## Success: 8 cases

For the Hotels-60 case, XLAM's complete\_score is lower than GPT's because it inefficiently calls functions in series, while GPT correctly groups them into a parallel call.

Specifically, after finding the hotel, GPT makes one parallel call for the description, payment features, and reviews. XLAM makes three separate, sequential calls for the same information, which is less optimal and results in a lower score.

## func\_error, "Do not call the correct function.", 6 cases

## param\_hallucination", "Parameter adults is hallucinated.", 16 cases

[{"error\_type": "param\_missing", "content": "Missing parameter adults in prediction."}]: 1

[{"error\_type": "value\_error", "content": "Parameter (adults) value do not equal to golden."}]: 134

In the case of Attraction-1, XLAM's score is lower because it provided an incorrect id in the Search\_Attractions function call.

While both models correctly first searched for the location "Oxford," GPT used the correct id (eyJ1ZmkiOi0yNjA0OTExfQ==) in the subsequent Search\_Attractions call. In contrast, XLAM provided a different and incorrect id (eyJwaW5uZWRQcm9kdWN0IjoiUFJXYkNGV2V3UERSIiwidWZpIjotMjYwNDkxMX0=), which resulted in a value\_error and a lower score.

For the case Cross-286:

XLAM's score is lower due to a value\_error in its second function call. It attempted to call Google Hotels\_By\_Coordinates with an imprecise longitude value of "4.745925". The correct, more precise longitude required was "4.745924599999999". This led to the error and premature termination of the tool-use sequence.

On the other hand, GPT, despite also making an error by not calling the correct subsequent function (Search\_Attraction\_Location), managed to successfully execute more steps with the correct parameters before failing, resulting in a higher score.

Parameter Errors: In cases Cross-39, Flights-27, Hotels-119, and Car-Rental-74, XLAM failed due to value\_error. This indicates it provided incorrect values for parameters. For instance, in Cross-39, it used a truncated latitude value, and in Flights-27, it provided an incorrect return date.

Inefficient Function Calling: XLAM often makes sequential, individual function calls where parallel calls are more efficient. In Hotels-65 and Cross-94, GPT correctly bundled multiple requests into a single turn, while XLAM separated them into multiple, slower turns, leading to a lower score.

Incomplete or Incorrect Task Execution: In several instances (Cross-312, Cross-235, Attraction-66, Hotels-29, etc.), XLAM stopped early or failed to execute all the required steps. For example, in Cross-312, it didn't proceed to find the nearby shopping mall, and in many Attraction and Hotels domain tasks, it used a generic query (e.g., "Liverpool") instead of the specific one required ("Beatles Liverpool"), leading to incorrect or incomplete results.

{"error\_type": "stop\_early", "content": "Stop early."}: 28  
  
  
  
  
  
  
  
  
CFB-Hard:  
Cross2: round, arg

Car-Rental-100: round, time

Cross-3: round, arg

Car-Rental-31: time

Cross-52: round, truncate

Car-Rental-9: date

Car-Rental-62: time

Cross-9: less call