Supplementary Material-1: Examples of Historical Entities

Table S1: Definitions of CC/HPI/PFSH concepts and associated examples according to the 1997 E/M Guidelines.

History	Concept	Definitions	Examples
CC		A brief statement on the reason for a medical encounter, usually stated in the patient's own words.	Migraine
HPI	Location	Anatomical location of the problem	Right-side of my head
	Quality	The nature of the problem	Constant, aching pain
	Severity	A degree or measurement of how bad the problem is	5 on a scale of 1-10
	Duration	How long the problem has been existing	It started two days ago
	Timing	The temporal pattern of the problem, such as when and how frequently it occurs.	One or two episodes of pains per day, all happened in afternoons
	Context	The patient's activities, environmental factors, and/or circumstances surrounding the problem.	I stayed up late every night to catch up a deadline last week
	Modifying Factors	Factors that make the problem better or worse	Better when heat is applied
	Associated Signs or Symptoms	Symptoms or signs that accompany the problem	Vomiting and fatigue
PFSH	Past Medical History	The patient's past experiences with illness, treatments, and operations	Pre-diabetes and hypertension
	Family History	Medical events in the patient's family, such as diseases which are hereditary or place the patient at risk	Father has migraine as well
	Social History	An age-appropriate review of the patient's social activities	Nonsmoker, drinks occasionally

Selected examples showing the polysemous nature of historical concepts. Entities are highlighted with **red, bold** font.

- HPI-Modifying Factors can include multiple sub-concepts like Medications, Environment Factors, and Actions:
  - Medication: "The patient returns to our office today because of continued problems with her headaches. … She takes **Motrin** 400 mg b.i.d., which helped." in sample_1128.txt.
  - Environmental Factor: "Her husband has been **hauling corn and this seems to aggravate things**." in sample_343.txt.
  - Action: "She has little bit of paraesthesias in the left toe as well and seem to involve all the toes of the right foot. They are not worse with walking. It seems to be **worse when she is in bed**." in sample_365.txt.
- HPI-Associated Signs and Symptoms can include sub-concepts like Problems and Signs, among which Signs may not be immediately apparent.
  - Problem: "When questioned further, she described **shortness of breath** primarily with ambulation." in sample_398.txt.
  - Sign: "… the patient continues to **lose weight**." in sample_378.txt.
- Past Medical History can include multiple sub-concepts like Conditions and Treatments:
  - Condition: "PAST MEDICAL HISTORY: "…, **diabetics** with a bad family history for **cardiovascular disease** such as this patient does have, …" in sample_1568.txt.
  - Treatments: "PAST MEDICAL HISTORY: … 5. **Removal of the melanoma from the right thigh** in 1984. …" in sample_380.txt.

Selected examples of entities that primarily consist of non-medical vocabulary.

- HPI-Context example: "This is a 32-year-old male who **had a piece of glass fall on to his right foot** today." in sample_2747.txt.

- Social History example: “SOCIAL HISTORY: … He has what sounds like a **data entry computer job**.” in sample_1152.txt.
- HPI-Quality example: “This is a one plus-month-old female with respiratory symptoms … This involved cough, … The coughing persisted and **worsened**.” in sample_1956.txt.
- HPI-Timing example: “This is an 18-month-old white male here with his mother for complaint of intermittent fever … Mother states that his temperature usually **elevates at night**.” in sample_439.txt

Selected examples of entities that did not occur in the designated headed sections.

- Example of an HPI-Duration entity in the CC section: “CHIEF COMPLAINT: Cough and abdominal pain for **two days**.” in sample_930.txt.
- Example of an HPI-Location entity in the CC section: “CHIEF COMPLAINT: Pressure decubitus, **right hip**. HISTORY OF PRESENT ILLNESS: This is a 30-year-old female patient presenting with the above chief complaint.” in sample_687.txt.
- Examples of Past Medical History entities in the HPI/Subjective section:
  - “SUBJECTIVE: This is a 78-year-old male who recently had his **right knee replaced** and also **back surgery** about a year and a half ago.” in sample_70.txt.
  - “BRIEF HISTORY OF PRESENT ILLNESS: ... The patient has history of **hypercholesterolemia**, …” in sample_96.txt.
  - “HISTORY OF PRESENT ILLNESS: This is the initial clinic visit for a 29-year-old man … He has **no upper extremity**.” in sample_223.txt.
- Examples of CC entities in the HPI/Subjective section:
  - “HISTORY OF PRESENT ILLNESS: This is the initial clinic visit for a 29-year-old man who is seen for new onset of **right shoulder pain**.” in sample_223.txt.
  - “HISTORY OF PRESENT ILLNESS: This is the initial clinic visit for a 41-year-old worker who is seen for a **foreign body to his left eye**.” in sample_225.txt.
- Examples of Social History entities in the HPI/Subjective section:
  - “CHIEF COMPLAINT: Blood-borne pathogen exposure. HISTORY OF PRESENT ILLNESS: The patient is a 54-year-old right-handed male who **works as a phlebotomist and respiratory therapist at Hospital**.” in sample_226.txt.
  - “HISTORY OF PRESENT ILLNESS: The patient is a 55-year-old Caucasian female … The patient is a **nonsmoker**.” in sample_380.txt.
- Examples of Family History entities in the HPI/Subjective section:
  - “SUBJECTIVE: The patient is a 7-year-old male who comes in today with a three-day history of emesis and a four-day history of diarrhea. Apparently, **his brother had similar symptoms**.” in sample_388.txt.
  - “HISTORY OF PRESENT ILLNESS: This 60-year-old white male is referred to us … There is a history of **gallstone pancreatitis** in the patient's sister.” in sample_2780.txt.

Supplementary Material-2: Model Development and Specifications

Prompt for GPT-4o

Task

Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare which are in doctor's note or clinical note.

The entities to be identified are: 'cc', 'hpi.location', 'hpi.quality', 'hpi.severity', 'hpi.duration', 'hpi.timing', 'hpi.context', 'hpi.modifyingFactors', 'hpi.assocSignsAndSymptoms', 'pastHistory','familyHistory', 'socialHistory'.

If a sentence has negation words, entities might not need to be identified.

Use HTML tags to highlight these entities. Each should have a class attribute indicating the type of the entity.

Entity Markup Guide

Use to denote a chief complain entity in the clinical note

Use to denote an entity related to the location of a symptom or condition in the history of present illness.

Use to denote an entity related to the quality or character of a symptom in the history of present illness.

Use to denote an entity related to the severity of a symptom or condition in the history of present illness.

Use to denote an entity related to how long a symptom or condition has been present in the history of present illness.

Use to denote an entity related to the timing or frequency of a symptom in the history of present illness.

Use to denote an entity related to the context or circumstances surrounding a symptom or condition in the history of present illness.

Use to denote an entity related to factors that make a symptom better or worse in the history of present illness.

Use to denote an entity related to associated signs and symptoms present with the condition.

Use to denote an entity related to the patient's past medical history.

Use to denote an entity related to the patient's family history.

Use to denote an entity related to the patient's social history, such as lifestyle or habits.

Leave the text as it is if no such entities are found.

cPLMs Evaluation

- **Bio+Discharge Summary BERT & Bio+Clinical BERT**. Both the models were initialized with BioBERT, specifically, starting with BioBERT's architecture and parameters ($\text{BERT}_{\text{BASE}}$), which include 12 transformer layers and approximately 110 million parameters. Afterward, Bio+Discharge Summary BERT was fine-tuned on the discharge summaries in the MIMIC-III EHR repository, while Bio+Clinical BERT underwent fine-tuning with all clinical notes from MIMIC-III. The fine-tuning process involved further adjusting the model's parameters based on the clinical corpora provided, enabling the models to specialize in understanding the semantics within the given categories of clinical text.
- **BiomedBERT and BiomedBERT large (formerly PubMedBERT & PubMedBERT large)**. The PubMedBERT model was created by training the $\text{BERT}_{\text{BASE}}$ architecture, which consists of 12 layers and 110 million parameters, on the abstracts from PubMed and the full-text articles from PubMed Central. The development of PubMedBERT Large was based on the $\text{BERT}_{\text{LARGE}}$ architecture with 24 layers and 340 million parameters, but utilizing only the abstracts from PubMed.
- **BioMegatron**. The BioMegatron model was initiated with Megatron-LM, then fine-tuned using the abstracts from PubMed and the full-text articles from PubMedCentral. While different sizes of BioMegatron are available, we employed the one with 24 layers and 345 million parameters in this study ($\text{BERT}_{\text{LARGE}}$ architecture).
- **GatorTronBase & GatorTronS**. Trained on a variety of copra including PubMed, WikiText, MIMIC-III clinical notes, and de-identified clinical notes from the University of Florida Health System (for GatorTronBase) or synthetic clinical words generated by GPT-3 model (for GatorTronS), GatorTronBase and GatorTronS stand out as two of the most extensively trained cLLMs. The GatorTronBase and GatorTronS models we used in this study, each had 24 layers and 345 million parameters ($\text{BERT}_{\text{LARGE}}$ architecture).

The Two Fine-Tuning Approaches

- **BIO Fine-Tuning**: It is achieved by building a fully connected output layer on top of a cLLM [4]. Specifically, the process begins with feeding textual data from clinical notes into the cLLM in batches, with each batch containing eight sentences. Given an m-token batch, the cLLM transforms the tokens into m vectors of n-dimensions (n depends on the specific cLLM used), denoted as $\{V_1, V_2, \dots, V_m\}$. Each vector represents the embedding and positional encoding of a token. These vectors are then passed to the output layer for classification. Given that 12 medical history entities (MHEs) resulted in a total of 25 *"B-<MHE>", "I-<MHE>",* and *"O"* tags, the output layer comprises 25 neurons corresponding to the 25 BIO tags. Each neuron contains n weights. The classification will select the BIO tag corresponding to the neuron that has the highest dot summation between the embedding vector and the weights. Thus, the output layer has $n \times 25$ weights, denoted as $W \in \mathbb{R}^{n \times 25}$, which need to be trained for optimal classification. Let $\{Y_1, Y_2, \dots, Y_m\}$ be the BIO tags of the m-token batch, the training process is directed towards determining the optimal W that minimizes the following cross-entropy loss function.

$$\min_{W \in \mathbb{R}^{n \times 25}} L = -\sum_{i=1}^{m} CrossEntropy_i(V_i, Y_i; W)$$

Furthermore, we applied the dropout technique to mitigate the potential overfitting [11]. In our implementation, dropout was used to randomly deactivate 10% elements of the embedding vector output by the cLLM.

- **BIO Fine-Tuning with Pre-Identified Basic Medical Entities (BMEs)**: The BMEs pre-identified by CLAMP can be categorized into two groups:
  - ***Group 1***: "problem", "test," "treatment", "drug"
  - ***Group 2***: "body location," "severity," and "temporal."

  These two groups of entities can overlap. For example, the term "chest pain" is classified as a "problem," whereas "chest" in this context refers to a "body location." To explore the potential of the preliminary BME information in enhancing the recognition of sophisticated CC/HPI/PFSH entities, we incorporated it into the cLLM fine-tuning as follows:

  1) ***New Tag Creation***: We added new BIO tags, specifically "*B-<BME>*," "*I-<BME>,"* and "*O,"* to each token based on the BME identification provided by CLAMP. Given the seven types of BMEs, this resulted in a total of 15 BIO tags. Note that, due to the overlapping of the two groups of CLAMP BMEs, it is possible that a token can have multiple tags. For instance, the "chest" in term "chest pain" can have both "B-problem" and "B-Body Location" tags.
  2) ***One-Hot Encoding***: We applied one-hot encoding to convert each token's BME-related BIO tag into a 14-dimensional vector. A value of 1 in a vector element indicates a specific "*B-<BME>*" or "*I-<BME>"* tag, while all 0s correspond to the *"O"* tag.
  3) ***Concatenation and Fine-Tuning***: The one-hot encoded vectors representing BME information were concatenated with the cLLM embeddings. Each concatenated vector was then fine-tuned towards the 25 historical entity-related BIO tags in a manner similar to the basic fine-tuning process described earlier.

All cPLMs were implemented using PyTorch and trained for 40 epochs. The code is available at https://github.com/isallexist/Patient-History-NER

Supplementary Material-3: Detail Results

Table S1: F1 scores among GPT-4o and 7 cPLMs. The F1 score follows the CoNLL-2003 Shared Task benchmark and is computed as a micro-averaged, entity-level F1. In addition to the standard Exact Match criterion, we also report performance under a Relaxed Match criterion, which credits partially overlapping entity spans. **Bold** scores indicate improvement with CLAMP support.

F1 score among 12 MHE concepts – EXACT MATCH															
	GPT-4o	**BioDischarge**		**BioClinical**		**Biomed**		**BioMegatron**		**BiomedLarge**		**GatortronBase**		**GatortronS**	
MHE	Zero-shot	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP
CC	0.22	0.51	**0.58**	0.45	**0.51**	0.50	**0.56**	0.54	**0.57**	0.54	**0.58**	0.64	0.64	0.60	**0.66**
HPI Location	0.13	0.33	0.30	0.24	**0.27**	0.27	0.23	0.29	0.26	0.32	**0.37**	0.35	**0.38**	0.37	**0.38**
HPI Quality	0.09	0.28	0.28	0.39	0.24	0.33	0.25	0.27	0.26	0.27	0.23	0.40	0.35	0.43	0.38
HPI Severity	0.32	0.14	0.14	0.20	0.13	0.18	0.05	0.13	0.11	0.10	0.00	0.24	0.16	0.22	0.05
HPI Duration	0.28	0.44	0.42	0.47	**0.48**	0.44	0.42	0.47	**0.52**	0.48	0.45	0.60	0.59	0.54	**0.59**
HPI Timing	0.16	0.43	0.37	0.45	0.44	0.37	0.26	0.44	0.36	0.35	**0.51**	0.48	0.44	0.43	0.41
HPI Context	0.03	0.04	0.08	0.00	0.04	0.00	0.05	0.00	0.00	0.00	0.00	0.19	0.12	0.08	0.04
HPI Modifying Factor	0.15	0.45	0.44	0.50	0.43	0.45	0.33	0.38	0.34	0.37	**0.44**	0.51	0.50	0.44	0.44
HPI Associated Sign and Symptom	0.18	0.30	0.30	0.29	0.25	0.31	0.27	0.32	0.25	0.35	0.31	0.38	0.35	0.38	0.34
Past Medical History	0.28	0.69	0.66	0.69	0.67	0.68	0.67	0.66	**0.67**	0.67	0.67	0.70	0.70	0.69	**0.71**
Family History	0.43	0.81	**0.83**	0.66	**0.76**	0.84	0.80	0.83	0.77	0.83	**0.84**	0.86	**0.87**	0.83	**0.86**
Social History	0.23	0.54	**0.56**	0.53	0.52	0.54	0.49	0.55	0.53	0.56	0.54	0.59	0.56	0.58	0.54

F1 score among 12 MHE concepts – RELAX MATCH															
	GPT-4o	**BioDischarge**		**BioClinical**		**Biomed**		**BioMegatron**		**BiomedLarge**		**GatortronBase**		**GatortronS**	
MHE	Zero-shot	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP	FT	FT w/ CLAMP
CC	0.31	0.60	**0.71**	0.51	**0.64**	0.58	**0.65**	0.59	**0.66**	0.64	**0.66**	0.70	**0.71**	0.66	**0.73**

HPI Location	0.26	0.56	0.54	0.40	**0.46**	0.42	0.37	0.50	0.48	0.52	0.47	0.66	0.65	0.61	0.59
HPI Quality	0.20	0.47	0.47	0.51	0.45	0.47	0.44	0.45	**0.46**	0.39	**0.42**	0.62	**0.65**	0.61	0.53
HPI Severity	0.40	0.26	0.22	0.24	0.24	0.30	0.26	0.35	0.25	0.15	0.10	0.35	**0.41**	0.42	0.15
HPI Duration	0.47	0.63	0.60	0.64	0.62	0.54	**0.55**	0.66	**0.70**	0.59	**0.62**	0.75	**0.76**	0.73	**0.75**
HPI Timing	0.32	0.43	0.37	0.47	0.44	0.37	0.26	0.47	0.36	0.38	**0.51**	0.60	0.52	0.46	**0.47**
HPI Context	0.12	0.28	**0.37**	0.33	0.22	0.32	0.21	0.25	0.09	0.28	**0.36**	0.48	0.47	0.41	0.30
HPI Modifying Factor	0.21	0.48	0.47	0.54	0.46	0.48	0.41	0.46	0.44	0.42	**0.46**	0.58	**0.59**	0.48	**0.51**
HPI Associated Sign and Symptom	0.23	0.46	0.39	0.45	0.39	0.43	0.37	0.44	0.37	0.46	0.41	0.50	0.47	0.48	0.47
Past Medical History	0.39	0.73	0.72	0.74	0.73	0.73	0.72	0.72	**0.73**	0.72	0.72	0.75	0.75	0.74	**0.77**
Family History	0.65	0.88	**0.90**	0.73	**0.77**	0.88	0.81	0.89	0.81	0.89	0.88	0.92	0.91	0.89	**0.90**
Social History	0.43	0.83	**0.84**	0.82	0.81	0.80	**0.82**	0.84	0.81	0.87	0.83	0.87	0.86	0.87	**0.82**

Table S2: Detailed counts and rates of matching results for each CC/HPI/PFSH concept. Highlights: Integration of BMEs identified by CLAMP reduced the errors. ***Red italic bold font***: The lowest error rate among all models assessed for each MHE concept.

Abbreviations: M.F. (modifying factors), Past H. (past medical history), Fam. H. (family history), Social H. (social history). EM (Exact Match), RM (Relaxed Match), MM(Mismatch), UD (Under Detection), OD (Over Detection).

		OpenAI GPT-4o					
	Total entities	EM	RM	MM	UD	OD	Error rate
CC	133	53	27	38	15	306	73.0%
Location	48	24	26	3	2	283	85.7%
Quality	53	8	12	9	25	124	74.9%
Severity	34	16	5	5	8	50	64.9%
Duration	66	25	21	5	15	85	61.4%
Timing	37	7	8	10	13	42	63.7%
Context	37	1	3	14	19	26	61.5%
M.F.	81	9	4	10	58	28	54.2%
Symptom	268	54	17	74	124	268	63.5%
Past H.	518	120	56	19	323	205	51.4%
Fam. H.	45	23	18	0	5	39	49.4%
Social H.	129	23	25	4	82	42	49.8%

		Bio_Discharge						**Bio_Discharge + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**
CC	133	51	14	40	29	16	39.0%	66	23	30	18	24	35.1%
Location	48	13	12	9	16	15	45.5%	11	12	9	19	11	44.8%
Quality	53	11	10	9	25	14	47.5%	11	10	7	26	14	47.0%
Severity	34	3	3	6	22	7	50.7%	3	2	11	18	7	51.4%
Duration	66	27	17	0	23	29	44.1%	25	15	0	27	26	44.5%
Timing	37	13	0	11	13	11	48.6%	11	0	14	12	12	50.7%
Context	37	1	7	17	12	13	53.2%	2	9	18	9	10	50.0%
M.F.	81	34	3	11	33	36	49.7%	31	3	14	33	29	48.4%
Symptom	268	74	49	27	122	143	52.1%	71	27	37	136	130	53.1%
Past H.	518	351	38	19	113	152	35.4%	338	48	15	125	158	36.5%
Fam. H.	45	34	5	2	5	4	19.6%	35	5	1	5	3	16.7%
Social H.	129	60	54	5	20	22	26.7%	61	52	4	20	20	25.4%

		Bio_Clinical						**Bio_Clinical + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**
CC	133	43	8	44	38	16	42.4%	56	20	36	24	27	39.5%
Location	48	9	8	8	25	17	51.0%	10	9	13	19	12	47.8%
Quality	53	17	7	7	23	17	47.0%	11	12	8	26	22	51.4%
Severity	34	5	1	10	18	10	52.8%	3	3	8	20	9	52.1%
Duration	66	31	16	2	20	32	45.0%	30	13	1	22	30	44.5%
Timing	37	15	1	9	12	15	49.3%	14	0	10	13	13	49.3%
Context	37	0	10	19	10	12	52.6%	1	5	18	14	10	53.2%
M.F.	81	39	5	2	35	37	47.7%	34	4	10	35	43	52.1%
Symptom	268	73	55	29	118	164	53.7%	60	43	45	126	151	54.6%
Past H.	518	346	40	27	109	137	34.5%	350	48	21	105	173	36.6%
Fam. H.	45	25	4	11	5	6	32.8%	31	1	8	5	6	29.7%
Social H.	129	58	53	8	21	19	27.1%	61	59	7	19	29	29.9%

		BioMedLarge						**BioMedLarge + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**

CC	133	56	16	37	27	17	37.9%	62	13	33	27	16	36.4%
Location	48	13	11	10	17	18	48.4%	13	5	7	25	8	45.5%
Quality	53	11	6	10	27	17	50.5%	9	9	10	27	13	48.5%
Severity	34	2	1	17	14	4	50.7%	0	2	10	22	6	52.8%
Duration	66	31	10	2	24	32	46.8%	27	15	2	22	28	44.1%
Timing	37	10	1	14	12	10	49.3%	16	0	10	11	10	45.6%
Context	37	0	8	18	12	11	52.6%	0	11	20	7	12	51.3%
M.F.	81	31	5	9	37	53	55.0%	33	2	3	43	37	50.6%
Symptom	268	83	36	30	123	125	50.9%	76	30	27	141	133	52.9%
Past H.	518	340	45	17	119	161	36.4%	336	40	19	126	150	36.3%
Fam. H.	45	36	5	3	3	4	18.2%	36	3	3	4	4	19.6%
Social H.	129	58	56	6	21	7	***20.9%***	57	53	6	25	13	25.4%
		BioMegatron						**BioMegatron + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**
CC	133	57	8	36	33	22	40.6%	64	15	32	25	26	38.4%
Location	48	12	12	8	19	21	50.0%	10	11	7	22	16	48.4%
Quality	53	10	9	7	29	10	46.5%	10	10	8	28	11	47.0%
Severity	34	3	6	6	20	8	50.0%	2	3	7	22	1	46.9%
Duration	66	30	19	1	20	29	43.1%	32	16	1	19	22	38.9%
Timing	37	14	1	11	11	12	47.9%	11	0	8	18	13	51.3%
Context	37	0	7	20	12	10	53.2%	0	2	17	18	4	51.3%
M.F.	81	30	8	13	33	45	52.9%	25	10	9	38	42	52.4%
Symptom	268	80	37	39	118	140	52.6%	57	33	34	145	127	53.3%
Past H.	518	344	48	15	117	167	36.6%	342	46	9	127	157	36.1%
Fam. H.	45	35	4	2	5	3	18.2%	32	3	5	5	6	26.2%
Social H.	129	55	50	7	24	10	24.1%	57	53	5	25	20	27.9%
		BioMed						**BioMed + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**
CC	133	47	11	47	29	8	38.7%	62	16	35	23	25	38.4%
Location	48	10	7	11	21	15	49.5%	9	7	11	22	21	52.9%
Quality	53	14	8	8	23	18	48.0%	10	10	10	26	15	49.0%
Severity	34	4	3	14	14	5	49.3%	1	5	10	19	6	50.7%
Duration	66	28	9	1	28	33	48.4%	27	11	2	27	34	48.8%
Timing	37	12	0	12	13	16	52.6%	7	0	14	16	9	51.3%
Context	37	0	9	20	10	9	51.3%	1	4	18	15	4	50.0%
M.F.	81	34	3	13	32	36	50.0%	23	7	15	39	31	51.2%
Symptom	268	80	38	30	123	158	53.7%	63	27	39	141	128	53.5%
Past H.	518	355	40	18	108	173	36.6%	348	44	15	115	171	36.8%
Fam. H.	45	35	3	4	3	3	18.2%	33	1	8	3	5	26.2%
Social H.	129	60	49	5	23	26	29.5%	55	62	5	23	25	29.1%
		GatortronBase						**GatortronBase + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**
CC	133	69	11	38	16	14	33.8%	72	12	31	19	20	34.5%
Location	48	14	18	7	16	10	***40.7%***	15	15	7	16	10	***40.7%***
Quality	53	17	14	5	19	14	41.8%	15	19	4	20	13	***41.1%***
Severity	34	6	3	9	16	9	50.0%	4	8	4	20	10	50.0%
Duration	66	40	16	1	12	24	35.9%	37	18	1	12	21	34.0%
Timing	37	16	6	9	8	12	***43.9%***	14	3	9	12	11	46.4%
Context	37	5	10	13	9	10	***46.4%***	3	12	13	12	9	47.9%
M.F.	81	39	7	10	27	30	45.3%	36	10	8	29	26	***43.8%***
Symptom	268	95	41	26	109	132	49.9%	77	37	35	124	94	***48.6%***
Past H.	518	343	45	18	113	122	32.8%	347	39	11	123	118	32.7%
Fam. H.	45	39	5	0	2	6	***15.1%***	39	4	0	2	6	***15.1%***
Social H.	129	63	53	6	17	12	21.3%	61	55	6	20	13	23.2%
		GatortronS						**GatortronS + CLAMP**					
	Total entities	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error rate**
CC	133	67	11	25	30	24	37.3%	75	14	24	22	19	***32.8%***
Location	48	14	13	8	17	10	42.2%	13	10	8	19	5	***40.0%***
Quality	53	19	11	4	21	14	42.4%	15	8	7	25	9	43.6%
Severity	34	5	6	8	16	6	***46.9%***	1	2	6	25	2	49.3%
Duration	66	34	19	0	18	21	37.1%	35	15	0	19	14	***33.3%***
Timing	37	13	1	9	14	10	47.1%	12	2	8	15	9	46.4%
Context	37	2	11	13	13	12	50.7%	1	7	15	15	8	50.7%
M.F.	81	32	4	7	39	32	49.1%	31	6	9	35	28	47.1%
Symptom	268	92	33	27	120	125	50.4%	78	40	23	132	110	49.7%
Past H.	518	341	40	14	124	126	33.8%	343	48	12	117	101	***30.7%***
Fam. H.	45	36	5	1	4	5	18.2%	38	3	0	4	5	16.7%
Social H.	129	64	55	4	17	16	22.3%	56	49	2	31	13	26.3%

Table S3: Detailed number of labels, length (in words), counts and error rate of matching result for each sample for GatorTronBase and GatorTronS models. Abbreviation: Anno #: number of annotations (labels); Word #: number of words in the sample

GatorTronS									
Sample	**Anno #**	**Word #**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error**	**Error rate**
sample_1128	33	492	18	2	2	11	0	13	39.39%
sample_1133	24	782	8	2	1	13	1	15	60.00%
sample_1152	18	593	10	0	4	4	2	10	50.00%
sample_1169	8	422	4	0	1	3	1	5	55.56%
sample_1242	11	536	3	0	1	7	4	12	80.00%
sample_1248	31	910	13	1	5	12	16	33	70.21%
sample_1252	19	585	11	5	2	2	3	7	30.43%
sample_1419	10	540	3	4	2	2	1	5	41.67%
sample_1439	19	896	3	2	4	10	1	15	75.00%
sample_1495	45	872	17	0	2	26	1	29	63.04%
sample_1505	16	784	6	3	3	4	6	13	59.09%
sample_1568	28	1112	19	3	2	4	13	19	46.34%
sample_1592	24	785	16	2	0	6	4	10	35.71%
sample_1921	21	789	14	3	1	3	10	14	45.16%
sample_1956	14	628	9	1	2	2	13	17	62.96%
sample_2129	13	427	4	4	1	4	12	17	68.00%
sample_214	14	586	5	4	4	2	6	12	57.14%
sample_2210	38	874	6	7	0	27	7	34	72.34%
sample_2218	35	771	18	12	2	5	10	17	36.17%
sample_223	11	567	4	2	2	4	7	13	68.42%
sample_225	13	468	4	4	1	4	3	8	50.00%
sample_226	8	1022	3	2	1	2	19	22	81.48%
sample_2275	27	612	8	2	5	12	3	20	66.67%
sample_2604	20	733	5	5	3	8	5	16	61.54%
sample_2623	14	480	6	1	2	5	10	17	70.83%
sample_2746	31	2066	24	3	1	3	18	22	44.90%
sample_2747	13	814	5	2	0	6	3	9	56.25%
sample_2780	48	994	20	17	6	13	6	25	40.32%
sample_2789	34	1082	26	5	1	3	18	22	41.51%
sample_2790	34	1123	21	5	3	6	24	33	55.93%
sample_2792	9	708	5	1	1	2	27	30	83.33%
sample_343	25	1075	13	3	4	5	11	20	55.56%
sample_365	57	1312	41	8	5	4	12	21	30.00%
sample_377	55	1420	22	9	9	16	11	36	53.73%
sample_378	24	939	10	2	2	10	8	20	62.50%
sample_380	50	1139	22	2	2	24	5	31	56.36%
sample_388	17	620	6	4	2	5	10	17	62.96%
sample_391	20	918	6	1	3	10	8	21	75.00%
sample_392	34	1004	14	2	2	16	3	21	56.76%
sample_393	50	1552	28	6	0	16	6	22	39.29%
sample_394	19	676	8	0	5	6	1	12	60.00%
sample_398	24	694	17	3	0	4	4	8	28.57%
sample_402	15	491	9	1	3	2	2	7	41.18%
sample_403	17	562	10	6	0	2	3	5	23.81%
sample_439	16	411	5	2	1	8	0	9	56.25%
sample_452	15	398	10	2	1	2	2	5	29.41%
sample_476	17	666	7	3	2	5	5	12	54.55%
sample_485	24	742	10	3	2	10	3	15	53.57%
sample_570	31	1038	16	9	3	5	12	20	44.44%
sample_579	30	932	19	2	3	6	7	16	43.24%
sample_583	29	927	15	6	0	8	9	17	44.74%
sample_664	19	885	13	1	0	5	2	7	33.33%
sample_666	40	873	22	6	3	11	4	18	39.13%
sample_687	20	426	7	8	1	7	0	8	34.78%
sample_70	7	519	3	0	0	4	0	4	57.14%
sample_71	12	562	3	3	0	7	2	9	60.00%
sample_782	10	355	2	1	0	7	0	7	70.00%
sample_930	32	853	27	4	1	0	6	7	18.42%
sample_942	10	285	8	0	0	2	0	2	20.00%
sample_945	24	766	16	3	1	4	7	12	38.71%
sample_96	23	759	12	5	0	7	4	11	39.29%
N	61		T-stat (e_count)	T-stat (e_rates)			correlation	0.635	-0.048
DF	59		6.313	-0.373			p-values	3.89E-08	0.711
GatorTronS + CLAMP									
Sample	**Anno #**	**Word #**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error**	**Error rate**

sample_1128	33	492	16	1	1	15	0	16	48.48%
sample_1133	24	782	10	1	1	12	1	14	56.00%
sample_1152	18	593	10	1	2	5	2	9	45.00%
sample_1169	8	422	3	2	1	3	0	4	44.44%
sample_1242	11	536	3	0	2	6	4	12	80.00%
sample_1248	31	910	13	0	5	13	10	28	68.29%
sample_1252	19	585	10	4	0	6	1	7	33.33%
sample_1419	10	540	2	5	2	2	1	5	41.67%
sample_1439	19	896	4	4	2	9	3	14	63.64%
sample_1495	45	872	17	1	4	23	1	28	60.87%
sample_1505	16	784	6	3	2	5	7	14	60.87%
sample_1568	28	1112	19	3	2	4	11	17	43.59%
sample_1592	24	785	16	2	0	6	3	9	33.33%
sample_1921	21	789	12	3	1	6	3	10	40.00%
sample_1956	14	628	5	1	1	7	10	18	75.00%
sample_2129	13	427	3	5	2	4	3	9	52.94%
sample_214	14	586	4	2	3	5	4	12	66.67%
sample_2210	38	874	5	3	0	30	0	30	78.95%
sample_2218	35	771	16	8	2	9	3	14	36.84%
sample_223	11	567	3	1	2	5	0	7	63.64%
sample_225	13	468	4	2	0	7	0	7	53.85%
sample_226	8	1022	1	2	1	4	10	15	83.33%
sample_2275	27	612	9	1	3	14	2	19	65.52%
sample_2604	20	733	4	2	0	15	1	16	72.73%
sample_2623	14	480	6	0	3	5	6	14	70.00%
sample_2746	31	2066	23	6	1	2	17	20	40.82%
sample_2747	13	814	5	1	0	7	2	9	60.00%
sample_2780	48	994	21	11	4	14	6	24	42.86%
sample_2789	34	1082	23	6	1	5	13	19	39.58%
sample_2790	34	1123	22	8	1	5	19	25	45.45%
sample_2792	9	708	5	1	1	2	28	31	83.78%
sample_343	25	1075	12	6	3	4	8	15	45.45%
sample_365	57	1312	39	9	6	4	8	18	27.27%
sample_377	55	1420	21	8	8	19	14	41	58.57%
sample_378	24	939	11	4	2	8	7	17	53.13%
sample_380	50	1139	21	3	1	26	4	31	56.36%
sample_388	17	620	5	3	3	6	12	21	72.41%
sample_391	20	918	2	4	4	10	6	20	76.92%
sample_392	34	1004	13	2	2	17	2	21	58.33%
sample_393	50	1552	31	6	1	14	9	24	39.34%
sample_394	19	676	8	0	4	7	3	14	63.64%
sample_398	24	694	18	4	0	2	3	5	18.52%
sample_402	15	491	9	2	3	1	2	6	35.29%
sample_403	17	562	11	4	0	2	3	5	25.00%
sample_439	16	411	5	2	2	7	1	10	58.82%
sample_452	15	398	9	1	1	4	1	6	37.50%
sample_476	17	666	6	6	2	3	4	9	42.86%
sample_485	24	742	10	2	2	10	4	16	57.14%
sample_570	31	1038	17	5	5	5	15	25	53.19%
sample_579	30	932	16	7	2	6	3	11	32.35%
sample_583	29	927	18	5	0	6	12	18	43.90%
sample_664	19	885	12	2	2	3	2	7	33.33%
sample_666	40	873	23	7	2	11	5	18	37.50%
sample_687	20	426	8	6	3	5	0	8	36.36%
sample_70	7	519	5	1	0	1	1	2	25.00%
sample_71	12	562	6	1	0	5	2	7	50.00%
sample_782	10	355	3	1	0	6	0	6	60.00%
sample_930	32	853	24	3	3	2	7	12	30.77%
sample_942	10	285	8	1	1	0	1	2	18.18%
sample_945	24	766	14	3	2	5	10	17	50.00%
sample_96	23	759	13	6	0	5	3	8	29.63%
N	61		T-stat (e_count)	T-stat (e_rates)			correlation	0.624	-0.030
DF	59		6.132	-0.232			p-values	7.79E-08	0.817
GatorTronBase									
Sample	**Anno #**	**Word #**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error**	**Error rate**
sample_1128	33	492	18	1	2	12	2	16	45.71%
sample_1133	24	782	11	2	1	10	5	16	55.17%
sample_1152	18	593	11	1	3	3	1	7	36.84%
sample_1169	8	422	4	0	1	3	0	4	50.00%
sample_1242	11	536	3	1	1	6	5	12	75.00%
sample_1248	31	910	9	5	7	10	16	33	70.21%
sample_1252	19	585	9	7	1	3	1	5	23.81%

sample_1419	10	540	3	4	2	2	1	5	41.67%
sample_1439	19	896	4	3	4	8	4	16	69.57%
sample_1495	45	872	18	0	4	23	2	29	61.70%
sample_1505	16	784	7	3	4	2	7	13	56.52%
sample_1568	28	1112	21	2	2	3	15	20	46.51%
sample_1592	24	785	19	2	0	3	3	6	22.22%
sample_1921	21	789	12	5	2	2	10	14	45.16%
sample_1956	14	628	9	2	0	3	14	17	60.71%
sample_2129	13	427	3	5	1	4	12	17	68.00%
sample_214	14	586	7	2	5	0	5	10	52.63%
sample_2210	38	874	9	2	1	26	2	29	72.50%
sample_2218	35	771	21	8	4	3	5	12	29.27%
sample_223	11	567	4	5	2	2	4	8	47.06%
sample_225	13	468	5	7	1	2	2	5	29.41%
sample_226	8	1022	3	1	2	2	13	17	80.95%
sample_2275	27	612	12	4	2	9	3	14	46.67%
sample_2604	20	733	5	3	4	9	6	19	70.37%
sample_2623	14	480	9	0	3	2	7	12	57.14%
sample_2746	31	2066	23	7	0	2	19	21	41.18%
sample_2747	13	814	4	0	0	9	1	10	71.43%
sample_2780	48	994	23	12	4	14	4	22	38.60%
sample_2789	34	1082	25	5	2	4	16	22	42.31%
sample_2790	34	1123	22	4	2	6	21	29	52.73%
sample_2792	9	708	4	1	2	2	26	30	85.71%
sample_343	25	1075	13	8	3	3	10	16	43.24%
sample_365	57	1312	39	12	5	3	16	24	32.00%
sample_377	55	1420	27	7	10	11	10	31	47.69%
sample_378	24	939	10	3	1	10	3	14	51.85%
sample_380	50	1139	23	2	3	22	2	27	51.92%
sample_388	17	620	7	3	4	3	9	16	61.54%
sample_391	20	918	4	4	6	6	8	20	71.43%
sample_392	34	1004	14	7	2	12	5	19	47.50%
sample_393	50	1552	32	2	3	13	10	26	43.33%
sample_394	19	676	6	1	4	8	2	14	66.67%
sample_398	24	694	17	2	1	4	2	7	26.92%
sample_402	15	491	9	3	2	1	2	5	29.41%
sample_403	17	562	9	6	3	0	2	5	25.00%
sample_439	16	411	7	1	3	5	2	10	55.56%
sample_452	15	398	8	4	2	2	2	6	33.33%
sample_476	17	666	5	7	2	3	5	10	45.45%
sample_485	24	742	7	6	3	10	6	19	59.38%
sample_570	31	1038	15	11	3	3	14	20	43.48%
sample_579	30	932	20	2	4	4	6	14	38.89%
sample_583	29	927	17	3	1	8	12	21	51.22%
sample_664	19	885	12	0	3	4	3	10	45.45%
sample_666	40	873	28	4	1	9	6	16	33.33%
sample_687	20	426	8	8	0	7	0	7	30.43%
sample_70	7	519	4	0	0	3	1	4	50.00%
sample_71	12	562	2	1	0	9	2	11	78.57%
sample_782	10	355	4	2	0	4	0	4	40.00%
sample_930	32	853	26	3	2	1	10	13	30.95%
sample_942	10	285	8	1	1	0	1	2	18.18%
sample_945	24	766	15	5	1	3	7	11	35.48%
sample_96	23	759	13	7	0	4	5	9	31.03%
N	61		T-stat (e_count)	T-stat (e_rates)			correlation	0.665	0.042
DF	59		6.838	0.324			p-values	5.08E-09	0.747
GatorTronBase + CLAMP									
Sample	**Anno #**	**Word #**	**EM**	**RM**	**MM**	**UD**	**OD**	**Error**	**Error rate**
sample_1128	33	492	21	3	1	8	3	12	33.33%
sample_1133	24	782	12	2	1	9	6	16	53.33%
sample_1152	18	593	8	3	3	4	3	10	47.62%
sample_1169	8	422	4	0	2	2	1	5	55.56%
sample_1242	11	536	3	2	1	6	3	10	66.67%
sample_1248	31	910	10	5	8	9	19	36	70.59%
sample_1252	19	585	9	7	1	3	1	5	23.81%
sample_1419	10	540	5	4	2	0	2	4	30.77%
sample_1439	19	896	5	6	4	5	6	15	57.69%
sample_1495	45	872	17	2	6	20	5	31	62.00%
sample_1505	16	784	7	3	4	2	7	13	56.52%
sample_1568	28	1112	20	4	2	3	18	23	48.94%
sample_1592	24	785	18	2	0	4	3	7	25.93%
sample_1921	21	789	12	3	1	5	6	12	44.44%

sample_1956	14	628	4	0	1	9	6	16	80.00%
sample_2129	13	427	3	4	2	4	10	16	69.57%
sample_214	14	586	3	4	5	2	3	10	58.82%
sample_2210	38	874	8	2	0	28	2	30	75.00%
sample_2218	35	771	18	6	2	9	4	15	38.46%
sample_223	11	567	4	3	1	4	2	7	50.00%
sample_225	13	468	4	8	0	5	1	6	33.33%
sample_226	8	1022	2	2	2	2	10	14	77.78%
sample_2275	27	612	12	2	2	11	3	16	53.33%
sample_2604	20	733	3	4	0	13	1	14	66.67%
sample_2623	14	480	8	0	2	4	3	9	52.94%
sample_2746	31	2066	23	6	2	1	18	21	42.00%
sample_2747	13	814	5	0	1	7	0	8	61.54%
sample_2780	48	994	22	9	4	15	5	24	43.64%
sample_2789	34	1082	24	6	2	4	14	20	40.00%
sample_2790	34	1123	20	8	1	7	20	28	50.00%
sample_2792	9	708	4	2	1	2	27	30	83.33%
sample_343	25	1075	13	5	2	5	9	16	47.06%
sample_365	57	1312	38	8	5	7	11	23	33.33%
sample_377	55	1420	20	9	10	17	7	34	53.97%
sample_378	24	939	11	3	0	10	1	11	44.00%
sample_380	50	1139	23	1	1	25	2	28	53.85%
sample_388	17	620	3	5	3	7	11	21	72.41%
sample_391	20	918	6	4	5	6	8	19	65.52%
sample_392	34	1004	11	5	3	17	3	23	58.97%
sample_393	50	1552	32	3	1	14	8	23	39.66%
sample_394	19	676	7	2	4	7	3	14	60.87%
sample_398	24	694	19	4	0	2	2	4	14.81%
sample_402	15	491	9	3	3	0	3	6	33.33%
sample_403	17	562	13	2	1	1	3	5	25.00%
sample_439	16	411	6	1	1	8	2	11	61.11%
sample_452	15	398	7	2	3	4	2	9	50.00%
sample_476	17	666	6	7	1	3	4	8	38.10%
sample_485	24	742	9	2	4	10	8	22	66.67%
sample_570	31	1038	15	12	3	4	15	22	44.90%
sample_579	30	932	18	4	2	6	5	13	37.14%
sample_583	29	927	19	3	1	6	7	14	38.89%
sample_664	19	885	13	1	4	1	2	7	33.33%
sample_666	40	873	23	7	2	11	3	16	34.78%
sample_687	20	426	7	8	1	7	0	8	34.78%
sample_70	7	519	5	1	0	1	1	2	25.00%
sample_71	12	562	3	6	0	5	3	8	47.06%
sample_782	10	355	3	1	0	6	0	6	60.00%
sample_930	32	853	27	0	2	3	6	11	28.95%
sample_942	10	285	8	0	1	1	1	3	27.27%
sample_945	24	766	15	3	2	4	7	13	41.94%
sample_96	23	759	13	8	0	4	2	6	22.22%
N	61		T-stat (e_count)	T-stat (e_rates)			correlation	0.615	0.004
DF	59		5.998	0.028			p-values	1.30E-07	0.978

Table S4: The occurrences of MMUD for entities inside and outside of dedicated sections with headers.

		GatorTronBase			GatorTronBase+CLAMP			GatorTronS			GatorTronS+CLAMP		
Entity	**Match**	**In**	**Out**	***p*-value**	**In**	**Out**	***p*-value**	**In**	**Out**	***p*-value**	**In**	**Out**	***p*-value**
CC	EM+RM	40	40	0.009	44	40	0.000	39	39	0.008	47	42	0.000
	MMUD	14	40		10	40		14	41		7	39	
HPI	EM+RM	268	79	0.945	243	80	0.475	237	72	0.872	214	62	1.000
	MMUD	227	69		254	72		259	75		279	82	
Fam. H.	EM+RM	40	4	0.208	40	3	0.172	39	2	0.006	39	2	0.034
	MMUD	1	1		1	1		2	3		2	2	
Past H.	EM+RM	280	108	0.000	283	103	0.000	282	99	0.000	285	106	0.000
	MMUD	75	56		73	61		73	65		71	58	
Social H.	EM+RM	96	20	0.000	97	19	0.000	97	22	0.000	90	15	0.000
	MMUD	9	14		11	15		9	12		14	19	