def create\_chatcompletion\_messages(row):

return [

{

"role": "system",

"content": (

"An Expert Biological Data Annotator with Domain Knowledge in Transcriptomics and Oncology, capable of interpreting biological and medical contexts, inferring missing metadata (e.g., organ regions or cell types), and providing consistent, concise annotations for transcriptomic datasets, tailored to experimental and biological origins."

"Annotate the Revised RNA Source for transcriptomic data by providing clear, concise, and consistent annotations based on the metadata given by ROC field (the combined info on'Organ\_Region,' 'Disease,' and 'RNA\_source' fields)."

"Your annotations should help distinguish between cell types, tissues, tumor sources, and metastatic cases."

)

},

{

"role": "user",

"content": f"Carefully review the info below: {row['RNA\_Source']}\n"

"Given Input:"

"The input metadata in 'ROD'(the combined Organ\_Region info, Disease info and the inital RNA\_source info in this format: Organ\_Region: xx and Disease: xx and RNA\_source: xx: xx) includes the following fields info:"

"Organ\_Region: The organ or tissue where the sample originates. If 'NA' or generic terms like 'Tumor' or 'Subcutaneous Tumor' are provided, infer from the Disease field."

"Disease: The associated disease or condition of the sample. Use it to guide inferences but do not repeat it in the RNA source."

"RNA\_source: The type of RNA source, which could be 'Tissue,' 'Cells,' or a 'Cell Line,' along with any specific cell type or cell line name."

"Desired Output:"

"A single Revised RNA Source field that:"

"Focuses on the sample type and specific details (e.g., organ, cell type, or cell line)."

"Avoids repeating information already in Disease or Organ\_Region."

"Infers organ and cell types when necessary to provide clear, concise annotations."

"Before starting the standardization, check the total number of Original Terms provided. Ensure that the order of terms in your Standardized terms matches the exact order in the input data. Standardize each Original Term sequentially without omission, and verify that the total number of Standardized Terms matches the count in the input.\n"

"Logic and Rules:"

"1. Tissue Samples:"

"1) Tumor Tissue with Specific Organ\_Region:"

"- If Disease indicates a tumor and RNA\_source is 'Tissue,' annotate as 'Tumor Tissue: [Organ\_Region]' or 'Tumor Tissue: [Inferred Organ]'."

"- If the sample is experimental (e.g., graft, implant), include the context in parentheses."

"Examples:"

"- Input: Organ\_Region: Tumor, Disease: Hepatoblastoma, RNA\_source: Tissue: Tumor"

"Revised RNA Source: 'Tumor Tissue: Liver'"

"- Input: Organ\_Region: Subcutaneous Tumor, Disease: Colorectal Cancer, RNA\_source: Tissue: MC38 Tumor"

"Revised RNA Source: 'Tumor Tissue: MC38 (Subcutaneous Implant)'"

"- Input: Organ\_Region: Brainstem, Disease: Diffuse Intrinsic Pontine Glioma, RNA\_source: Tissue: Tumor"

"Revised RNA Source: 'Tumor Tissue: Brainstem'"

"- Input: Organ\_Region: Tumor, Disease: Ewing Sarcoma, RNA\_source: Tissue: Tumor"

"Explanation: Since Ewing Sarcoma typically originates in bone or soft tissue, and no specific organ information is provided, generalize it to 'Tumor Tissue: Bone or Soft Tissue'."

"Revised RNA Source: 'Tumor Tissue: Bone or Soft Tissue'"

"- Input: Organ\_Region: Tumor, Disease:Hepatoblastoma , RNA\_source: Tissue: Tumor"

"Explanation: Since Hepatoblastoma is a liver-origin tumor, even though Organ\_Region is generically labeled as 'Tumor,' specify the primary tissue based on the disease."

"Revised RNA Source: 'Tumor Tissue: Liver'"

"- Input: Organ\_Region: Pancreas and Disease: Pancreatic Cancer and RNA\_source: Tissue: Tumor"

"Revised RNA Source: 'Tumor Tissue: Pancreas' "

"2). Normal or Non-Cancerous Tissue:"

"Annotate simply as 'Tissue: [Organ\_Region]' or 'Tissue: [Inferred Organ]'."

"Example:"

"- Input: Organ\_Region: Liver, Disease: None, RNA\_source: Tissue: Liver"

"Revised RNA Source: 'Tissue: Liver'"

"- Input: Organ\_Region: Rectum, Disease: Congenital Diarrhoea, RNA\_source: Tissue: Rectum"

"Revised RNA Source: 'Tissue: Rectum'"

"3). Tumor Tissue with Generic or Non-specific Organ\_Region:"

"When Organ\_Region is generic (e.g., ‘Tumor’, ‘Subcutaneous Tumor’, ‘Melanoma’): "

"- If Disease specifies a cancer type with a known primary organ, infer the organ from the Disease. "

"Annotate as: "

"'Tumor Tissue: [Inferred Organ]' if Organ\_Region is ‘Tumor’ or a cancer type. "

"'Tumor Tissue: [Inferred Organ] (Subcutaneous Implant)' if Organ\_Region is ‘Subcutaneous Tumor’ or indicates a subcutaneous location. "

"Examples: "

"- Input: Organ\_Region: Tumor and Disease: Melanoma and RNA\_source: Tissue: Tumor"

"Explanation: Melanoma primarily originates in the skin. "

"Revised RNA Source: 'Tumor Tissue: Skin'"

"- Input: Organ\_Region: Melanoma and Disease: Melanoma and RNA\_source: Tissue: Melanoma "

"Explanation: Since ‘Melanoma’ is both the disease and the Organ\_Region, and it originates in the skin. "

"Revised RNA Source: 'Tumor Tissue: Skin' "

"- Input: Organ\_Region: Subcutaneous Tumor and Disease: Triple-Negative Breast Cancer and RNA\_source: Tissue: Tumor "

"Explanation: Triple-Negative Breast Cancer originates in the breast; the subcutaneous location indicates an implant. "

"Revised RNA Source: 'Tumor Tissue: Breast (Subcutaneous Implant)' "

"- Input: Organ\_Region: Subcutaneous Tumor and Disease: Subependymal Giant Cell Astrocytoma and RNA\_source: Tissue: Subcutaneous Tumor "

"Explanation: Subependymal Giant Cell Astrocytoma is a brain tumor. "

"Revised RNA Source: 'Tumor Tissue: Brain (Subcutaneous Implant)' "

"- If Disease is also generic or does not specify a primary organ, annotate as: "

"'Tumor Tissue: Unknown'"

"2. Cell Line Samples:"

"1) Cell Lines:"

"- When RNA\_source is a cell line, annotate as 'Cell Line: [Cell Line Name]'. Remove any redundant disease terms if they repeat information in the Disease field."

"Example:"

"- Input: Organ\_Region: Skin, Disease: Melanoma, RNA\_source: Cell Line: B16 Melanoma"

"Revised RNA Source: 'Cell Line: B16'"

"- Input: Organ\_Region: NA, Disease: Myeloid Leukemia, RNA\_source: Cell Line: MLL-AF9 murine leukemia cell line"

"Revised RNA Source: 'Cell Line: MLL-AF9'"

"- Only include disease when Cellline name is not available."

"- Input: Organ\_Region: Lung, Disease: Small Cell Lung Cancer, RNA\_source: Cell Line: SCLC derived cell line"

"Revised RNA Source: 'Cell Line: SCLC'"

"3. Cellular Samples:"

"1). Tumor Cells:"

"If Disease indicates a cancer or tumor, annotate as 'Tumor Cells: [Specific Cell Type]' if a specific cell type is given, otherwise just 'Tumor Cells: [Organ\_Region]'."

"Example:"

"- Input: Organ\_Region: Lung, Disease: Lung Cancer, RNA\_source: Cells: Tumor Cells (Alveolar Cells)"

" Revised RNA Source: 'Tumor Cells: Alveolar Cells'"

"- Input: Organ\_Region: Bone Marrow, Disease: Acute Myeloid Leukemia, RNA\_source: Cells: Leukemic Myeloid Cells"

"Revised RNA Source: 'Tumor Cells: Myeloid Cells'"

"- Input: Organ\_Region: Spleen, Disease: Acute Myeloid Leukemia, RNA\_source: Cells: Leukemic Splenocytes"

"Revised RNA Source: 'Tumor Cells: Splenocytes'"

"Leukemias and Lymphomas: "

"- When the Disease indicates a leukemia or lymphoma, and the RNA\_source includes the disease name, replace it with the specific abnormal cell type in the annotation. "

"Leukemia Samples: "

"Use 'Tumor Cells: [Specific Abnormal Cell Type]'. "

"Examples: B-ALL (B-cell Acute Lymphoblastic Leukemia): Use 'B-Lymphoblasts'. "

" Acute Myeloid Leukemia: Use 'Myeloid Cells'. "

"Lymphoma Samples: "

"Annotate as 'Tumor Tissue: Lymphoid Tissue' or 'Tumor Cells: Lymphoid Cells'. "

"If the RNA\_source includes the disease name, replace it with the specific cell type associated with that disease. "

"Hepatoblastoma: Use 'Hepatoblasts' instead of 'Hepatoblastoma'. "

"Examples: "

"Input: Organ\_Region: NA, Disease: B-ALL, RNA\_source: Cells: B-ALL"

"Revised RNA Source: 'Tumor Cells: B-Lymphoblasts'"

"Input: Organ\_Region: Liver, Disease: Hepatoblastoma, RNA\_source: Tumor Cells: Hepatoblastoma"

"Revised RNA Source: 'Tumor Cells: Hepatoblasts'"

"Input: Organ\_Region: Subcutaneous Tumor, Disease: Lymphoma, RNA\_source: Tissue: Subcutaneous Tumor"

"Revised RNA Source: 'Tumor Tissue: Lymphoid Tissue (Subcutaneous Implant)'"

"Include experimental context if applicable (e.g., '(Subcutaneous Implant)')"

"2). Metastatic Cancer Cells:"

"If the sample is metastatic (Organ\_Region differs from Disease's primary organ), annotate as 'Metastatic Cells: [Specific Cell Type] (Origin: [Disease primary organ])'."

"Example:"

"- Input: Organ\_Region: Bone, Disease: Prostate Cancer, RNA\_source: Cells: Tumor Cells (Osteoclasts)"

"Revised RNA Source: 'Metastatic Cells: Osteoclasts (Origin: Prostate)'"

"3) Normal or Non-Cancerous Cells:"

"For normal (non-cancerous) cells, include the specific cell type and annotate as 'Cells: [Specific Cell Type]'."

"Example:"

"- Input: Organ\_Region: Blood, Disease: None, RNA\_source: Cells: Peripheral Blood Mononuclear Cells"

"Revised RNA Source: 'Cells: Peripheral Blood Mononuclear Cells'"

"4). Disease-Specific Non-Cancerous Cells:"

"For cells associated with non-cancerous diseases, include the specific cell type, without 'Tumor' or 'Metastatic' labels."

"Example:"

"- Input: Organ\_Region: Lung, Disease: Asthma, RNA\_source: Cells: Alveolar Macrophages"

"Revised RNA Source: 'Cells: Alveolar Macrophages'"

"4. Additional Notes:"

"- If Organ\_Region or Disease are unclear, make logical inferences based on commonly known origins of the disease."

"- For experimental contexts (e.g., subcutaneous grafts or implants), include descriptors like '(Subcutaneous Implant)' or '(Graft)' in parentheses."

"- Use medical knowledge to infer the primary organ associated with the Disease when necessary. "

"- If the sample involves a subcutaneous implant or graft, include '(Subcutaneous Implant)' in the annotation. "

"4. Tissue Samples:\n"

"- Remove unnecessary words such as 'Whole' (e.g., 'Whole Blood' -> 'Blood') or 'Adult' (e.g., 'Adult Skin Fibroblast' -> 'Skin Fibroblast').\n"

"- Format example: 'Organ\_Region: Blood and Disease: NA and RNA\_source: Tissue: Whole Blood -> Revised RNA Source: Tissue: Blood'\n\n"

"5. Cell Lines:\n"

"- Standardize cell line names to consistent formats (e.g., Kasumi-1, MyLa, OCI-Ly7) by normalizing capitalization, hyphenation, and suffixes.\n"

"- Example: 'Cell Line: Kasumi-1 AML Cell Line' -> 'Cell Line: Kasumi-1'\n"

"- Remove redundant prefixes and disease names from cell line entries.\n\n"

"6. Cells:\n"

"- Standardize singular/plural forms of cellular descriptions (e.g., 'Cells: Bronchial Epithelium' and 'Cells: Bronchial Epithelial Cells').\n"

"- Use singular for consistency where appropriate (e.g., 'Bronchial Epithelium').\n"

"- Replace terms like 'Cells: Glioblastoma' with concise descriptions, e.g., 'Glioblastoma Cells'.\n"

"- Maintain specific, clear cell type descriptors (e.g., 'Cells: Endothelial Cells' remains unchanged).\n\n"

"7. Summary of Format Rules:"

"- Tumor Tissue with Specific Organ\_Regionferred Organ):"

"1) If Organ\_Region is 'Tumor': 'Tumor Tissue: [Inferred Organ]'"

"2) If Organ\_Region is 'Subcutaneous Tumor':'Tumor Tissue: [Inferred Organ] (Subcutaneous Implant)'"

"- Tumor Tissue with Unknown Organ:'Tumor Tissue: Unknown'"

"- Normal or Non-Cancerous Tissue: 'Tissue: [Organ\_Region]'"

"- Cell Lines: 'Cell Line: [Cell Line Name]'"

"- Tumor Cells: 'Tumor Cells: [Specific Cell Type]'"

"Leukemia Samples:'Tumor Cells: [Specific Abnormal Cell Type]'"

"Lymphoma Samples: 'Tumor Tissue: Lymphoid Tissue'"

"- Metastatic Cells: 'Metastatic Cells: [Specific Cell Type] (Origin: [Disease primary organ])'"

"- Normal or Non-Cancerous Cells: 'Cells: [Specific Cell Type]'"

"8. Create the final outputs as a table with exactly 2 columns, corresponding to the following headers in this exact order. For example, some sample outputs are shown as below.\n"

"|ROD| Revised\_RNA\_Source|\n"

"| Organ\_Region: Liver and Disease: NA and RNA\_source: Tissue: Liver| Tissue: Liver|\n"

"| Organ\_Region: Blood and Disease: NA and RNA\_source: Tissue: Whole Blood | Tissue: Blood |\n\n"

"ROD: List each original term in ORD column separately.\n"

"Revised\_RNA\_Source: Provide the standardized term corresponding to each original term.\n"

"Ensure accuracy and consistency based on biological standards and domain expertise.\n"

"Do not include any explanations in any format."

}

]