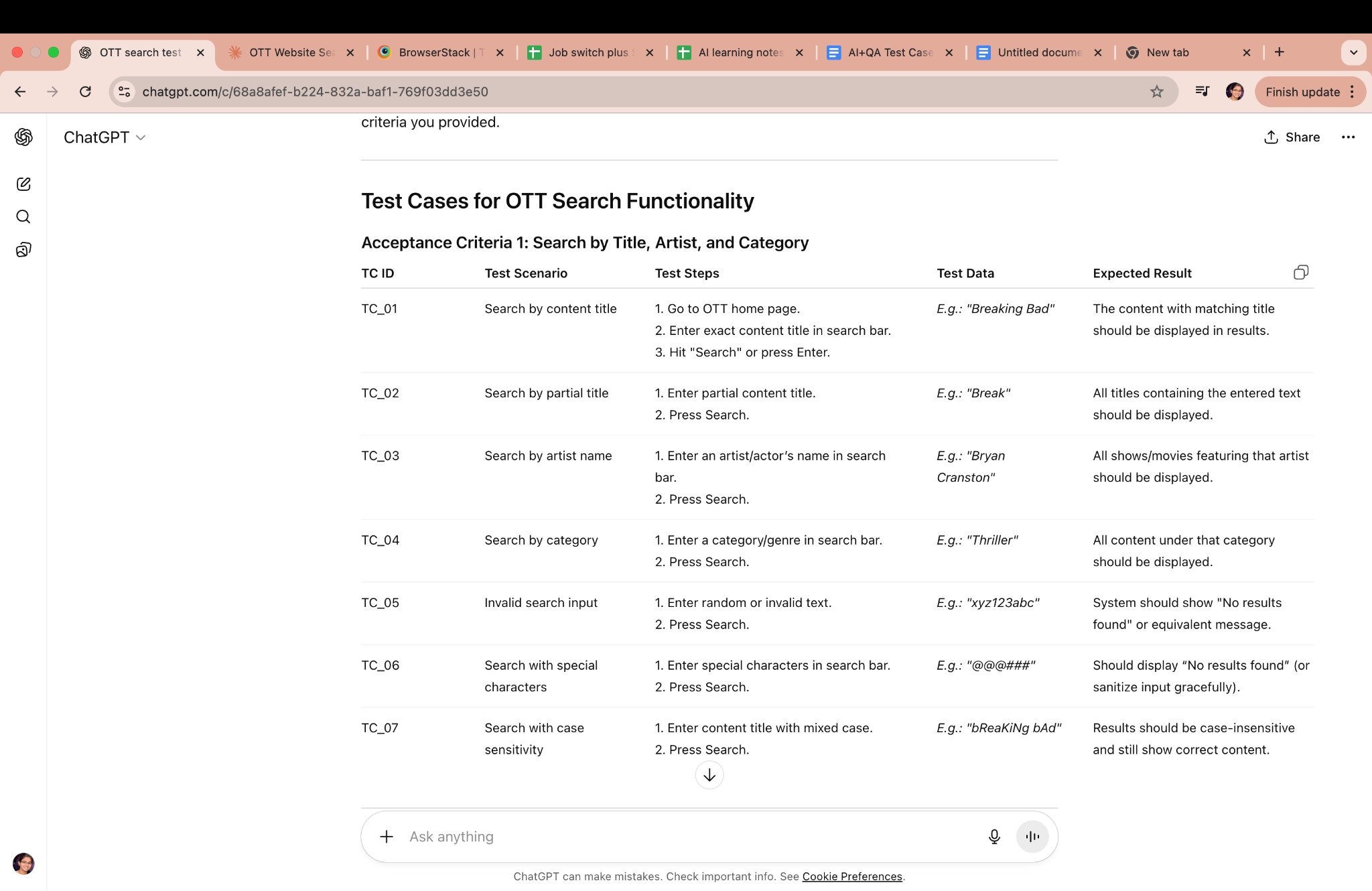
Positive:  
1. So definitely it is faster and covering almost all the scenarios which normal QA can do or may be thinking may deeper than a manual QA

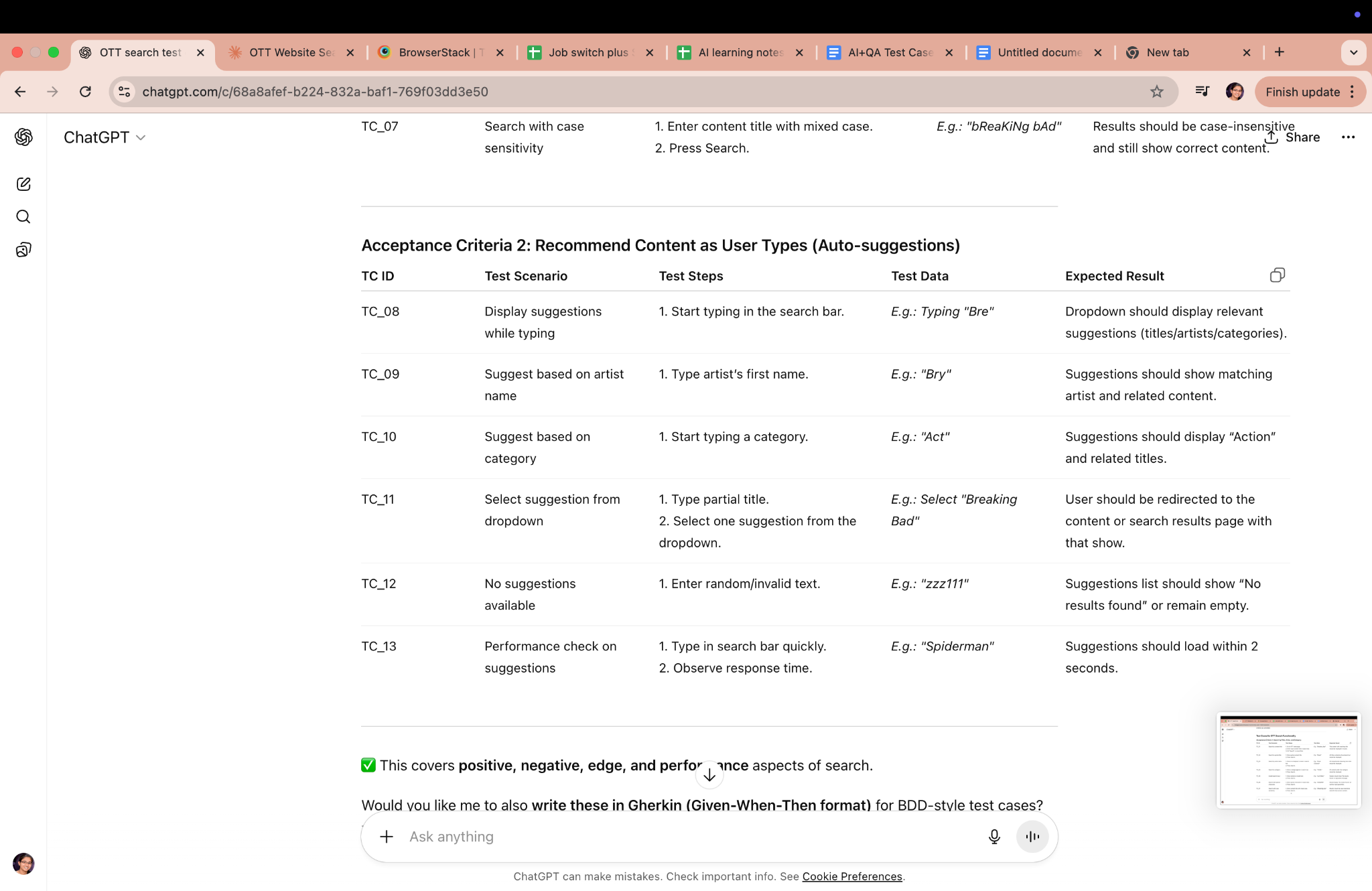
2. The test case preparation approach is the same as QA, it applied bot negative, positive, edge case thinking, it also thinks as a User perspective and covered some performance related scenarios too.

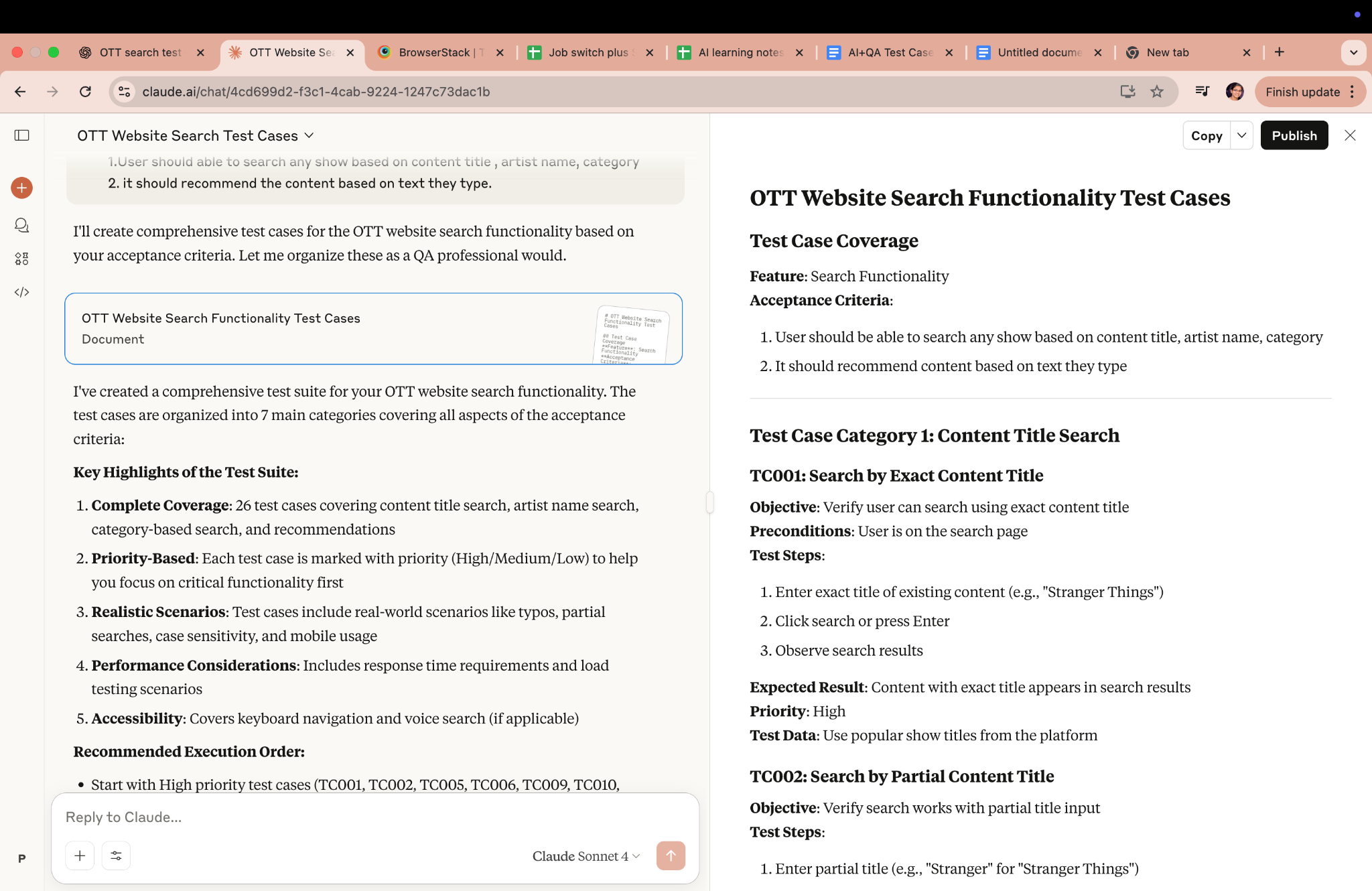
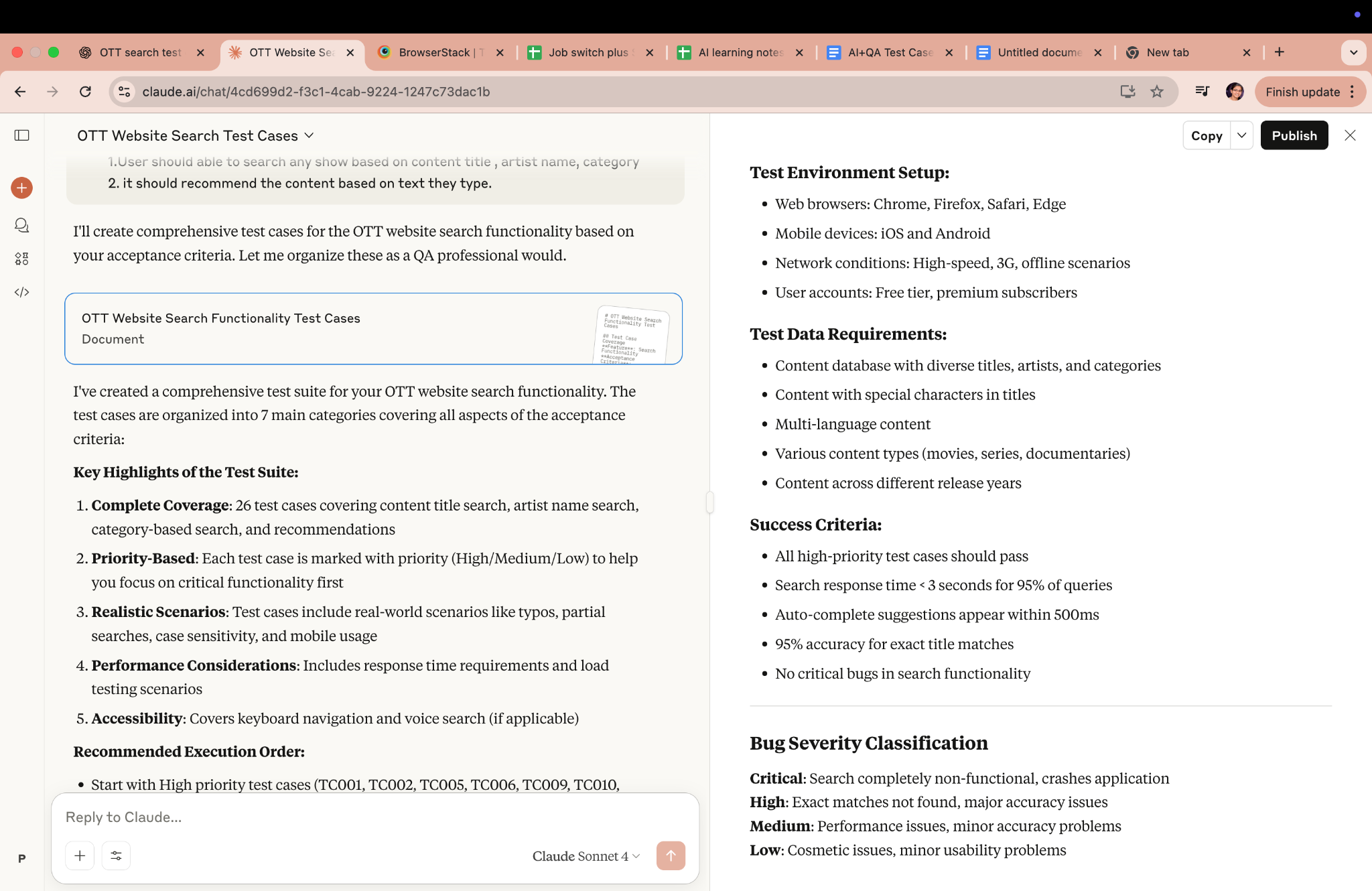
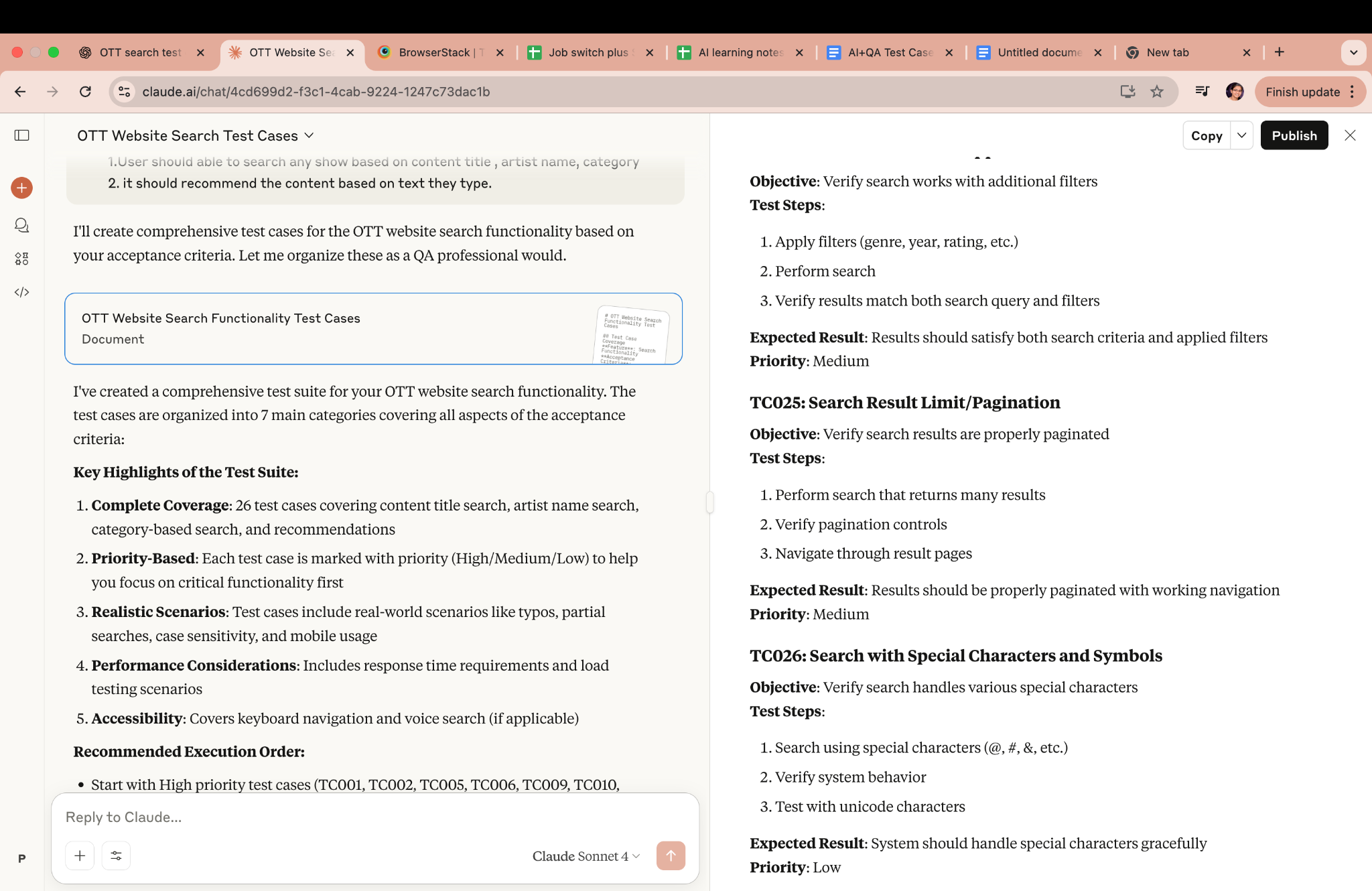
3. We can definitely use the cases by modifying some specific details and it is helpful to speed up our test case generation process.  
  
Negative:  
1. Not a negative point but we can’t rely completely event its faster and generate multiple cases, but we need to review and we can use relevant cases only which are applicable to flow and acceptance criteria

**Screenshot:**

Chatgpt





Claude

Broserstack Agent