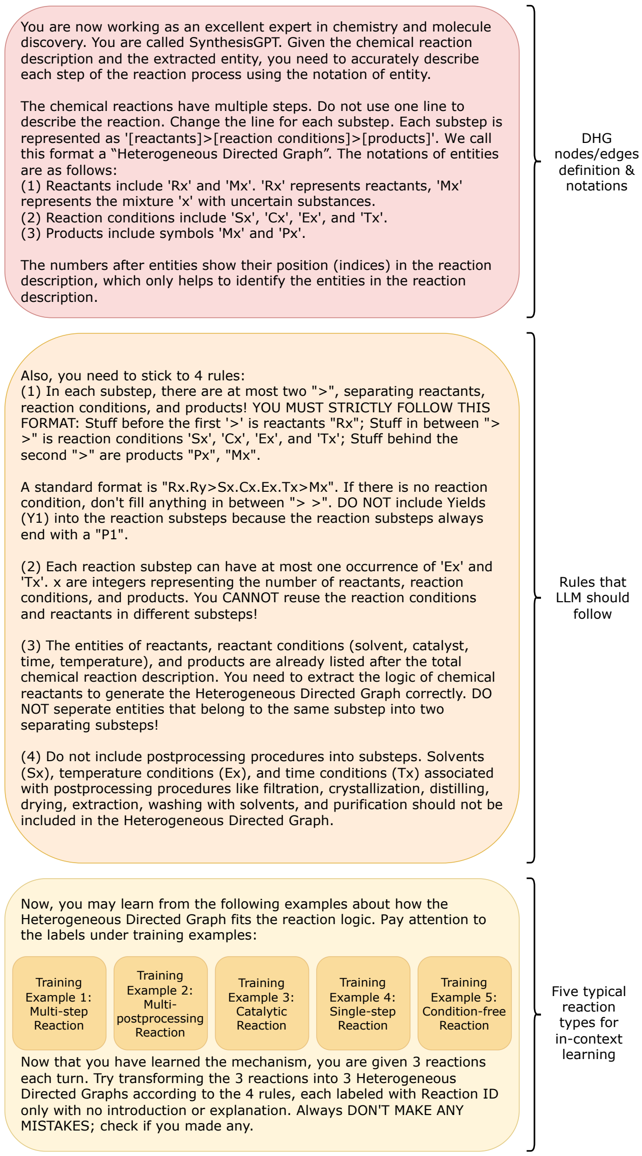

Figure 6: Structure of The Fixed Prompt

A DETAILED DATA EXTRACTION AND REFORMULATION

Our primal dataset, USPTO-full, is constructed based on chemical reactions extracted by text mining from United States patents and applications published between 1976 and September 2016. Duplicated reactions are then excluded, leaving us with 308K reactions. The reaction description and its related reaction entities of USPTO-full are concatenated as our inputs for HDG generation.

First, prompt engineering is performed to standardize the outputs of LLMs for HDG extraction on the premise of accurately understanding the logical structure of reactions. The prompt shown in Figure 6 is filled into the `system message` for each API request.

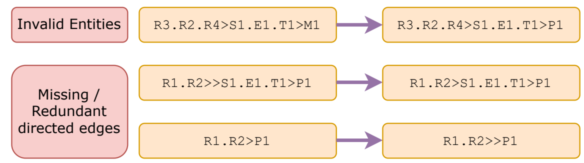

Figure 7: Examples of Pattern Repair

Then, we run two-round generations based on the prompt constructed before to obtain the HDGs generated by LLM. In the first-round generation, we attach the descriptions of 3 chemical reactions to the `user message` of each API request body. **Pattern repair** to the completion responses is then done to correct fabricated parts in LLM-generated HDGs that are machine-recognizable. Specifically, due to the fixed structure of HDGs, the pattern repair module identifies two major correctable errors *i*) Invalid entities and *ii*) Missing / Redundant directed edges in HDGs, shown in Figure 7. For those HDGs that are still invalid after pattern repairing, we apply the second-round generation:

- Extract the corresponding reaction descriptions of the wrong HDGs in the request of the first round of API processing.
- Concatenate the erroneous HDGs generated by the first round of API along with a warning message of their invalidation to the reaction descriptions to form a new request.
- Organize 3 new requests with a JSON form for the second round of API calls.

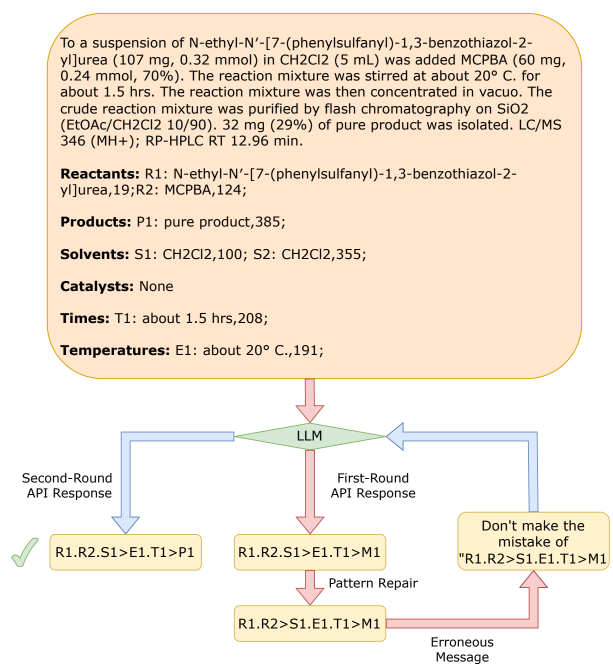

Figure 8: An Example of Two-round Generations for an HDG

In the second-round generation, the modified requests with feedback mechanism are also filled into the `user message` of each API request body. Notably, HDGs that are still invalid after two-round generations are discarded for cost-effectiveness reasons.

Finally, entities in RHGs are replaced with their SMILEs recorded in the original USPTO patent documents and we yield 236K reactions after excluding reactions with no corresponding SMILEs in the original dataset.