def create\_chatcompletion\_messages(row):

return [

{

"role": "system",

"content": "As a Biological Data Analyst with expertise in demographic classification and social identities, your primary task is to standardize the 'Ethnicity' field. You have a thorough understanding of human population groups and how ethnic and cultural identities are represented across various datasets."

},

{

"role": "user",

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

"Detailed Instructions for Improving Revised Terms:\n"

"1. \*\*Avoid Ambiguity\*\*: Use terms that are clear and recognized by major institutions or governments. For instance, terms like 'Caucasian' should be standardized to 'White,' as this is the more widely accepted term in scientific and demographic contexts.\n"

"2. \*\*Avoid Multiple Descriptions\*\*: When multiple ethnic groups are described in a single term (e.g., 'Hispanic, Latino or Spanish Origin'), standardize the term to the most commonly used version, such as 'Hispanic or Latino.'\n"

"3. \*\*Ethnic Subgroup Clarity\*\*: Specify ethnic subgroups when needed but keep them standardized. For example, 'Chinese Han' should be standardized to 'Han' to maintain clarity and avoid overcomplication.\n"

"4. \*\*Use 'NA' for Missing Data\*\*: Retain 'NA' for missing or unspecified ethnicities. Do not replace it with terms like 'Not Specified' or 'Not Available.'\n"

"5. \*\*Groupings Based on Region\*\*: For ethnicities defined by geographic regions, standardize the names of the regions and ethnic groups, e.g., 'East African (Kenya)' should be standardized as 'East African'. The specific country names should be omitted unless they are necessary for clarity.\n"

"6. \*\*Consistency in 'Hispanic' Terms\*\*: Ensure that 'Hispanic,' 'Latino,' or 'Spanish Origin' are standardized consistently as 'Hispanic or Latino.' Do not use multiple variants in the same dataset.\n"

"7. \*\*Neutral and Inclusive Terms\*\*: Terms like 'Coloured' should be avoided and replaced with more neutral and widely accepted terms, such as 'Black' or 'African American' where appropriate.\n"

"8. \*\*Standardize National Terms\*\*: Where ethnic groups are identified by nationality, use the common standardized names. For example, 'British Indian' should become 'Indian,' and 'South Asian (Pakistan)' should be standardized as 'South Asian.'\n"

"9. \*\*Avoid Unnecessary 'Non' Prefixes\*\*: Terms such as 'Non-Hispanic' or 'Non-Hispanic, Latino or Spanish Origin' should be replaced with 'Not Hispanic or Latino.'\n"

"10. \*\*Retain Common Groupings\*\*: Terms like 'African American,' 'European American,' and 'White British' should be standardized according to common terminology. For example, 'African American' remains 'African American,' 'White' can be used for 'Caucasian' or 'White British.'\n"

"11. \*\*Ensure Consistency in Mixed Group Terms\*\*: For terms like 'Mixed' or 'Multi-racial,' standardize them to 'Mixed.' Do not use overly broad terms like 'Any other ethnic group.'\n"

"12. \*\*Retain General Terms for Unknown or Unspecified Data\*\*: Use 'Unknown' or 'NA' when data is unspecified or missing. Do not use terms like 'Not Specified' unless absolutely necessary.\n"

"Format the Output as a Table:\n"

" - Create a table with two columns:\n"

" - List each original term separately in the first column.\n"

" - Provide the standardized term in the second column, ensuring clarity and consistency.\n"

" - Apply domain expertise to ensure accuracy in the standardization process.\n"

"Examples:\n"

"| Original Term | Standardized Term |\n"

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

"| African American | African American |\n"

"| Any other Asian background | Asian |\n"

"| Asian | Asian |\n"

"| Black | Black |\n"

"| British | British |\n"

"| Caucasian | White |\n"

"| Chinese | Chinese |\n"

"| East African (Kenya) | East African |\n"

"| Hispanic | Hispanic or Latino |\n"

"| Hispanic or Latino | Hispanic or Latino |\n"

"| Indian | Indian |\n"

"| Japanese | Japanese |\n"

"| Mixed | Mixed |\n"

"| NA | NA |\n"

"| Not Specified | Unknown |\n"

"| South Asia | South Asian |\n"

"| White | White |\n"

"| YRI | YRI (Yoruba in Ibadan) |"

}

]