Table S1. List of data-to-paper research steps with their Provided prior products, Performer system prompt, Performer mission prompt, created products, Reviewer system prompt and Reviewer mission prompt.

List of research steps

Prompts color scheme

Prompts are color-coded according to the following scheme:

General mission description
Detailed mission instructions
Product formatting instructions
One shot examples
Code output formatting guidelines
Reference to prior products

Data exploration

LLM	gpt-4
Provided prior products	Data (only provided as data file for code), General description of dataset, Data file description
Performer system prompt	You are a brilliant data scientist. You are writing a Python code to analyze data.
Performer mission prompt	As part of a data-exploration phase, please write a complete short Python code for getting a first sense of the data. Your code should create an output text file named "data_exploration.txt", which should contain a summary of the data. The output file should be self-contained; any results you choose to save to this file should be accompanied with a short header. The output file should be formatted as follows: ```output # Data Size <Measure of the scale of our data (e.g., number of rows, number of columns)> # Summary Statistics <Summary statistics of all or key variables> # Categorical Variables <As applicable, list here categorical values and their most common values> # Missing Values <Counts of missing, unknown, or undefined values> <As applicable, counts of special numeric values that stand for unknown/undefined if any (check in the "Description of the Dataset" above for any)> # <other summary you deem relevant, if any> <summary> ``` If needed, you can use the following packages which are already installed: ('pandas', 'numpy', 'scipy') Do not provide a sketch or pseudocode; write a complete runnable code. Do not create any graphics, figures or any plots. Do not send any presumed output examples.
Product	Data exploration – code: Python code Data exploration – output: Numerical data
Reviewer system prompt	You are a brilliant data scientist. You are writing a Python code to analyze data.
Reviewer mission prompt	I ran your code. Here is the content of the output file(s) that the code created: ``` <Data exploration code - output> ```

	Please follow these two steps: (1) Check the code and the output for any issues, and return a bullet-point response addressing these points: * Are there any unexpected NaN values in the output. * Can results be understood from the output file? In particular, do we have a short label for each result? * Are there any results that are missing. Check that under each header in the output file there is a corresponding meaningful result. * Any other issues you find. (2) Based on your assessment above, return a Python Dict[str, str] mapping the issues you have noted above (dict keys) to specific suggested corrections/improvements in the code (dict values). For example: { "The result of the average of variable ... is missing": "Add the missing calculation of ... to the code.", "The average of the variable <xxx> is `Nan`": "Remove missing values in the calculation." } Try to be as specific as possible when describing the issues and proposed fixes. Include in the dict as many issues as you find. If there are no issues, and the code and tables are just perfect and need no corrections or enhancements, then return an empty dict: {} Important: * Do not return the revised code, only the issues and suggested fixes. * If there are no critical issues, then return an empty dict: `{}`. * Do not create positive issues that require no change in the code. In particular, do not write {"No issues found": "No corrections or improvements are needed."}, return an empty dict instead.

Data exploration explanation

LLM	gpt-4
Provided prior products	General description of dataset, Data file description, Data exploration - code
Performer system prompt	You are a scientist who needs to write explanation of the Data Exploration code.
Performer mission prompt	Please return a triple-backtick Latex Block explaining what the code above does. Do not provide a line-by-line explanation, rather provide a high-level explanation of the code in a language suitable for a Methods section of a research paper. Focus on analysis steps. There is no need to explain trivial parts, like reading/writing a file, etc. Also explain what does the code write into the "data_exploration.txt" file. Your explanation should be written in LaTeX, and should be enclosed within a LaTeX Code Block, like this: ```latex \section{Code Explanation} <your code explanation here> ``` Remember to enclose your explanation within a LaTeX Code Block, so that I can easily copy-paste it!
Product	Data exploration – code explanation: LaTex text

Research goal

LLM	gpt-3.5-turbo-0613, with temperature=1
Provided prior products	General description of dataset, Data file description, Data exploration - code output, Data exploration - code explanation
Performer system prompt	You are a helpful scientist.
Performer mission prompt	Please suggest a research goal and an hypothesis that can be studied using only the provided dataset. The goal and hypothesis should be interesting and novel. Guidelines: * Try to avoid trivial hypotheses (like just testing for simple linear associations). Instead, you could perhaps explore more complex associations and relationships, like testing for moderation effects or interactions between variables. * Make sure that your suggested hypothesis can be studied using only the provided dataset, without requiring any additional data. In particular, pay attention to using only data available based on the provided headers of our data files (see "Description of the Original Dataset", above). * Avoid goals and hypotheses that involve ethic issues like sociodemographic (Income, Education, etc.) and psychological (Mental Health) variables. Note though that you can, and should, still use these as confounding variables if needed. * Do not suggest methodology. Just the goal and an hypothesis. INSTRUCTIONS FOR FORMATTING YOUR RESPONSE: Please return the goal and hypothesis enclosed within triple-backticks, like this: ``` Research Goal: <your research goal here> Hypothesis: <your hypothesis here> ```
Product	Research goal: Free text
Reviewer system prompt	You are a scientific reviewer for a scientist who needs to suggest research goal and hypothesis.
Reviewer mission prompt	Here is the research goal and hypothesis: *<Research goal product>* Please provide constructive bullet-point feedback on the above research goal and hypothesis. Specifically:

	* If the hypothesis cannot be tested using only the provided dataset (without requiring additional data), suggest how to modify the hypothesis to better fit the dataset. * If the hypothesis is not interesting and novel, suggest how to modify it to make it more interesting. * If the hypothesis is broad or convoluted, suggest how best to focus it on a single well defined question. Do not provide positive feedback; if these conditions are all satisfied, just respond with: "The research goal does not require any changes". If you feel that the initial goal and hypothesis satisfy the above conditions, respond solely with "The research goal does not require any changes".

Literature search I

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data file description, Research goal
Performer system prompt	You are a scientist who needs to write literature search queries.
Performer mission prompt	Please write literature-search queries that we can use to search for papers related to our study. You would need to compose search queries to identify prior papers covering these 2 areas: "dataset": papers that use the same or similar datasets as in our study "questions": papers that ask questions similar to our study Return your answer as a `Dict[str, List[str]]`, where the keys are the 2 areas noted above, and the values are lists of query string. Each individual query should be a string with up to 5-10 words. For example, for a study reporting waning of the efficacy of the covid-19 BNT162b2 vaccine based on analysis of the "United Kingdom National Core Data (UK-NCD)", the queries could be: { "dataset": ['The UK-NCD dataset', 'covid-19 vaccine efficacy dataset'] "questions": ['covid-19 vaccine efficacy over time', 'covid-19 vaccine waning'] }
Product	Literature search I – queries: Structured text Literature search I – citations: Citations

Similar citation search

LLM	gpt-4
Provided prior products	General description of dataset, Data file description, Research goal, Literature search I - citations (Dataset and Question scopes)
Performer system prompt	You are a scientist who needs to find most similar papers.
Performer mission prompt	From the literature search above, list up to 5 key papers whose results are most similar/overlapping with our research goal and hypothesis. Return your response as a Python Dict[str, str], where the keys are bibtex ids of the papers, and the values are the titles of the papers. For example: { "Smith2020TheAB": "A title of a paper most overlapping with our goal and hypothesis", "Jones2021AssortedCD": "Another title of a paper that is similar to our goal and hypothesis", }
Product	Similar papers: Citations

Goal validation

LLM	gpt-4
Provided prior products	General description of dataset, Data file description, Research goal, Similar papers
Performer system prompt	You are a scientist who needs to check research goal and hypothesis.
Performer mission prompt	Given the related papers listed above, please follow these 3 steps: (1) Provide a bullet-point list of potential similarities between our goal and hypothesis, and the related papers listed above. (2) Determine in what ways, if any, our stated goal and hypothesis are distinct from the related papers listed above. (3) Given your assessment above, choose one of the following two options: a. Our goal and hypothesis offer a significant novelty compared to existing literature, and will likely lead to interesting and novel findings {'choice': 'OK'}. b. Our goal and hypothesis have overlap with existing literature, and I can suggest ways to revise them to make them more novel {'choice': 'REVISE'}. Your response for this part should be formatted as a Python dictionary mapping 'choice' to either 'OK' or 'REVISE'. Namely, return either: {'choice': 'OK'} or {'choice': 'REVISE'}
Product	Goal validation: Binary decision

Hypothesis testing plan

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data file description, Data exploration - code, Data exploration - code output, Research goal
Performer system prompt	You are a scientist who needs to write hypothesis testing plan.
Performer mission prompt	We would like to test the specified hypotheses using the provided dataset. Please follow these two steps: (1) Return a bullet-point review of relevant statistical issues. Read the "Description of the Original Dataset" and the "Data Exploration Code and Output" provided above, and then for each of the following generic statistical issues determine if they are relevant for our case and whether they should be accounted for: * multiple comparisons. * confounding variables (see available variables in the dataset that we can adjust for). * dependencies between data points. * missing data points. * any other relevant statistical issues. (2) Create a Python Dict[str, str], mapping each hypothesis (dict key) to the statistical test that would be most adequate for testing it (dict value). The keys of this dictionary should briefly describe each of our hypotheses. The values of this dictionary should specify the most adequate statistical test for each hypothesis, and describe how it should be performed while accounting for any issues you have outlined above as relevant. For each of our hypotheses, suggest a *single* statistical test. If there are several possible ways to test a given hypothesis, specify only *one* statistical test (the simplest one). Your response for this part should be formatted as a Python dictionary, like this: { "xxx is associated with yyy and zzz": "linear regression with xxx as the independent variable and yyy and zzz as the dependent variables while adjusting for aaa, bbb, ccc", "the association between xxx and yyy is moderated by zzz": "repeat the above linear regression, while adding the interaction term between yyy and zzz", } These of course are just examples. Your actual response should be based on the goal and hypotheses that we have specified above (see the "Research Goal" above). Note how in the example shown the different hypotheses are connected to each other, building towards a single study goal. Remember to return a valid Python dictionary Dict[str, str].
Product	Hypothesis testing plan: Structured text

Data analysis code

LLM	gpt-4
Provided prior products	Data (only provided as data file for code), General description of dataset, Data file description, Data exploration - code output, Research goal, Hypothesis testing plan
Performer system prompt	You are a brilliant data scientist. You are writing a Python code to analyze data.
Performer mission prompt	Write a complete Python code to analyze the data and create dataframes as basis for scientific Tables for our paper. The code must have the following sections (with these exact capitalized headers): `# IMPORT` `import pickle` You can also import here any other packages you need from the following list: ('pandas', 'numpy', 'scipy', 'statsmodels', 'sklearn', 'pickle') `# LOAD DATA` Load the data from the original data files described above (see "Description of the Original Dataset"). `# DATASET PREPARATIONS` * Join dataframes as needed. * Dealing with missing, unknown, or undefined values, or with special numeric values that stand for unknown/undefined (check in the "Description of the Original Dataset" for any such values, and consider also the "Output of the Data Exploration Code"). * Create new columns as needed. * Remove records based on exclusion/inclusion criteria (to match study goal, if applicable). * Standardization of numeric values with different units into same-unit values. If no dataset preparations are needed, write below this header: `# No dataset preparations are needed.` `# DESCRIPTIVE STATISTICS` * In light of our study goals and the hypothesis testing plan (see above "Research Goal" and "Hypothesis Testing Plan"), decide whether and which descriptive statistics are needed to be included in the paper and create a relevant table. For example: `## Table 0: "Descriptive statistics of height and age stratified by sex"` Write here the code to create a descriptive statistics dataframe `df0` and save it using: `df0.to_pickle('table_0.pkl')` If no descriptive statistics are needed, write: `# No descriptive statistics table is needed.` # PREPROCESSING Perform any preprocessing steps needed to further prepare the data for the analysis. For example, as applicable:

	* Standardization and normalization of numeric values (as needed). * Creating dummy variables for categorical variables (as needed). * Any other data preprocessing you deem relevant. If no preprocessing is needed, write: `# No preprocessing is needed, because <your reasons here>.` # ANALYSIS Considering our "Research Goal" and "Hypothesis Testing Plan", decide on 1-3 tables (in addition to the above descriptive statistics, if any) we should create for our scientific paper. Typically, we should have at least one table for each hypothesis test. For each such scientific table: [a] Write a comment with a suggested table's caption. Choose a caption that clearly describes the table's content and its purpose. For example: `## Table 1: "Test of association between age and risk of death, accounting for sex and race"` Avoid generic captions such as `## Table 1: "Results of analysis"`. [b] Perform analysis - Perform appropriate analysis and/or statistical tests (see above our "Hypothesis Testing Plan"). - The statistical analysis should account for any relevant confounding variables, as applicable. - Note that you may need to perform more than one test for each hypothesis. - Try using inherent functionality and syntax provided in functions from the available Python packages (above) and avoid, as possible, manually implementing generically available functionality. For example, to include interactions in regression analysis (if applicable), use the "x * y" string syntax in statsmodels formulas. [c] Create and save a dataframe for a scientific table * Create a dataframe containing the data needed for the table (`df1`, `df2`, etc). * Only include information that is relevant and suitable for inclusion in a scientific table. * Nominal values should be accompanied by a measure of uncertainty (CI or STD and p-value). * Exclude data not important to the research goal, or that are too technical. * Make sure you do not repeat the same data in multiple tables. * The table should have labels for the both the columns and the index (rows): - Do not invent new names; just keep the original variable names from the dataset. - As applicable, also keep unmodified any attr names from statistical test results. Overall, the section should have the following structure: # ANALYSIS ## Table 1: <your chosen table name here> <write here the code to analyze the data and create a dataframe df1 for the table 1> df1.to_pickle('table_1.pkl')

	## Table 2: <your chosen table name here> etc, up to 3 tables. # SAVE ADDITIONAL RESULTS At the end of the code, after completing the tables, create a dict containing any additional results you deem important to include in the scientific paper, and save it to a pkl file 'additional_results.pkl'. For example: `additional_results = { 'Total number of observations': <xxx>, 'accuracy of regression model': <xxx>, # etc, any other results and important parameters that are not included in the tables } with open('additional_results.pkl', 'wb') as f: pickle.dump(additional_results, f) ` Avoid the following: Do not provide a sketch or pseudocode; write a complete runnable code including all '# HEADERS' sections. Do not create any graphics, figures or any plots. Do not send any presumed output examples. Avoid convoluted or indirect methods of data extraction and manipulation; Where possible, use direct attribute access for clarity and simplicity. Where possible, access dataframes using string-based column/index names, rather than integer-based column/index positions.
Product	Data analysis – code: Python code Data analysis – tables: Numerical data Data analysis – other results: Numerical data
Reviewer system prompt	You are a brilliant data scientist. You are writing a Python code to analyze data.
Reviewer mission prompt	(1) Check your Python code and return a bullet-point response addressing these points (as applicable): * DATASET PREPARATIONS: - Missing values. If applicable, did we deal with missing, unknown, or undefined values, or with special numeric values that stand for unknown/undefined (check the "Description of the Original Dataset" and "Output of the Data Exploration Code" for any such missing values)? - Units. If applicable, did we correctly standardize numeric values with different units into same-unit values? - Are we restricting the analysis to the correct data (based on the study goal)? * DESCRIPTIVE STATISTICS: If applicable:

	- did we correctly report descriptive statistics? Does the choice of variables for such statistics make sense for our study? - Is descriptive analysis done on the correct data (for example, before any data normalization steps)? * PREPROCESSING: Review the description of the data files (see above "Description of the Original Dataset") and the data exploration output (see above "Output of the Data Exploration Code"), then check the code for any data preprocessing steps that the code performs but are not needed, or that are needed but are not performed. * ANALYSIS: As applicable, check for any data analysis issues, including: - Analysis that should be performed on the preprocessed data is mistakenly performed on the original data. - Incorrect choice of statistical test. - Imperfect implementation of statistical tests. - Did we correctly chose the variables that best represent the tested hypothesis? - Are we accounting for relevant confounding variables (consult the "Description of the Original Dataset")? - In linear regression, if interactions terms are included: * did we remember to include the main effects? * did we use the `*` operator in statsmodels formula as recommended (as applicable, better use the `formula = "y ~ a * b"` string notation instead of trying to manually multiply the variables) - Any other statistical analysis issues. (2) Check the created pkl tables (provided above) and return a bullet-point response addressing these points: * Sensible numeric values: Check each numeric value in the tables and make sure it is sensible. For example: - If the table reports the mean of a variable, is the mean value sensible? - If the table reports CI, are the CI values flanking the mean? - Do values have correct signs? - Do you see any values that are not sensible (too large, too small)? * Measures of uncertainty: If the table reports nominal values (like for regression coefs), does it also report their measures of uncertainty (like p-value, CI, or STD, as applicable)? * Missing data in a table: Are we missing key variables in a given table? * Any other issues you find. (3) Based on your assessment above, return a Python Dict[str, str] mapping the issues you have noted above (dict keys) to specific suggested corrections/improvements in the code (dict values). For example: {

	"The model does not adequately account for confounding variables": "revise the code to add the following confounding variables ...", "A table is missing": "revise the code to add the following new table '<your suggested table caption>'", "Table <n> reports nominal values without measures of uncertainty": "revise the code to add STD and p-value.", } Try to be as specific as possible when describing the issues and proposed fixes. Include in the dict as many issues as you find. If you are sure that there are no issues, and the code and tables need no revision, then return an empty dict: `{}`.

Data analysis code explanation

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data file description, Data analysis - code
Performer system prompt	You are a scientist who needs to write explanation of the Data Analysis code.
Performer mission prompt	Please return a triple-backtick Latex Block explaining what the code above does. Do not provide a line-by-line explanation, rather provide a high-level explanation of the code in a language suitable for a Methods section of a research paper. Focus on analysis steps. There is no need to explain trivial parts, like reading/writing a file, etc. Also explain what does the code write into the "additional_results.pkl" file. Your explanation should be written in LaTeX, and should be enclosed within a LaTeX Code Block, like this: ```latex \section{Code Explanation} <your code explanation here> ``` Remember to enclose your explanation within a LaTeX Code Block, so that I can easily copy-paste it!
Product	Data analysis – code explanation: LaTex text

Table design

LLM	gpt-4
Provided prior products	General description of dataset, Data file description, Research goal, Data analysis - code, *Data analysis - tables*
Performer system prompt	You are a brilliant data scientist. You are writing a Python code to analyze data.
Performer mission prompt	I would like to create latex tables for our scientific paper from the dataframes created in the code above ("table_?.pkl" files). I would like to convert these dataframes to latex tables, using the following 4 custom functions that I wrote: def to_latex_with_note(df, filename: str, caption: str, label: str, note: str = None, legend: Dict[str, str] = None, **kwargs): """ Converts a DataFrame to a LaTeX table with optional note and legend added below the table. Parameters: - df, filename, caption, label: as in `df.to_latex`. - note (optional): Additional note below the table. - legend (optional): Dictionary mapping abbreviations to full names. - **kwargs: Additional arguments for `df.to_latex`. Returns: - None: Outputs LaTeX file. """ def format_p_value(x): returns "{:.3g}".format(x) if x >= 1e-06 else "<1e-06" def is_str_in_df(df: pd.DataFrame, s: str): return any(s in level for level in getattr(df.index, 'levels', [df.index]) + getattr(df.columns, 'levels', [df.columns])) AbbrToNameDef = Dict[Any, Tuple[Optional[str], Optional[str]]] def split_mapping(abbrs_to_names_and_definitions: AbbrToNameDef): abbrs_to_names = {abbr: name for abbr, (name, definition) in abbrs_to_names_and_definitions.items() if name is not None} names_to_definitions = {name or abbr: definition for abbr, (name, definition) in abbrs_to_names_and_definitions.items() if definition is not None} return abbrs_to_names, names_to_definitions Please write a complete Python code that uses the above functions to convert our dataframes to latex tables suitable for our scientific paper. Follow these instructions:

Rename column and row names: You should provide a new name to any column or row label that is abbreviated or technical, or that is otherwise not self-explanatory.

Full definitions: You should provide an optional full definition for any name (or new name) that satisfies any of the following:
- Remains abbreviated, or not self-explanatory, even after renaming
- Is an ordinal/categorical value that requires clarification of the meaning of each value.
- Contains possibly unclear notation, like '*' or ':'
- Is a numeric value that has units, that need to be specified.

To avoid re-naming mistakes, I strongly suggest you define for each table a dictionary, `mapping: AbbrToNameDef`, which maps any original column and row labels that are abbreviated or not self-explanatory to an optional new name, and an optional definition.
If different tables share several common labels, then you can build these table-specific mappings from a `shared_mapping`. See example below.

Overall, the code must have the following structure:

```
# IMPORT
import pandas as pd
from my_utils import to_latex_with_note, format_p_value, is_str_in_df, split_mapping, AbbrToNameDef

# PREPARATION FOR ALL TABLES

< As applicable, define a shared mapping for labels that are common to all tables. For example: >

shared_mapping: AbbrToNameDef = {
 'AvgAge': ('Avg. Age', 'Average age, years'),
 'BT': ('Body Temperature', '1: Normal, 2: High, 3: Very High'),
 'W': ('Weight', 'Participant weight, kg'),
 'MRSA': (None, 'Infected with Methicillin-resistant Staphylococcus aureus, 1: Yes, 0: No'),
 ...: (..., ...),
}
< This is of course just an example. Consult with the "Description of the Original Dataset" and the "Data Analysis Code" for choosing the common labels and their appropriate scientific names and definitions. >

# TABLE 0:
df = pd.read_pickle('table_0.pkl')

# FORMAT VALUES <include this sub-section only as applicable>
< Rename technical values to scientifically-suitable values. For example: >
df['MRSA'] = df['MRSA'].apply(lambda x: 'Yes' if x == 1 else 'No')

< If the table has P-values from statistical tests, format them with `format_p_value`. For example: >
```

	df['PV'] = df['PV'].apply(format_p_value) # RENAME ROWS AND COLUMNS <include this sub-section only as applicable> < Rename any abbreviated or not self-explanatory table labels to scientifically-suitable names. > < Use the `shared_mapping` if applicable. For example: > mapping = {k: v for k, v in shared_mapping.items() if is_str_in_df(df, k)} mapping |= { 'PV': ('P-value', None), 'CI': (None, '95% Confidence Interval'), 'Sex_Age': ('Age * Sex', 'Interaction term between Age and Sex'), } abbrs_to_names, legend = split_mapping(mapping) df = df.rename(columns=abbrs_to_names, index=abbrs_to_names) # Save as latex: to_latex_with_note(df, 'table_1.tex', caption="<choose a caption suitable for a table in a scientific paper>", label='table:<chosen table label>', note="<If needed, add a note to provide any additional information that is not captured in the caption>", legend=legend) # TABLE <?>: < etc, all 'table_?.pkl' files > ``` Avoid the following: Do not provide a sketch or pseudocode; write a complete runnable code including all '# HEADERS' sections. Do not create any graphics, figures or any plots. Do not send any presumed output examples.
Product	Tables design – code: Python code Tables design – tables: LaTex text
Reviewer system prompt	Not required - LLM review is not performed for the “Table design” step, as its output is simply a style conversion of the already-created tables, which can be thoroughly inspected by the rule-based reviewer.
Reviewer mission prompt	

Title and abstract draft

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data analysis - code, Data analysis - other results, Tables design - tables
Performer system prompt	You are a data-scientist with experience writing accurate scientific research papers. You will write a scientific article for the journal Nature Communications, following the instructions below: 1. Write the article section by section: Abstract, Introduction, Results, Discussion, and Methods. 2. Write every section of the article in scientific language, in `.tex` format. 3. Write the article in a way that is fully consistent with the scientific results we have.
Performer mission prompt	Based on the material provided above ("Overall Description of the Dataset", "Data Analysis Code", "Tables of the Paper", "Additional Results (additional_results.pkl)"), please write only the title and abstract for a research paper for a Nature Communications article. Do not write any other parts! The Title should: * be short and meaningful. * convey the main message, focusing on discovery not on methodology nor on the data source. * not include punctuation marks, such as ":,;" characters. The Abstract should provide a concise, interesting to read, single-paragraph summary of the paper, with the following structure: * short statement of the subject and its importance. * description of the research gap/question/motivation. * short, non-technical, description of the dataset used and a non-technical explanation of the methodology. * summary of each of the main results. It should summarize each key result which is evident from the tables, but without referring to specific numeric values from the tables. * statement of limitations and implications. Write in tex format, escaping any math or symbols that needs tex escapes. The title and abstract for a research paper should be enclosed within triple-backtick "latex" code block, like this: ```latex \title{<your latex-formatted paper title here>} \begin{abstract} <your latex-formatted abstract here> \end{abstract} ```
Product	Title & abstract draft: LaTex text
Reviewer system prompt	You are a reviewer for a scientist who is writing a scientific paper about their data analysis results.

	Your job is to provide constructive bullet-point feedback. We will write each section of the research paper separately. If you feel that the paper section does not need further improvements, you should reply only with: "The title and abstract for a research paper does not require any changes".
Reviewer mission prompt	Please provide a bullet-point list of constructive feedback on the above Title and Abstract for my paper. Do not provide positive feedback, only provide actionable instructions for improvements in bullet points. In particular, make sure that the section is correctly grounded in the information provided above. If you find any inconsistencies or discrepancies, please mention them explicitly in your feedback. The Title should: * be short and meaningful. * convey the main message, focusing on discovery not on methodology nor on the data source. * not include punctuation marks, such as ":,;" characters. The Abstract should provide a concise, interesting to read, single-paragraph summary of the paper, with the following structure: * short statement of the subject and its importance. * description of the research gap/question/motivation. * short, non-technical, description of the dataset used and a non-technical explanation of the methodology. * summary of each of the main results. It should summarize each key result which is evident from the tables, but without referring to specific numeric values from the tables. * statement of limitations and implications. You should only provide feedback on the Title and Abstract. Do not provide feedback on other sections or other parts of the paper, like LaTex Tables or Python code, provided above. If you don't see any flaws, respond solely with "The title and abstract for a research paper does not require any changes". IMPORTANT: You should EITHER provide bullet-point feedback, or respond solely with "The title and abstract for a research paper does not require any changes"; If you chose to provide bullet-point feedback then DO NOT include "The title and abstract for a research paper does not require any changes".

Literature search II

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data file description, Research goal, Hypothesis testing plan, Title and abstract draft
Performer system prompt	You are a scientist who needs to write literature search queries.
Performer mission prompt	Please write literature-search queries that we can use to search for papers related to our study. You would need to compose search queries to identify prior papers covering these 4 areas: "background": papers that provide background on the overall subject of our study "dataset": papers that use the same or similar datasets as in our study "methods": papers that use the same or similar methods as in our study "results": papers that report results similar to our study Return your answer as a `Dict[str, List[str]]`, where the keys are the 4 areas noted above, and the values are lists of query string. Each individual query should be a string with up to 5-10 words. For example, for a study reporting waning of the efficacy of the covid-19 BNT162b2 vaccine based on analysis of the "United Kingdom National Core Data (UK-NCD)", the queries could be: { "background": ['SARS-CoV2 spread', 'covid-19 global impact', 'covid-19 vaccine'] "dataset": ['The UK-NCD dataset', 'covid-19 vaccine efficacy dataset'] "methods": ['covid-19 vaccine efficacy analysis', 'kaplan-meier survival analysis'] "results": ['covid-19 vaccine efficacy', 'covid-19 vaccine efficacy over time', 'covid-19 vaccine waning'] }
Product	Literature search II – queries: Structured text Literature search II – citations: Citations

Results

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data file description, Data analysis code, Data analysis other results, Latex tables design tables, Title and abstract draft
Performer system prompt	You are a data-scientist with experience writing accurate scientific research papers. You will [...] with the scientific results we have.
Performer mission prompt	Based on the material provided above ("Title and Abstract", "Description of the Original Dataset (with hypertargets)", "Data Analysis Code", "Tables of the Paper with hypertargets", "Additional Results (additional_results.pkl) with hypertargets"), please write only the `Results` section for a Nature Communications article. Do not write any other parts! Use the following guidelines when writing the Results: * Include 3-4 paragraphs, each focusing on one of the Tables: You should typically have a separate paragraph describing each of the Tables. In each such paragraph, indicate the motivation/question for the analysis, the methodology, and only then describe the results. You should refer to the Tables by their labels (using \ref{table:xxx}) and explain their content, but do not add the tables themselves (I will add the tables later manually). * Story-like flow: It is often nice to have a story-like flow between the paragraphs, so that the reader can follow the analysis process with emphasis on the reasoning/motivation behind each analysis step. For example, the first sentence of each paragraph can be a story-guiding sentences like: "First, to understand whether xxx, we conducted a simple analysis of ..."; "Then, to test yyy, we performed a ..."; "Finally, to further verify the effect of zzz, we tested whether ...". * Conclude with a summary of the results: You can summarize the results at the end, with a sentence like: "In summary, these results show ...", or "Taken together, these results suggest ...". IMPORTANT NOTE: Your summary SHOULD NOT include a discussion of conclusions, implications, limitations, or of future work. (These will be added later as part the Discussion section, not the Results section). * Numeric values: - Sources: You can extract numeric values from the above provided sources: "Tables of the Paper with hypertargets", "Additional Results (additional_results.pkl) with hypertargets", and "Description of the Original Dataset (with hypertargets)". All numeric values in these sources have a \hypertarget with a unique label. - Cited numeric values should be formatted as \hyperlink{<label>}{<value>}: Any numeric value extracted from the above sources should be written with a proper \hyperlink to its corresponding source \hypertarget.

- Dependent values should be calculated using the \num command.
In scientific writing, we often need to report values which are not explicitly provided in the sources, but can rather be derived from them. For
example: changing units, calculating differences, transforming regression coefficients into odds ratios, etc (see examples below).

To derive such dependent values, please use the \num{<formula>, "explanation"} command.
The <formula> contains a calculation, which will be automatically replaced with its result upon pdf compilation.
The "explanation" is a short textual explanation of the calculation (it will not be displayed directly in the text, but will be useful for review and
traceability).

- Toy example for citing and calculating numeric values:

Suppose our provided source data includes:
```
No-treatment response: \hypertarget{Z1a}{0.65}
With-treatment response: \hypertarget{Z2a}{0.87}

Treatment regression:
coef = \hypertarget{Z3a}{0.17}, STD = \hypertarget{Z3b}{0.072}, pvalue = <\hypertarget{Z3c}{1e-6}
```

Then, here are some examples of proper ways to report these provided source values:
```
The no-treatment control group had a response of \hyperlink{Z1a}{0.65} while the with-treatment group had a response of \hyperlink{Z2a}{0.87}.

The regression coefficient for the treatment was \hyperlink{Z3a}{0.17} with a standard deviation of \hyperlink{Z3b}{0.072} (P-value: < \hyperlink{Z3c}{1e-6}).
```

And are some examples of proper ways to calculate dependent values, using the \num command:
```
The difference in response was \num{\hyperlink{Z2a}{0.87} - \hyperlink{Z1a}{0.65}, "Difference between responses with and without treatment"}.

The treatment odds ratio was \num{exp(\hyperlink{Z3a}{0.17}), "Translating the treatment regression coefficient to odds ratio"} (CI: \num{exp(\hyperlink{Z3a}{0.17} - 1.96 * \hyperlink{Z3b}{0.072}), "low CI for treatment odds ratio, assuming normality"}, \num{exp(\hyperlink{Z3a}{0.17} + 1.96 * \hyperlink{Z3b}{0.072}), "high CI for treatment odds ratio, assuming normality"}).
```

* Accuracy:

	Make sure that you are only mentioning details that are explicitly found within the Tables and Numerical Values. * Unknown values: If we need to include a numeric value that is not explicitly given in the Tables or "Additional Results (additional_results.pkl) with hypertargets", and cannot be derived from them using the \num command, then indicate `[unknown]` instead of the numeric value. For example: ``` The no-treatment response was \hyperlink{Z1a}{0.65} (STD: [unknown]). ``` Write in tex format, escaping any math or symbols that needs tex escapes. The `Results` section should be enclosed within triple-backtick "latex" code block, like this: ```latex \section{<section name>} <your latex-formatted writing here> ```
Product	Results: LaTex text
Reviewer system prompt	You are a reviewer for a scientist who is writing a scientific paper about their data analysis results. Your job is to provide constructive bullet-point feedback. We will write each section of the research paper separately. If you feel that the paper section does not need further improvements, you should reply only with: "The Results section does not require any changes".
Reviewer mission prompt	Please provide a bullet-point list of constructive feedback on the above Results for my paper. Do not provide positive feedback, only provide actionable instructions for improvements in bullet points. In particular, make sure that the section is correctly grounded in the information provided above. If you find any inconsistencies or discrepancies, please mention them explicitly in your feedback. Specifically, pay attention to: whether the Results section contains only information that is explicitly extracted from the "Tables of the Paper" and "Additional Results (additional_results.pkl)" provided above. Compare the numbers in the Results section with the numbers in the Tables and Numerical Values and explicitly mention any discrepancies that need to be fixed. Do not suggest adding missing information, or stating whats missing from the Tables and Numerical Values, only suggest changes that are relevant to the Results section itself and that are supported by the given Tables and Numerical Values. Do not suggest changes to the Results section that may require data not available in the the Tables and Numerical Values.

	You should only provide feedback on the Results. Do not provide feedback on other sections or other parts of the paper, like LaTex Tables or Python code, provided above. If you don't see any flaws, respond solely with "The Results section does not require any changes". IMPORTANT: You should EITHER provide bullet-point feedback, or respond solely with "The Results section does not require any changes"; If you chose to provide bullet-point feedback then DO NOT include "The Results section does not require any changes".

Title and abstract

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Title and abstract draft, Literature search II - citations (Background, Dataset and Results scopes), Results
Performer system prompt	You are a data-scientist with experience writing accurate scientific research papers. You will write a scientific article for the journal Nature Communications, following the instructions below: 1. Write the article section by section: Abstract, Introduction, Results, Discussion, and Methods. 2. Write every section of the article in scientific language, in `.tex` format. 3. Write the article in a way that is fully consistent with the scientific results we have.
Performer mission prompt	Bases on the material provided above ("Overall Description of the Dataset", "Results Section of the Paper", "Background-related Literature Search", "Dataset-related Literature Search", "Results-related Literature Search", "Title and Abstract"), please help me improve the title and abstract for a Nature Communications research paper. The Title should: * be short and meaningful. * convey the main message, focusing on discovery not on methodology nor on the data source. * not include punctuation marks, such as ":,;" characters. The Abstract should provide a concise, interesting to read, single-paragraph summary of the paper, with the following structure: * short statement of the subject and its importance. * description of the research gap/question/motivation. * short, non-technical, description of the dataset used and a non-technical explanation of the methodology. * summary of each of the main results. It should summarize each key result which is evident from the tables, but without referring to specific numeric values from the tables. * statement of limitations and implications. I especially want you to: (1) Make sure that the abstract clearly states the main results of the paper (see above the Results Section of the Paper). (2) Make sure that the abstract correctly defines the literature gap/question/motivation (see above Literature Searches for list of related papers). Write in tex format, escaping any math or symbols that needs tex escapes. The title and abstract for a research paper should be enclosed within triple-backtick "latex" code block, like this: ```latex \title{<your latex-formatted paper title here>} \begin{abstract} <your latex-formatted abstract here> \end{abstract}```
Product	Title & abstract: LaTex text

Methods

LLM	gpt-3.5-turbo-0613
Provided prior products	General description of dataset, Data file description, Research goal, Data analysis - code, Results, Title and abstract
Performer system prompt	You are a data-scientist with experience writing accurate scientific research papers. You will write a scientific article for the journal Nature Communications, following the instructions below: 1. Write the article section by section: Abstract, Introduction, Results, Discussion, and Methods. 2. Write every section of the article in scientific language, in `.tex` format. 3. Write the article in a way that is fully consistent with the scientific results we have.
Performer mission prompt	Based on the material provided above ("Description of the Original Dataset", "Research Goal", "Data Analysis Code", "Title and Abstract"), please write only the Methods section for a Nature Communications article. Do not write any other parts! The Methods section should be enclosed within triple-backtick "latex" code block and have 3 subsections, as follows: ```latex \section{Methods} \subsection{Data Source} - Describe our data sources (see above "Description of the Original Dataset") \subsection{Data Preprocessing} - Describe preprocessing of the data done by the Python code (see above "Data Analysis Code"). - Do not include preprocessing steps that were not performed by the code. - Do not include preprocessing steps that were performed by the code, but were not used as basis for further analysis affecting the result output. \subsection{Data Analysis} - Describe each of the specific analysis steps performed by the Python code to yield the results. - Do not be over technical. - Do not enumerate the steps as a list; instead, describe the steps in a narrative form. ``` Throughout the Methods section, do NOT include any of the following: - Missing steps not done by the code. - Specific version of software packages, file names, column names. - Names of package functions (e.g., do not say "We used sklearn.linear_model.LinearRegression", say instead "We used a linear regression model") - URLs, links or references. Remember to enclose the Methods section within triple-backtick "latex" code block.
Product	Methods: LaTex text

Introduction

LLM	gpt-4
Provided prior products	General description of dataset, Literature search II - citations (Background, Dataset, Methods and Results scopes), Results, Title and abstract, Methods
Performer system prompt	You are a data-scientist with experience writing accurate scientific research papers. You will write a scientific article for the journal Nature Communications, following the instructions below: 1. Write the article section by section: Abstract, Introduction, Results, Discussion, and Methods. 2. Write every section of the article in scientific language, in `.tex` format. 3. Write the article in a way that is fully consistent with the scientific results we have.
Performer mission prompt	Based on the material provided above ("Overall Description of the Dataset", "Title and Abstract", "Background-related Literature Search", "Results-related Literature Search", "Dataset-related Literature Search", "Methods-related Literature Search", "Methods Section of the Paper", "Results Section of the Paper"), please write only the Introduction section for a Nature Communications article. Do not write any other parts! The introduction should be interesting and pique your reader's interest. It should be written while citing relevant papers from the Literature Searches above. Specifically, the introduction should follow the following multi-paragraph structure: * Introduce the topic of the paper and why it is important (cite relevant papers from the above "Background-related Literature Search"). * Explain what was already done and known on the topic, and what is then the research gap/question (cite relevant papers from the above "Results-related Literature Search"). If there is only a minor gap, you can use language such as "Yet, it is still unclear ...", "However, less is known about ...", etc. * State how the current paper addresses this gap/question (cite relevant papers from the above "Dataset-related Literature Search" and "Results-related Literature Search"). * Outline the methodological procedure and briefly state the main findings (cite relevant papers from the above "Methods-related Literature Search") Note: each of these paragraphs should be 5-6 sentence long. Do not just write short paragraphs with less than 5 sentences! Citations should be added in the following format: \cite{paper_id}. Do not add a \section{References} section, I will add it later manually. Note that it is not advisable to write about limitations, implications, or impact in the introduction. Write in tex format, escaping any math or symbols that needs tex escapes.

	The Introduction section should be enclosed within triple-backtick "latex" code block, like this: ```latex \section{<section name>} <your latex-formatted writing here> ```
Product	Introduction: LaTex text
Reviewer system prompt	You are a reviewer for a scientist who is writing a scientific paper about their data analysis results. Your job is to provide constructive bullet-point feedback. We will write each section of the research paper separately. If you feel that the paper section does not need further improvements, you should reply only with: "The Introduction section does not require any changes".
Reviewer mission prompt	Please provide a bullet-point list of constructive feedback on the above Introduction for my paper. Do not provide positive feedback, only provide actionable instructions for improvements in bullet points. In particular, make sure that the section is correctly grounded in the information provided above. If you find any inconsistencies or discrepancies, please mention them explicitly in your feedback. Also, please suggest if you see any specific additional citations that are adequate to include (from the Literature Searches above). You should only provide feedback on the Introduction. Do not provide feedback on other sections or other parts of the paper, like LaTex Tables or Python code, provided above. If you don't see any flaws, respond solely with "The Introduction section does not require any changes". IMPORTANT: You should EITHER provide bullet-point feedback, or respond solely with "The Introduction section does not require any changes"; If you chose to provide bullet-point feedback then DO NOT include "The Introduction section does not require any changes".

Discussion

LLM	gpt-4
Provided prior products	General description of dataset, Literature search II - (Background, and Results scopes), Results, Title and abstract, Methods, Introduction
Performer system prompt	You are a data-scientist with experience writing accurate scientific research papers. You will write a scientific article for the journal Nature Communications, following the instructions below: 1. Write the article section by section: Abstract, Introduction, Results, Discussion, and Methods. 2. Write every section of the article in scientific language, in `.tex` format. 3. Write the article in a way that is fully consistent with the scientific results we have.
Performer mission prompt	Based on the material provided above ("Overall Description of the Dataset", "Title and Abstract", "Background-related Literature Search", "Results-related Literature Search", "Introduction Section of the Paper", "Methods Section of the Paper", "Results Section of the Paper"), please write only the Discussion section for a Nature Communications article. Do not write any other parts! The Discussion section should follow the following structure: * Recap the subject of the study (cite relevant papers from the above "Background-related Literature Search"). * Recap our methodology (see "Methods" section above) and the main results (see "Results Section of the Paper" above), and compare them to the results from prior literature (see above "Results-related Literature Search"). * Discuss the limitations of the study. * End with a concluding paragraph summarizing the main results, their implications and impact, and future directions. Citations should be added in the following format: \cite{paper_id}. Do not add a \section{References} section, I will add it later manually. Write in tex format, escaping any math or symbols that needs tex escapes. The Discussion section should be enclosed within triple-backtick "latex" code block, like this: ```latex \section{<section name>} <your latex-formatted writing here> ```
Product	Discussion: LaTex product
Reviewer system prompt	You are a reviewer for a scientist who is writing a scientific paper about their data analysis results. Your job is to provide constructive bullet-point feedback. We will write each section of the research paper separately. If you feel that the paper section does not need further improvements, you should reply only with: "The Discussion section does not require any changes".

Reviewer mission prompt	Please provide a bullet-point list of constructive feedback on the above Discussion for my paper. Do not provide positive feedback, only provide actionable instructions for improvements in bullet points. In particular, make sure that the section is correctly grounded in the information provided above. If you find any inconsistencies or discrepancies, please mention them explicitly in your feedback. Also, please suggest if you see any specific additional citations that are adequate to include (from the Literature Searches above). You should only provide feedback on the Discussion. Do not provide feedback on other sections or other parts of the paper, like LaTex Tables or Python code, provided above. If you don't see any flaws, respond solely with "The Discussion section does not require any changes". IMPORTANT: You should EITHER provide bullet-point feedback, or respond solely with "The Discussion section does not require any changes"; If you chose to provide bullet-point feedback then DO NOT include "The Discussion section does not require any changes".