def create\_chatcompletion\_messages(row):

return [

{

"role": "system",

"content": "As a Researcher in Mouse Genomics, your task is to annotate mouse model terms for strain and genotype. The primary objective is to ensure accurate identification and standardization of mouse strains and genotypes based on the provided terms. The output should be in a table format with three columns: Original Term, Strain, and Genotype. Use 'NA' if the strain or genotype cannot be determined from the term."

},

{

"role": "user",

"content": f"Carefully review the original strain names in the strain column below: {row['Strain']}\n\n"

"You have been provided with a dataset containing mouse genetic strains utilized in various experiments. However, the strain names within the dataset lack uniformity, necessitating standardization for clarity and consistency.\n"

"Guidelines for Standardization:\n"

"1. Check the Term:\n"

" - Determine if the term contains both strain and genotype information.\n"

"2. Strain Annotation:\n"

" - Use a clear, standardized form for all strains, including substrains or sublines, ignoring the maintaining institution.\n"

" - Recognize and standardize variations in strain names, including different capitalizations, case sensitivity, and placements of slashes.\n"

" - use the following Common Mouse Strains for their corresponding variations:\n"

" - Use C57BL/6: Includes all variations like C57BL6x, C57Black6, CD57BL/6, Bl/6, BL6, BL6/C57, etc.\n"

" - BALB/c: Includes all variations like BALB/cJ, BALB/cByJ, etc.\n"

" - 129: Includes all variations like 129S1/SvImJ, 129P3/J, 129S2/SvPas, etc.\n"

" - 129/Sv: Includes all variations like 129SV, 129/SV, 129/sV, 129sV, 129Sv, 129/sv, 129sv, etc.\n"

" - FVB: Includes all variations like FVB/NJ, FVB/NCrl, etc.\n"

" - DBA: Includes all variations like DBA/1J, DBA/2J, etc.\n"

" - CD-1: includes all variations like CD-1 (ICR) or any variations like cd1, CD1, ICR.\n"

" - PWK: Refers to the PWK strain.\n"

" - 129/Ola: includes all variations like 129ola, 129 ola, 129 Ola.\n"

" - 129/SvEv: Includes all variations like 129 SvEv, 129SvEv ect.\n"

" - For recombinant inbred strains, use the format 'Mixed: Strain1 x Strain2', with the lowercase ‘x’.\n"

" - Mixed: For terms 'mixed', 'mix', 'Mix', 'Mixed, 'mixed background' etc without specifics for the parent strains.\n"

" - For hybrid strains, the maternal strain is listed first, followed by the paternal strain.\n"

" - If the strain cannot be determined, annotate as 'NA'.\n"

" - Remove specific institution identifiers such as 'J' from strain names to maintain standardization. Ensure all instances of institution-specific identifiers, such as 'J', are removed from strain names to prevent inconsistencies.\n"

" - Double-check all strain names to confirm that all lab or company identifiers have been removed, ensuring the standardized output is consistent.\n"

"3. Genotype Annotation:\n"

" - Categorize the genotype as one of the following: Overexpression, Transgenic, Knockout, Knockdown, Conditional knockout, Reporter, Heterozygous.\n"

" - Follow the category with the specific gene involved, separated by a colon (e.g., Knockout: Myc).\n"

" - If no specific gene information is provided, annotate as 'NA'.\n"

"4. Identify Specific Markers:\n"

" - Include specific genetic markers or modifications in the genotype annotation. For example, 'GFP' or 'mCherry' should be annotated as 'Reporter: GFP' or 'Reporter: mCherry'.\n"

"5. Compound Terms:\n"

" - Separate compound terms into their respective strain and genotype components.\n"

"6. Use of Symbols:\n"

" - Use '+' for heterozygous (e.g., 'Heterozygous: Gene') and '-' for knockout (e.g., 'Knockout: Gene').\n"

"7. Unknown Information:\n"

" - Use 'NA' for parts of the term that are unclear or cannot be determined.\n"

"8. Case Sensitivity:\n"

" - Normalize case sensitivity across all annotations.\n"

"9. Abbreviation Expansion:\n"

" - Expand common abbreviations to their full form for clarity (e.g., 'KO' to 'Knockout').\n"

"10. Hyphen and Slash Normalization:\n"

" - Normalize the use of hyphens and slashes in strain names.\n"

"11. Standardize Typographical Errors:"

"If you encounter common typographical errors or variations in strain names (e.g., 'CH3' instead of 'C3H'), correct them to the standard strain name."

"Example: 'CH3' should be corrected to 'C3H'."

"12. Cell Lines:"

"For terms that are known cell lines (e.g., 'CGR8', 'CJ7'), annotate the strain as 'NA' since they are not considered strains but rather specific cell lines."

"Example: 'CGR8' and 'CJ7' should be annotated as 'NA'."

"13. Mixed Genetic Populations:"

"For strains that are mixed genetic populations or labeled as outbred (e.g., 'Diversity Outbred', 'Outbred'), annotate them as 'Mixed'."

"Example: 'Diversity Outbred' or 'Outbred' should be annotated as 'Mixed'."

"14. Recombinant Inbred Strains:"

"For recombinant inbred strains, use the format 'Mixed: Strain1 X Strain2', ensuring that the 'X' is capitalized."

"Example: For a cross between C57BL/6 and DBA/2, annotate as 'Mixed: C57BL/6 X DBA/2'.\n"

"15. Double Check Results:\n"

" - Double check your results before providing the final revised terms to ensure all the rules are applied correctly, and the final results are accurate.\n"

" - Double check to make sure the 'X' is capitalized for all the crossed strains in the Mixed: Strain 1 X Strain 2 format.\n"

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

"Create a table with three columns:\n"

"Column 1: List each original term exactly as it appears in the input.\n"

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

"Column 3: Provide the Genotype for each corresponding original term.\n"

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

"For each annotation, keep a concise annotation only.\n"

"Don't provide any explanations in any forms.\n"

"Example Entries:\n"

"Original Term Strain Genotype\n"

"---------------------------------- ----------------------- ----------------------\n"

"BDF1(mother)XPWK(father) |Mixed: C57BL/6 X DBA/2 X PWK |NA\n"

"LFng-GFP BAC Transgenic Mouse (B6;FVB-Tg(Lfng-EGFP)HM340Gsat/Mmucd) |Mixed: C57BL/6 X FVB |Overexpression: Lfng-EGFP\n"

"Foxp3-GFP-DTR\_CD4CreER\_R26tdTomato |NA |Overexpression: Foxp3-GFP, Conditional Knockout: CD4, Reporter: R26-tdTomato\n"

"Cx3cr1CreER +/- |NA |Conditional Knockout: Cx3cr1, Heterozygous: Dicer\n"

"FVB-Tg(Yac128)53Hay/J |FVB |Overexpression: Yac128\n"

"F1 |NA |NA\n"

"BXD-11/TyJ |Mixed: C57BL/6 X DBA/2 |NA\n"

"Bxa11 |Mixed: C57BL/6 X A/J |NA\n"

"Amigo2-GFP C57BL/6 |C57BL/6 |Overexpression: Amigo2-GFP\n"

"C57BL6x |C57BL/6 |NA\n"

"C57Black6 |C57BL/6 |NA\n"

"CD57BL/6 |C57BL/6 |NA\n"

"Bl/6 |C57BL/6 |NA\n"

"BL6 |C57BL/6 |NA\n"

"B6NCrl |C57BL/6 |NA\n"

"129S2/SvPas |129S2/Sv |NA\n"

"Bmal -/- |NA |Knockout: Bmal\n"

"AP4-mCherry c-MYC-GFP knockin |NA |Knockin: c-MYC-GFP, Reporter: AP4-mCherry\n"

"BKS.Cg-Dock7m+/+Leprdb/J |BKS |Knockout: Leprdb, Dock7m+/+\n"

"AXB-2/PgnJ |Mixed: A/J X C57BL/6 |NA\n"

}

]