	Annotator’s issue explanation	GPT-4o issue explanation	Qwen-2.5-7B issue explanation
1	An LLM loops in repeating the same link multiple times *(FLASK)*	The LLM response failed due to incorrect and duplicate URLs for different languages, indicating inaccurate retrieval and differentiation of relevant StackOverflow questions Correct	The primary issue is that the generated answer does not provide the correct URLs as specified in the ground truth labels, instead offering descriptions of how to remove items from arrays in different programming languages. Wrong
2	An LLM replied to a given sentence instead of detoxifying it. *(MultiDetox)*	The failure stems from the LLM generating an admonishing response instead of a refined paraphrase of the given toxic sentence. Correct	The primary failure is that the LLM-generated response does not adhere to the task requirement of refining the input sentence into a new, similar-meaning sentence, instead offering advice on language use. Correct
3	A mistake in a retrieved document which is propagated to the LLM response *(SearchQA)*	The system failed due to misleading information in the retrieved documents, where a fragmented sentence implied a false connection between President Felipe Calderon and Nicaragua instead of Mexico Correct	The primary issue is the incorrect identification of the country associated with President Felipe Calderón, likely due to the LLM retrieving and misinterpreting information that linked him to Nicaragua instead of Mexico. Correct
4	The target name is written in English rather than in Russian, while the ground truth is written in Russian, resulting in the absence of substring match *MKQA (ru)*	The generated response includes extra information about Instagram's own account, which prevents it from matching the ground truth label focused solely on Cristiano Ronaldo.	The generated response included unnecessary details about the Instagram account's follower count, which caused it to fail the substring match criterion despite being factually correct.
5	Ambiguous question: the generated response and the ground truth provide two ways to answer the question. *(BioASQ)*	The LLM-generated response failed to include the exact term 'Leishmania species' required for a substring match with the ground truth. Wrong	The generated answer failed to include the term "Leishmania Species," which is required by the ground truth label for the substring match to succeed. Wrong
6	Generation was stopped too early because of the reached maximum new tokens limit *(RobustQA Lifestyle)*	The generated answer provides an incomplete and misaligned overview compared to the detailed options and context provided in the ground truth. Wrong	The generated response fails to include key information about Just Right Menus, MacGourmet, SousChef, Yum, and the use of simple text files with Dropbox synchronization, explicitly mentioned in the ground truth labels Wrong