

# benchmark\_report.html

Report generated on 20-Feb-2026 at 16:11:44 by [pytest-html](#) v4.2.0

## Environment

Python	3.12.4
Platform	Windows-10-10.0.19045-SP0
Packages	<ul style="list-style-type: none"><li>pytest: 9.0.2</li><li>pluggy: 1.6.0</li></ul>
Plugins	<ul style="list-style-type: none"><li>anyio: 4.12.1</li><li>html: 4.2.0</li><li>metadata: 3.1.1</li></ul>

## Summary

9 tests took 00:01:46.

(Un)check the boxes to filter the results.

1 Failed,  8 Passed,  0 Skipped,  0 Expected failures,  0 Unexpected passes,  0 Errors,  0 Runas,  0 Retried,

Show all details / Hide all details

Result	Test	Duration	Links
Failed	test_local_benchmark.py::test_model_speed_and_accuracy[item2-qwen2.5.0.5b]	819 ms	

```
model_name = 'qwen2.5.0.5b', item = {'category': 'Logic', 'expected': ['hot', 'warm'], 'question': 'If ice is cold, fire is what?'}

@pytest.mark.parametrize("model_name", MODELS_TO_TEST)
@pytest.mark.parametrize("item", load_test_data())
def test_model_speed_and_accuracy(model_name, item):
    category = item['category']
    question = item['question']
    expected_list = item['expected']

    print(f"\n[START] Model: {model_name} | Category: {category}")

    # Start the timer
    start_time = time.time()

    # Call the specific model via Ollama
    response = ollama.chat(model=model_name, messages=[{"role": "user", "content": f"Answer with one word only. {question}"}])

    # Stop the timer
    duration = time.time() - start_time

    # Clean the response
    answer = response['message'][['content']].lower().strip().replace(".", "")

    # Print execution report
    print(f"[RESULT] Time: {(duration:.2f)s} | Answer: '{answer}'")

    # Verify if AT LEAST ONE of the expected words is in the answer
    success = any(expected in answer for expected in expected_list)

>   assert success, f"FAILURE: {model_name} failed on {category}. Got: '{answer}'"
E   AssertionError: FAILURE: qwen2.5.0.5b failed on Logic. Got: 'fire'
E   assert False

test_local_benchmark.py:44: AssertionError
```

Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item0-llama3.2.1b]	00:00:14
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item0-phi3.mini]	00:00:24
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item0-qwen2.5.0.5b]	00:00:04
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item1-llama3.2.1b]	00:00:11
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item1-phi3.mini]	00:00:19
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item1-qwen2.5.0.5b]	00:00:06
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item2-llama3.2.1b]	00:00:09
Passed	test_local_benchmark.py::test_model_speed_and_accuracy[item2-phi3.mini]	00:00:14