def create\_chatcompletion\_messages(row):

return [

{

"role": "system",

"content": "As a Biological Data Analyst, your task is to standardize and categorize the 'Genotype' entries under appropriate genetic modification categories. Each genotype should follow a consistent format with the correct categorization, gene naming conventions, and uniform use of symbols. Ensure accuracy, consistency, and clarity in the representation of the genetic data."

},

{

"role": "user",

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

"Standardization Rules:\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"

"1. Major Categories: Categorize each genotype into a 'major type' such as 'Overexpression, Knock-In, Knockdown, Conditional Knockout, Knockout, Heterozygous, Homozygous, Mutant, Control, Transgenic', etc. Each entry should follow the format: 'Type: Gene'.\n"

"2. Case Sensitivity: Ensure uniform capitalization. The type (e.g., Knockout, Overexpression) should be capitalized, and gene symbols should follow proper nomenclature conventions (capitalize human genes, lowercase for mouse genes). Consistent capitalization should be applied across similar entries (e.g., `tdTomato+` and `TdTomato+` should be standardized to `TdTomato+`).\n"

"3. Symbol Standardization: Use consistent symbols across all entries. Ensure the correct placement of symbols like '+' or '-/-', with spaces where appropriate (e.g., `GFP+` → `GFP +`). Symbols must follow proper spacing, and redundant symbols (e.g., `-/-`) should be removed unless critical for conveying genetic status.\n"

"4. Formatting: Ensure that the output format is consistent: 'Type: Gene'. Use semicolons to separate multiple genes or modifications within the same entry (e.g., `Knockout: p53; Ink4a/Arf`).\n"

"5. Clarification of Complex Genotypes: For complex genotypes (e.g., those involving multiple genes with different modifications), each component should be categorized separately within the same entry. Example: '\*\*Conditional Knockout: Apc min/+; Kras LSL-G12D/+; Villin-Cre; p53 KO\*\*'.\n"

"6. Parenthetical Information: Remove unnecessary parenthetical abbreviations, but retain important contextual information such as mutations or reporter genes (e.g., `GFP+`).\n"

"7. Redundant Symbol Removal: Remove redundant symbols unless they are needed to indicate a specific genetic modification (e.g., keep `-/-` for knockouts). Example: \*\*TdTomato+\*\* and \*\*tdTomato+\*\* should be standardized as `TdTomato +`.\n"

"8. Gene Nomenclature: Follow proper gene naming conventions. Human genes should be written in uppercase (e.g., \*\*MYC\*\*), while mouse genes should be written in title case or lowercase as appropriate (e.g., \*\*Myc\*\*).\n"

"Examples:\n"

"- Original Term: Knockout: linc-Enc1 -/-\n"

"- Standardized Output: Knockout: Linc-Enc1 -/-\n\n"

"- Original Term: Overexpression: EGFR+\n"

"- Standardized Output: Overexpression: EGFR +\n\n"

"- Original Term: TdTomato+, tdTomato+\n"

"- Standardized Output: TdTomato + \n"

"9. Consistent Terminology for Conditional Knockouts: Ensure the full term 'Conditional Knockout' is included in the standardized term. Example: '\*\*Conditional Knockout: c-Maf flox/flox CD4Cre\*\*' should be standardized as '\*\*Conditional Knockout: c-Maf flox/flox; CD4-Cre\*\*'.\n\n"

"10. Heterozygous and Homozygous Mutations: Ensure that heterozygous mutations are standardized with the correct '+' notation (e.g., `Heterozygous: Ctnnb1 +/fl`) and homozygous mutations with the '\*\*-/-\*\*' notation (e.g., `Knockout: p53 -/-`).\n\n"

"11. Handling of Complex Entries: For entries with multiple genes or modifications, use semicolons to separate components clearly. Example: '\*\*Conditional Knockout: Apc min/+; Kras LSL-G12D/+; Villin-Cre; p53 KO\*\*'.\n\n"

"12. Uniform Use of Floxed Genes: Convert all instances of '\*\*fl/fl\*\*' to '\*\*flox/flox\*\*' for clarity and consistency. Ensure correct spacing between components.\n\n"

"- Original Term: Conditional Knockout: c-Maf fl/fl\n"

"- Standardized Output: Conditional Knockout: c-Maf flox/flox\n\n"

"13. Keeping Complex Genetic Modifications Unified: When a genotype involves multiple genetic modifications (e.g., knockouts, overexpression, knockdowns) within the same entity and is connected by a semicolon (;), the entire complex modification should be treated as a single term. This includes combinations of knockouts, overexpression, knockdowns, and other genetic alterations."

"Do not split terms connected by a semicolon. If a semicolon is present, it indicates that the term is complex and should remain unified as one single term in subsequent steps for clarity and consistency."

"Examples:"

"'Knockout: MyD88 -/-; Knockout: TRIF -/-' should be treated as one single term."

"'Knockout: Oct1; Mutant: BRCA1-I26A' should be treated as one single term."

"'Knockout: SAMHD1; Knockout: IFNAR' should be treated as one single term."

"'Mutant: BRCA1-I26A; Knockout: Oct1' should be treated as one single term."

"'Mutant: NRAS(V12); Mutant: Mll-AF9' should be treated as one single term."

"'Overexpression: K19-Ptgs2; Overexpression: K19-Ptges' should be treated as one single term."

"'Overexpression: NRAS(V12); Overexpression: Mll-AF9' should be treated as one single term."

"'Reporter: Btg2GFP; Reporter: Tubb3GFP' should be treated as one single term."

"'Overexpression: AcKRS-TAG-Dendra2; Overexpression: H3.3 K56TAG' should be treated as one single term."

"'Knockdown: Nkx2-1 +/-; Knockdown: Foxa2 +/-; Knockdown: Cdx2 +/-' should be treated as one single term."

"'Knockdown: Nkx2-1; Mutant: KrasG12/+; Knockout: Trp53-/-' should be treated as one single term."

"'Double Enhancer Knockout: E1 -/- E2 -/-; Knockdown: SET1A shRNA' should be treated as one single term, not split into separate entries for each modification."

"'Knockdown: Nkx2-1; Knockdown: Foxa2; Knockdown: Cdx2; Mutant: KrasG12/+; Knockout: Trp53 -/-' should be treated as one single term."

"'Knockout: MBNL1 -/-; Knockout: MBNL2 -/-; Overexpression: GFP+' should remain a single term despite involving both knockouts and overexpression."

"'Knockout: TTP -/-; Overexpression: GFP-TTP+' represents a combined modification of knocking out one gene and overexpressing another, and should be kept as one single term."

"'Overexpression: HrasG12V; Knockdown: p53' involves both gene overexpression and gene knockdown, and should be standardized as a single term."

"'Knockout: TCRαβ-/-; Knockout: RAG1-/-; Reporter: Foxp3GFP' should be treated as one single term."

"'Knockout: Tet2-/-; Mutant: Flt3ITD'should be treated as one single term."

"14. Use 'db/db' to refer to 'Knockout: Lepr' for homozygous leptin receptor deficiency."

"Use 'ob/ob' to refer to 'Knockout: Lep' for homozygous leptin deficiency."

"Use 'Foxn1nu/nu' to refer to 'Knockout: Foxn1' for homozygous nude phenotype."

"Add the strain background before the genotype if applicable (e.g., 'C57BL/6J-db/db')."

"For transgenic or knockout models, explicitly state the functional implication, e.g., 'Apoe-/-' for 'Knockout: Apoe'."

"15. Gene Naming Conventions: Mouse genes must be written in title case (e.g., 'Eed') according to mouse gene nomenclature conventions."

"If a gene is capitalized (e.g., 'EED'), it should be converted to title case (e.g., 'Eed')."

"Examples: 'Conditional Knockout: EED' → 'Conditional Knockout: Eed'; 'Conditional Knockout: EZH2' → 'Conditional Knockout: Ezh2'."

"16. Cre Systems: Cre driver lines must remain intact and properly formatted, with hyphens separating 'Cre' from the prefix (e.g., 'Zp3-Cre')."

"Separate multiple Cre lines with semicolons, ensuring consistent formatting."

"Examples: 'Cre: AlbCre' → 'Cre: Alb-Cre'; 'Cre: Pax8rtTA; Cre: TetO-cre' → 'Cre: Pax8-rtTA; Cre: TetO-Cre'."

"17. Beta-catenin: The term 'Beta-catenin' should be standardized to its mouse gene name 'Ctnnb1'."

"Examples: 'Conditional Knockout: Beta-catenin' → 'Conditional Knockout: Ctnnb1'."

"18. Conditional Knockout Formatting: The term 'Conditional Knockout' should remain unchanged."

"Separate multiple genetic modifications using semicolons."

"Examples: 'Conditional Knockout: Ctnnb1' → 'Conditional Knockout: Ctnnb1'; 'Conditional Knockout: Apc; Mutant: Kras G12D' → 'Conditional Knockout: Apc; Mutant: Kras G12D'."

"19. Heterozygosity and Homozygosity: Use '+/-' to represent heterozygous knockouts and '-/-' to represent homozygous knockouts."

"Examples: 'Conditional Knockout: Ctnnb1 +/-' → 'Conditional Knockout: Ctnnb1 +/-'."

"20. Reporter Genes: Standardize reporter genes with proper formatting and capitalization."

"Convert common terms like 'Reporter' to their expanded forms where appropriate."

"Examples: 'Reporter: RLP22HA +' → 'Reporter: Rosa26-Rlp22ha +'; 'Reporter: Ai9 fl(Stop)TdTomato' → 'Reporter: Ai9-Fl(Stop)-TdTomato'."

"21. Floxed Genes: Convert 'fl/fl' to 'flox/flox' for clarity."

"Examples: 'Conditional Knockout: Apc fl/fl' → 'Conditional Knockout: Apc flox/flox'."

"22. Mutant Allele Notation: Retain specific mutation designations such as '<sup>' or numeric/letter substitutions for alleles."

"Examples: 'Conditional Knockout: Kras<sup>G12D</sup>' → 'Conditional Knockout: Kras G12D'."

"23. Transgenic Terms: Ensure proper capitalization and retain location descriptions where applicable."

"Examples: 'Transgenic: Flag-CLC-mRuby 2 at Tigre' → 'Transgenic: Flag-Clc-mRuby 2 at Tigre'."

"24. Knock-In Genes: Use 'Knock-In' to describe genes with targeted insertions."

"Examples: 'Knock-In: ROSA26-StopFL-Ikk2ca +/L' → 'Knock-In: Rosa26-StopFL-Ikk2ca +/L'."

"25. Multiple Genetic Modifications: Separate multiple genetic modifications using semicolons for clarity and consistency."

"Examples: 'Conditional Knockout: Ctnnb1; Knockout: Pten -/-' → 'Conditional Knockout: Ctnnb1; Knockout: Pten -/-'."

"26. Transgenic and Conditional Knockouts in Same Term: Separate transgenic and knockout components with semicolons, retaining proper capitalization."

"Examples: 'Transgenic: Flag-LCL-mRuby 2 at Tigre; Conditional Knockout: Ctcfl' → 'Transgenic: Flag-Lcl-Mruby 2 at Tigre; Conditional Knockout: Ctcfl'."

"27. General Formatting: Use semicolons to separate multiple modifications or gene components, and ensure consistent capitalization."

"Examples: 'Conditional Knockout: Apc; Mutant: Kras G12D' → 'Conditional Knockout: Apc; Mutant: Kras G12D';"

"'Conditional Knockout: Ctnnb1; Knockout: Pten -/-' → 'Conditional Knockout: Ctnnb1; Knockout: Pten -/-'."

"28. Capitalization for Mouse Genes: Ensure all mouse gene names are in title case, even if they appear capitalized in the original input."

"Examples: 'Conditional Knockout: ABCA1' → 'Conditional Knockout: Abca1'; 'Conditional Knockout: ACSL4' → 'Conditional Knockout: Acsl4'."

"29. Final Formatting: Present each standardized term in the format \*\*Type: Gene\*\* or \*\*Type: Gene; Gene2\*\* if multiple genes are involved. Ensure that each entry follows the appropriate capitalization, symbol usage, and spacing conventions.\n\n"

"Final Format:\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 splitting, 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/revised term for each corresponding original term.\n\n"

"Ensure accuracy and consistency based on genetic nomenclature standards and domain expertise. Do not add explanations in any format."

}

]