

1. The sources or owners of the datasets.

Medical records retrieved by Gemini Legal. Approximately 20,000 total records.

2. A description of how the datasets further the intended purpose of the artificial intelligence system or service.

Medical records are used to fine-tuned open-weight LLMs to identify medical appointments, for the purposes of creating a chronology of appointments.

3. The number of data points included in the datasets, which may be in general ranges, and with estimated figures for dynamic datasets.

Approximately 20,000 documents, totalling approximately 38,000,000 pages of medical record text.

4. A description of the types of data points within the datasets. For purposes of this paragraph, the following definitions apply:

(A) As applied to datasets that include labels, “types of data points” means the types of labels used.

(B) As applied to datasets without labeling, “types of data points” refers to the general characteristics.

Scanned copies of medical records provided by treating facilities, in PDF format.

5. Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.

Neither protected by copyright, trademark, patent, nor in the public domain.

6. Whether the datasets were purchased or licensed by the developer.

Neither.

(7) Whether the datasets include personal information, as defined in subdivision (v) of Section 1798.140. That definition is as follows:

"Information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household. Personal information includes, but is not limited to, the following if it identifies, relates to, describes, is reasonably capable of being associated with, or could be reasonably linked, directly or indirectly, with a particular consumer or household:

(A) Identifiers such as a real name, alias, postal address, unique personal identifier, online identifier, Internet Protocol address, email address, account name, social security number, driver's license number, passport number, or other similar identifiers.

(B) Any personal information described in subdivision (e) of Section 1798.80.

(C) Characteristics of protected classifications under California or federal law.

(D) Commercial information, including records of personal property, products or services purchased, obtained, or considered, or other purchasing or consuming histories or tendencies.

(E) Biometric information.

(F) Internet or other electronic network activity information, including, but not limited to, browsing history, search history, and information regarding a consumer's interaction with an internet website application, or advertisement.

(G) Geolocation data.

(H) Audio, electronic, visual, thermal, olfactory, or similar information.

(I) Professional or employment-related information.

(J) Education information, defined as information that is not publicly available personally identifiable information as defined in the Family Educational Rights and Privacy Act (20 U.S.C. Sec. 1232g; 34 C.F.R. Part 99).

(K) Inferences drawn from any of the information identified in this subdivision to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes.

(L) Sensitive personal information."

Personal Health Information - including names, addresses, emails, phone numbers, SSNs, and extensive medical history data.

8. Whether the datasets include aggregate consumer information, as defined in subdivision (b) of Section 1798.140. That definition is as follows: Information that relates to a group or category of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a device.

'Aggregate consumer information' does not mean one or more individual consumer records that have been deidentified."

No.

9. Whether there was any cleaning, processing, or other modification to the datasets by the developer, including the intended purpose of those efforts in relation to the artificial intelligence system or service.

Yes, extensive cleaning and processing to format the data for AI training use.

10. The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing.

2023

11. The dates the datasets were first used during the development of the artificial intelligence system or service.

April 1, 2024

12. Whether the generative artificial intelligence system or service used or continuously uses synthetic data generation in its development. A developer may include a description of the functional need or desired purpose of the synthetic data in relation to the intended purpose of the system or service.

Yes, synthetic data generation is used for medical appointment classification.