Supplementary Dataset A. "Health Indicators" dataset.

Supplementary Dataset B. "Social network" dataset.

Supplementary Dataset C. "Treatment Policy" dataset.

Supplementary Dataset D. "Treatment Optimization" dataset.

Supplementary Manuscripts A1-5, B1-5, C1-10, Da1-10, Db1-10, Dc1-10, Dai1-10, Dbi1-10, Dci1-10. All manuscripts produced by data-to-paper based on the "Health Indicators" dataset (A1-5), the "Social Network" dataset (B1-5), the "Treatment Policy" dataset (C1-10) and the "Treatment Optimization" dataset with its respective goals (Da1-10, Db1-10, Dc1-10, Dai1-10, Dbi1-10, Dci1-10). Text has been manually highlighted as follows: green for good practice, e.g. putting findings into context or good representations of results, and correctly used numerical values in the text; yellow for atypical, but not erroneous practice, orange for minor mistakes which do not affect the overall message of the paper, and red for fundamental mistakes, affecting the results or conclusions of the paper.

Supplementary Runs A1-5, B1-5, C1-10, Da1-10, Db1-10, Dc1-10, Dai1-10, Dbi1-10, Dci1-10. Run files are color-coded terminal output files of the conversations of each of the created papers (html format). They include all data-to-paper research steps, each represented by one or two (in case of a review step) distinct LMM conversations (step starts are marked with 'Starting conversation', name of the conversation, purple). Each message in a conversation starts with a header, which includes (a) message number in the conversation (square brackets); (b) message attribution (Methods; in capital letters: SYSTEM, USER, ASSISTANT, SURROGATE or COMMENTER); (c) casting agent (agent name enclosed in curly brackets used for system testing); (d) conversation name (preceded by `->` symbol). Triple-backtick text within messages is highlighted (brighter for text, or Python code syntax pigmentation). Messages are classified and color-coded as follows: (a) SYSTEM and USER messages in green; (b) SURROGATE messages in turquoise (Methods); (c) COMMENTER messages in blue, which are meant only as comments and are not sent to ChatGPT; (d) ASSISTANT message in bright turquoise, including a header listing all prior messages included in the conversation as sent to the API(*30*). Of note, to save space, messages that have already been presented before appear as a single line with "[...]". In addition to the conversation messages, the file also contains the following alerts (in red): (a) API(*30*) calls including the LLM model and number of tokens; (b) indications of conversation message deletions; (c) API(*30*) failed responses; (d) model bumping; (e) check of numerical values comparison; and (in blue) (f) Citation retrieved from Semantic Scholar Academic Graph API(*28*), presenting the information about a successful retrieval of citations or a summary of the information retrieved for each of the citations.

Supplementary Coding Runs. Run files of evaluation of coding capabilities of different LLMs (Fig. S3B). File names indicate the underlying LLM. Same annotation as for Supplementary Runs A-D.

Supplementary Data-chained Manuscripts A-D. Data-chained example manuscripts produced by data-to-paper for each dataset. All numerical values are hyperlinked from the text to the code which produced it. For values which have been transformed in the text, an

explanation is provided in the “Notes” section. Automatically created hyperlinks can be found in the “Results” section in the text, in the “Notes” appendix, in the table headers, and in the appendix output file headers, which ultimately lead to the relevant code section.

Supplementary Human Co-piloted Manuscripts 1-3. Human co-piloted example manuscripts. These manuscripts were created in fixed-goal modality with “Treatment Optimization” dataset (*27*) (Supplementary Data Description Da; Supplementary Dataset D; Methods).