def create\_chatcompletion\_messages(row):

return [

{

"role": "system",

"content": (

"As a Biological Data Analyst, your primary task is to standardize the 'Perturbation\_Type' field. "

"Focus on uniformity, clarity, and scientific accuracy in presenting each perturbation type. "

"This includes standardizing compound names, biological agents, and treatment methods, ensuring "

"consistency across entries."

)

},

{

"role": "user",

"content": (

f"Carefully review the original genotype terms below: {row['Perturbation\_Name']}\n"

"Standardization Guidelines:\n"

"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\n"

"1. Inferring Perturbation\_Type from Perturbation\_Name:\n"

"Use the detailed 'Perturbation\_Name' to logically infer the 'Perturbation\_Type'. This involves understanding the nature "

"of the agents and their classifications based on their usage and effects in biological contexts. Choose the appropriate "

"'Perturbation\_Type' from the following categories:\n\n"

"- Genetic: Includes gene-level modifications, interventions like siRNAs, and viral vectors used for gene delivery "

"(but not the viral infections themselves).\n"

"- Biological: Includes proteins, antibodies, and other biological agents.\n"

"- Chemical: Includes drugs, small molecules, and pharmacological interventions.\n"

"- CTL (Control/Untreated): Includes untreated samples, vehicle controls, and common solvents like DMSO, saline, and PBS.\n"

"- CP (Chemical Perturbation): Encompasses drugs, inhibitors, toxins, and any synthetic chemical compounds.\n"

"- BIO (Biological Perturbation): Covers biological agents like cytokines, growth factors, antibodies, peptides, live cells, "

"and biological extracts.\n"

"- KO (Knockout): Genetic perturbations where a gene is completely inactivated or deleted.\n"

"- KD (Knockdown): Reduction of gene expression using methods like siRNA, shRNA, or antisense oligonucleotides.\n"

"- OE (Overexpression): Perturbations involving increased expression of a gene or protein.\n"

"- ES (Environmental Perturbation): Changes in environmental conditions like diet, hypoxia, or physical injury.\n"

"- VIR (Viral Infection): Perturbations involving infection with viruses or viral particles.\n"

"- OTHER: Perturbations that do not fit into the above categories or lack sufficient information.\n\n"

"1.1. Direct Association: Directly associate common agents with their respective types based on their established usage and effects "

"in biological studies. For example, classify agents like 'doxycycline' or 'Tamoxifen' as 'CP', and specific viruses (e.g., 'Influenza Virus') as 'VIR'.\n"

"1.2. Handling Complex Combinations: For names indicating combinations of different agents (e.g., 'doxycycline + 4' OH-tamoxifen + mepazine + cycloheximide'), "

"break down each component and classify accordingly. Ensure each agent is correctly categorized using the updated types, such as 'CP + CP + CP + CP'.\n"

"1.3. Clarifying the 'OTHER' Category Handling:\n"

"General Handling: For perturbations that do not fit into the predefined categories, classify these perturbations simply as 'OTHER' without specifying additional details. "

"This ensures uniformity and avoids variability from non-standard specifics.\n"

"Handling Combined Perturbations: For combined perturbations where one part falls under 'OTHER', such as '4-Thiouridine + UV Irradiation', format the output to reflect "

"the presence of a standard category plus 'OTHER', without additional detail on the non-standard part. For example:\n"

"'24-Hour Food Deprivation' should be standardized to 'EP'.\n"

"'4-Thiouridine + UV Irradiation' should be standardized to 'CP + OTHER'.\n"

"Consistency: Always ensure that the standardized term for 'OTHER' does not carry additional descriptors or specifics, maintaining broad categorization to enhance consistency across data entries.\n\n"

"1.4. Chemical Inducers and Modifiers: When the perturbation agent is a known chemical inducer or modifier, categorize it under 'Chemical'. "

"If the chemical specifically induces a genetic modification or a disease model, consider specifying this further if applicable. For example:\n"

"'Chemical Induced Disease Model: DSS-Induced Colitis' should be categorized as 'Chemical' with additional context if needed.\n\n"

"1.5. Generalize Perturbation Types: When annotating the perturbation type, avoid including specific details about the perturbation agent or its concentration. "

"The output should reflect only the general category of perturbation, such as 'OE' or 'CP', without additional descriptors or details about the agent used. For instance:\n"

"Convert specific entries like 'Genetic: Lentiviral MyoD Overexpression' to 'OE'.\n"

"Convert 'Small Molecule: Retinoic Acid' to 'CP'.\n"

"This approach ensures a high-level, consistent categorization suitable for broader analyses and comparisons across studies.\n\n"

"1.6. Treat Complex Multi-Component Terms as Single Perturbation Types:\n"

"Guideline: When the term involves multiple components connected by either a '+' or ';', treat the entire set of terms as one single entry during standardization for the 'Pert\_Type' field.\n"

"\*\*Usage of Symbols:\n"

"Use the '+' Symbol: Represents multiple agents or substances combined in a single perturbation. Each agent connected by '+' is to be individually categorized but kept within the same entry.\n"

"Use the ';' Symbol: Separates distinct genetic modifications or complex components within a single entry. Components connected by ';' are to be grouped as one complete perturbation unit.\n"

"\*\*Implementation Instructions:\n"

"For each '+' in the original term, ensure each connected agent is categorized separately, while still presenting them as part of a unified perturbation.\n"

"Do not split components separated by either '+' or ';'. Instead, keep the entire set as a single term when standardizing but apply individual types as needed.\n"

"Output Format for Combined Perturbations: Provide a standardized output that respects the '+' and ';' separations, using distinct but combined types where relevant (e.g., 'CP + CP' or 'KO; KO; BIO').\n"

"\*\*Examples:\n"

"Combined Agents with '+':\n"

"Doxycycline + Tamoxifen should be categorized as 'CP + CP'.\n"

"Bicuculline Methiodide + DL-Norepinephrin Hydrochloride + Carbamolylcholine Chloride becomes 'CP + CP + CP'.\n"

"Aflibercept + AMG386 + Anti-PD1 remains as 'BIO + BIO + BIO'.\n"

"\*\*Multiple Genetic Modifications with ';':\n"

"Conditional Knockout: Yap; Conditional Knockout: Taz; Cre: VE-cadherin-CreERT2 is treated as 'KO; KO; BIO'.\n"

"Conditional Knockout: Mst1; Conditional Knockout: Mst2; Cre: alb-CreF becomes 'KO; KO; BIO'.\n"

"Conditional Knockout: NIKdeltaT3flSTOP; Conditional Knockout: Notch2ICN; Cre: CD19Cre is presented as 'KO; KO; BIO'.\n\n"

"1.7. Formatting Guidelines:\n"

"- Use only short forms (e.g.,Genetic, Biological, Chemical,CTL, CP, BIO, KO, KD, OE, ES, VIR, OTHER) for standardization. Do not use full names like 'Chemical Perturbation' or 'Biological Perturbation' etc.\n"

"- Maintain consistent spacing and capitalization.\n"

"- Ensure the output reflects the exact order and sequence of perturbation components in the input.\n\n"

"1.8. Format the Output as a Table:\n"

"Create a table with two columns:\n"

"Column 1: List each Original Term exactly as it appears in the input. Preserve Original Term and do not apply any formatting or transformations to the original input term. "

"Simply copy it directly into the first column of the output.\n"

"Column 2: Provide the standardized term for each corresponding original term.\n\n"

"- Apply pharmacological and biological standards to ensure accuracy in the standardization process.\n\n"

)

}

]