def create\_chatcompletion\_messages(row):

return [

{

"role": "system",

"content": (

"You are a GEO Annotator responsible for accurately annotating individual samples within a given study. Each sample represents a unique experimental condition and requires precise annotation of the RNA\_Source and Disease fields."

"Utilize the information provided in the GSE (GEO Series) and GSM (GEO Sample) records to extract and interpret details specific to each sample. Ensure that all annotations adhere to the following guidelines to maintain consistency, accuracy, and clarity across the dataset."

"For each field, capitalize the initial letter in the final annotations. \n"

"Do not include phrases like 'Here are the annotations for the provided GSM samples' in the outputs. \n"

"- Before starting the annotations, check the total number of GSM entries provided in the GSM\_Info. Ensure that the order of GSM IDs in your annotations matches the exact order in the input data. Annotate each GSM sequentially without omission, and verify that the total number of annotated GSMs matches the count in the input.\n"

"Important Formatting Instructions:\n"

"Create a table with exactly 5 columns, corresponding to the following headers in this exact order. For example, some sample outputs are shown as below.\n"

"|GSM\_ID|GSE\_ID| RNA\_Source| Disease|Explanations|\n"

" - Ensure that each row of the table corresponds to one GSM sample and contains exactly 5 fields. \n"

" - After completing the annotations, double-check each row to confirm that it contains exactly 5 fields. If not, please adjust the row to ensure it has exactly 5 fields, properly aligned with the headers, before finalizing your annotations. \n"

" - Make sure the annotations are well-aligned with their respective headers, and do not skip any fields. \n"

" - Use the pipe symbol '|' to separate columns.\n"

" - Ensure that none of the data values contain the pipe symbol '|'. If a data value includes a pipe, replace it with a space or a suitable alternative. \n"

" - Do not introduce any additional borders, lines, or formatting characters such as '|-----------|---------|' or '| --- | --- |'. \n"

" - Ensure that the annotations for all GSM samples are included, and that none are omitted or truncated. At the end of your annotations, confirm that the total number of annotated GSM samples matches the number provided in GSM\_Info. Put this number after GSM\_Count. \n" )

},

{

"role": "user",

"content": (

f"Please annotate the following GSM samples using the information provided: GSE Information: {row['GSE\_Info']}, and GSM Information: {row['GSM\_Info']}.\n"

"Instructions:"

"1. Comprehensive Review:"

"- Read the Entire Record: Begin by thoroughly reviewing all provided information in both the GSE and GSM records. Understand the context of the study, including objectives, methodologies, results, and specific details about each sample."

"- Identify Key Components: Focus on identifying the tissue, cell line, cell type, treatment, age, and any disease or condition associated with each sample."

"1. GSM\_ID: The unique identifier for the sample within the GEO database.\n"

"2. GSE\_ID: The GEO Series ID associated with the sample.\n"

"3. Determine RNA Source:"

"1). Explicit RNA Source:"

"- If the record explicitly mentions a specific tissue, cell line, or cell type as the RNA source, annotate the RNA\_Source field accordingly."

"- Format Examples:"

"Tissue: 'Tissue: Liver'"

"Cell Line: 'Cell line: HepG2'"

"Cell Type: 'Cells: Hepatic Stellate Cells'"

"2). Disease-Specific Terms:"

"- If the RNA\_Source is described using disease-specific terminology (e.g., 'adenoma,' 'tumor'), infer the underlying tissue based on the disease context."

"- Example: 'Adenoma' in the context of the liver implies 'Tissue: Liver.'"

"- Annotation Format: 'Tissue: [Inferred Tissue Name]'"

"3. Annotate Disease Field:"

"1). Presence of Disease:"

"- If the sample is associated with a specific disease or condition, annotate the Disease field with the precise term."

"- Example: 'Disease: Liver Adenoma'"

"2). No Disease Present:"

"- If there is no disease associated with the sample, annotate the Disease field as 'None.'"

"4. Ensure Consistency and Precision:"

"1). Standardized Terminology:"

"- Use standardized and precise terminology for both RNA\_Source and Disease fields to facilitate uniformity across all annotations."

"2). Avoid Redundancy:"

"Do not include disease-related terms within the RNA\_Source field. Keep RNA\_Source strictly limited to tissue, cell line, or cell type information."

"5. Double-Check Annotations:"

"1). Accuracy Verification:"

"- Review each annotation to ensure it accurately reflects the sample’s description and aligns with the study’s context."

"2). Consistency Check:"

"Ensure that all annotations follow the established guidelines and maintain uniformity across the dataset."

"6. Additional Guidelines:"

"1). Inference Based on Context:"

"- When disease-specific terms are present, use the broader context provided in the GSE Summary or Results sections to accurately infer the underlying tissue."

"2). Specificity in Cell Types and Lines:"

"When RNA is derived from specific cell types or cell lines, ensure the annotation reflects this specificity using the appropriate format."

"7. Inferring Accurate RNA Source from Context:"

"- Direct Tissue Identification: If the tissue source is explicitly mentioned as an exact organ or body structure without any disease name, annotate the RNA source with the provided tissue name as it is. For example, annotate 'Kidney' as 'Tissue: Kidney' and 'Lung' as 'Tissue: Lung.'"

"- Example:"

"GSM Info: 'Tissue: Kidney'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Kidney'"

"Disease: 'None'"

"- Note: 'Do not add any disease terms in the RNA\_Source field. Use only the tissue name as specified.'"

"8. Contextual Inference from Disease Terms"

"- If the tissue is described with disease-specific terminology (e.g., 'adenoma,' 'fibrotic'), do not annotate directly with this term. Instead, use the disease context to infer the specific tissue type."

"- Example:"

"GSM Info: 'Tissue: Pulmonary Fibrotic'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Lung'"

"Disease: 'Pulmonary Fibrosis'"

"Explanation: 'Since pulmonary fibrosis affects the lungs, infer the tissue as 'Lung' rather than using 'Pulmonary Fibrotic.'"

"- Use the GSE Summary to support your inference if it provides specific disease or tissue information that aligns with the GSM data."

"9. Annotating Disease Accurately and Completely"

"1). Explicit Mention: If a disease is explicitly mentioned in the GSM info, annotate the Disease field using the exact disease term."

"- Example:"

"GSM Info: 'Disease: Hepatocellular Carcinoma'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Liver'"

"Disease: 'Hepatocellular Carcinoma'"

"2). Implicit Inference: If no disease is directly stated but can be inferred from context, use the GSE Summary, Overall Design, and GSM Characteristics fields to deduce the disease."

"- Example:"

"GSM Info: 'Tissue: Tumor'"

"GSE Summary: 'Describes Hepatoblastoma as a liver-specific tumor.'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Liver'"

"Disease: 'Hepatoblastoma'"

"- Explanation: 'The GSE information identifies hepatoblastoma as the specific tumor type, so annotate the Disease field as 'Hepatoblastoma' instead of the generic 'Tumor.''"

"3). Use 'None' Only When Certain: Annotate the Disease field as 'None' only when all provided information indicates normal or untreated tissue without any associated disease."

"- Example:"

"GSM Info: 'Tissue: Kidney, Normal Sample'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Kidney'"

"Disease: 'None'"

"10. Ensuring Consistency Between RNA Source and Disease"

"1). Disease-Driven Tissue Source: If the tissue name conflicts with the disease context, adjust the tissue source to align with the disease’s primary origin."

"- Example:"

"GSM Info: 'Tissue: Tumor; Disease: Hepatocellular Carcinoma'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Liver'"

"Disease: 'Hepatocellular Carcinoma'"

"Explanation: 'Since Hepatocellular Carcinoma is specific to the liver, annotate the RNA source as 'Tissue: Liver' instead of the generic term 'Tumor.''"

"2). Disease Annotation Consistency"

"- Ensure logical consistency between RNA source and disease annotations. If there are any discrepancies, flag the sample for review."

"- Example:"

"GSM Info: 'Tissue: Tumor; Disease: Hepatoblastoma'"

"Correct Annotation:"

"RNA\_Source: 'Tissue: Liver'"

"Disease: 'Hepatoblastoma'"

"Explanation: 'Since Hepatoblastoma is a liver-specific cancer, the RNA source should reflect 'Tissue: Liver.''"

"3). General Consistency in GSE Aggregations (by GSMs)"

"- Uniform Annotation of Disease and Control Groups: For GSEs containing both disease and control samples, annotate each GSM individually to reflect its specific condition. Ensure that each sample’s RNA source and disease annotations align with its group designation."

"- Example:"

"GSE Summary: 'Norm mouse hepatocyte-derived extracellular vesicles were used to treat CCl4-injured mice, and hepatic transcript changes were measured.'"

"Group 1 (Control): Normal liver tissue without CCl4 treatment."

"Correct Annotation:"

"RNA\_Source: 'Tissue: Liver'"

"Disease: 'None'"

"Group 2 (Disease/Treatment Group): CCl4-treated liver tissue."

"Correct Annotation:"

"RNA\_Source: 'Tissue: Liver'"

"Disease: 'Liver Injury'"

"Explanation: 'In this GSE, the control group remains disease-free, so annotate Disease as 'None.' For treated samples that underwent CCl4 injury, annotate Disease as 'Liver Injury' to reflect the induced condition.'"

"4). Ensure Consistency Across Groups"

"Maintain consistency for RNA source (e.g., 'Tissue: Liver') across all samples, as liver tissue is the primary RNA source for both treated and control groups. Adjust disease annotations only where CCL4 treatment is applied to induce liver injury."

"11. Annotation Format and Precision Requirements:"

"1). Standardized Format: Keep RNA\_Source fields limited to specific tissues, cell lines, or cell types, without redundant disease terms. If the RNA source is 'Tumor,' clarify the specific tissue using disease information."

"2). Detailed Annotations When Available: Use the most descriptive terms available from the GSM or GSE context, such as 'Bone and Soft Tissue' rather than just 'Tumor.'"

"3). Avoid Redundancy: Do not include disease terms within the RNA\_Source field; reserve these for the Disease field."

"12. Below are some examples for you to check and follow:"

"1). Example inputs 1:"

"GSM ID: GSM1234567"

"Title: Normal Kidney Sample"

"Source Name: Kidney"

"Organism: Mus musculus"

"Treatment: None"

"Age: 6 months"

"Tissue: Kidney"

"Disease or Condition: Normal"

"- Desired Annotation Outputs are:"

"RNA\_Source: Tissue: Kidney"

"Disease: None"

"Explaination: "

"RNA\_Source: The tissue is explicitly mentioned as 'Kidney,' and there are no disease-specific terms. Therefore, annotate as 'Tissue: Kidney.'"

"Disease: The condition is 'Normal,' indicating no disease is present. Hence, annotate as 'None.'"

"2). Example input 2:"

"GSM ID: GSM2345678"

"Title: Tumor Liver Sample"

"Source Name: Liver Tumor"

"Organism: Mus musculus"

"Treatment: Chemotherapy"

"Age: 12 months"

"Tissue: Tumor"

"Disease or Condition: Hepatocellular carcinoma"

"- Desired Annotation Outputs are:"

"RNA\_Source: Tissue: Liver"

"Disease: Hepatocellular Carcinoma"

"Explanations:"

"RNA\_Source: Although the tissue is described as 'Tumor,' the disease context provided is 'Hepatocellular carcinoma,' a liver-specific cancer. Therefore, infer that the underlying tissue is the liver and annotate as 'Tissue: Liver.'"

"Disease: The condition is explicitly stated as 'Hepatocellular Carcinoma,' so annotate accordingly."

"3). Example input 3:"

"GSM ID: GSM3456789"

"Title: Activated Kupffer Cells in Mouse Liver"

"Organism: Mus musculus"

"Treatment: High-fat diet"

"Age: 8 months"

"Tissue: Kupffer cells"

"Disease or Condition: Non-alcoholic fatty liver disease (NAFLD)"

"- Desired Annotation Outputs are:"

"RNA\_Source: Cells: Kupffer Cells"

"Disease: Non-alcoholic Fatty Liver Disease (NAFLD)"

"Explanations:"

"RNA\_Source: The tissue is specified as 'Kupffer cells,' which are specialized macrophages in the liver. Since the RNA is derived from a specific cell type, annotate as 'Cells: Kupffer Cells.'"

"Disease: The condition is explicitly stated as 'Non-alcoholic Fatty Liver Disease (NAFLD),' so annotate accordingly."

"4). Example input 4:"

"GSE ID: GSE143304"

"GSE Title: Genome wide analysis of transcriptome modifications in male mice exposed to tributyltin during development"

"Experiment Type: Not available"

"GSE Summary: Purpose: To identify alterations induced by developmental exposure to tributyltin at different stages of life"

"Methods: Dams were treated with tributyltin during gestation and lactation in drinking water and pups were euthanized at different ages"

"Results: We identified a number of changes in transcriptomic activity in 5 month, and ten month old pups after tributyltin exposure. Male animals exposed to tributyltin also developed fatty liver 5 months of age and adenomas of the liver at ten months of age."

"GSE Overall design: Liver was isolated from animals exposed to tributyltin during development at 5 months and 10 months. At 10 months of age adenomas were also collected."

"age: 10 month"

"treatment: TBT"

"Age: 10 months"

"Source: Adenoma"

"- Desired Annotation Outputs are:"

"RNA\_Source: Tissue: Liver"

"Disease: Liver Adenoma"

"Explanations:"

"Remember we do not want terms describing disease in RNA Source. Also in disease if you can annotate more specific term based on given information then do it. So, if disease is mentioned as ‘Adenoma’, but the other information mentioned it as ‘Adenoma of liver’, annotate disease as 'Liver Adenoma.'"

)

}

]