UNIVERSITY OF CAMBRIDGE

BAND: Biomedical Alert News Dataset

Zihao Fu, Meiru Zhang, Zaiqiao Meng,

Yannan Shen, David Buckeridge, Nigel Collier

https://github.com/fuzihaofzh/BAND/

Language Technology Lab, University of Cambridge
School of Computing Science, University of Glasgow
School of Population and Global Health, McGill University

Introduction

- Problem: Lack of sophisticated epidemiological data sets for natural language processing.
- Solution: Introducing the Biomedical Alert News Dataset (BAND).
- BAND's composition: 1,508 samples, 30 epidemiology-related questions.

Infectious Disease Surveillance

- Continued threat of infectious disease outbreaks.
- Mentioned Systems: BioCaster, GPHIN, ProMED-mail, HealthMap, EIOS.
- Limitations:
    - Focus mainly on detection, not in-depth epidemiological analysis.
    - Cannot identify cases with special scenarios (e.g., deliberate release, vulnerable populations).
    - Limited data for training machine learning systems.
    - Cannot do good detection without in-depth epidemiological analysis.

BAND Dataset

- 1,508 samples: News articles, emails, alerts.
- 30 epidemiology-related questions: Event-related queries and more detailed inquiries.
- Aim: Enhance NLP capability to support epidemiological surveillance by focusing on important details that matter to human analysts.

News

Las Vegas public health officials say dozens of people linked to a tuberculosis outbreak at a neonatal unit have tested positive for the disease. The Southern Nevada Health District reported on Monday that of the 977 people tested, 59 showed indications of the disease and 2 showed signs of being contagious…

Which infectious disease caused the outbreak?	tuberculosis
In which country is the outbreak taking place?	US
In which province is the outbreak taking place?	Nevada
In which city/town is the outbreak taking place?	Las Vegas
Did the outbreak involve the intentful release?	No
Are the victims healthcare workers?	Cannot Infer
Did victims acquire the disease from animals?	No
Did the outbreak happen after a natural disaster?	No
…	…

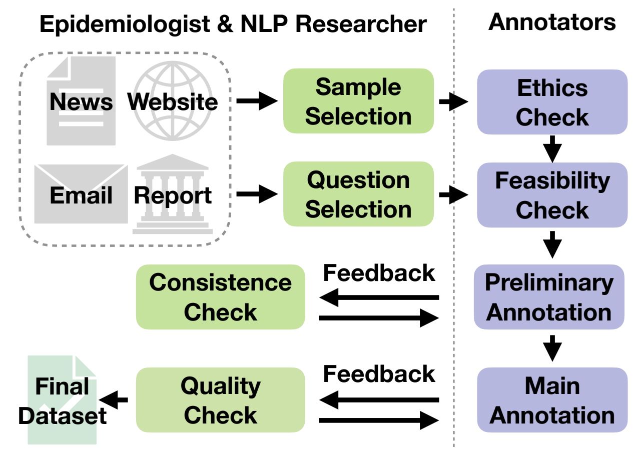
Benchmarks on BAND

- NLP Benchmarks on BAND
- Tasks highlighted:
  - Question Answering (QA)
  - Named Entity Recognition (NER)
  - Event Extraction (EE)
- Objective: Assess state-of-the-art models' capabilities.

Data Annotation

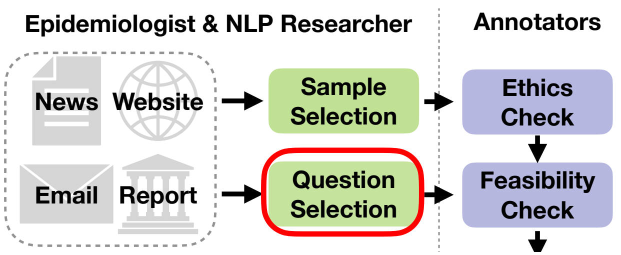
Data Annotation Process

- Question Selection
- Samples Selection
- Annotation
- Consistency Check
- Quality Check
- Ethics Check



Data Annotation Process

- Question Selection
  - Collaboration with experts in epidemiology and public health.
  - Categorized into:
    - Event questions
    - Epidemiology questions
    - Ethics questions
- Samples Selection
- Annotation
- Consistency Check
- Quality Check
- Ethics Check

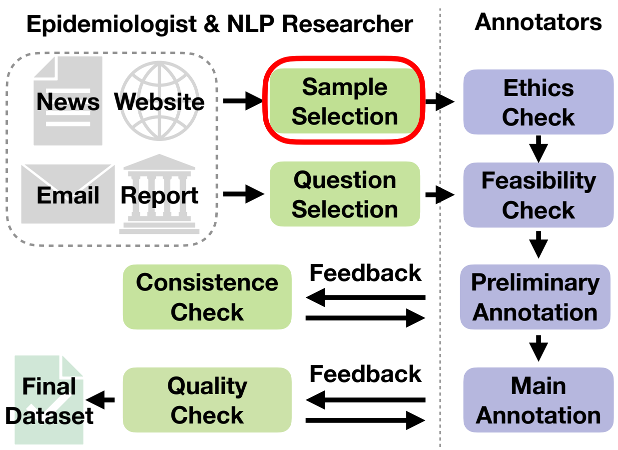


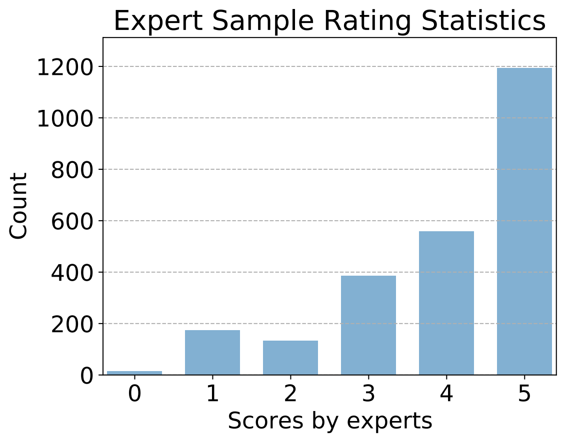
Questions	Short name	Category	Options	Sparse
1) Which infectious disease caused the outbreak?	Disease	Event	-	-
2) In which country is the outbreak taking place?	Country	Event	-	-
3) In which province is the outbreak taking place?	Province	Event	-	-
4) In which city/town is the outbreak taking place?	City	Event	-	-
5) Check and fill country Geo Code (e.g. 1794299):	Countrycode	Event	-	-
6) Check and fill province Geo Code (e.g. 1794299):	Provincecode	Event	-	-
7) Check and fill city Geo Code (e.g. 1815286):	Citycode	Event	-	-
8) Which virus or bacteria caused the outbreak?	Virus	Event	-	-
9) What symptoms were experienced by the infected victims?	Symptoms	Epidemiology	-	-
10) Which institution reported this outbreak?	Reporter	Epidemiology	-	-
11) What is the type of the victims?	Victimtype	Epidemiology	Human/Animal/Plant	-
12) How many new infected cases are reported in the specific event in the report? (please input digits like 1, 34, etc.)	Casesnum	Epidemiology	-	-
13) Has the victim of the disease travelled across international borders?	Borders	Epidemiology	YES/NO/Cannot Infer	YES
14) Does the outbreak involve the intentful release?	Intentful	Epidemiology	YES/NO/Cannot Infer	YES
15) Did human victims acquire the infectious disease from an animal?	Fromanimal	Epidemiology	YES/NO/Cannot Infer/Not Applicable	-
16) Did the victim fail to respond to a drug?	Faildrug	Epidemiology	YES/NO/Cannot Infer/Not Applicable	-
17) Are healthcare workers included in the infected victims?	Healthcareworkers	Epidemiology	YES/NO/Cannot Infer	YES
18) Are animal workers included in the infected victims?	Animalworkers	Epidemiology	YES/NO/Cannot Infer	YES
19) Is the victim of the disease a military worker?	Militaryworkers	Epidemiology	YES/NO/Cannot Infer	YES
20) Did the outbreak involve a suspected contaminated blood product or vaccine?	Vaccine	Epidemiology	YES/NO/Cannot Infer	YES
21) Are the victims in a group in time and place?	Group	Epidemiology	YES/NO/Cannot Infer/Not Applicable	-
22) Did the victim catch the disease during a hospital stay?	Hospitalstay	Epidemiology	YES/NO/Cannot Infer	YES
23) Is the victim of the disease a child?	Child	Epidemiology	YES/NO/Cannot Infer	-
24) Is the victim of the disease an elderly person?	Elderly	Epidemiology	YES/NO/Cannot Infer	-
25) Is the victim of the disease a pregnant woman?	Pregnant	Epidemiology	YES/NO/Cannot Infer	YES
26) Has the victim of the disease been in quarantine?	Quarantine	Epidemiology	YES/NO/Cannot Infer	YES
27) Did the outbreak take place during a major sporting or cultural event?	Event	Epidemiology	YES/NO/Cannot Infer	YES
28) Did the outbreak take place after a natural disaster?	Disaster	Epidemiology	YES/NO/Cannot Infer	YES
29) When did the outbreak happen? (Relative to article completion time)	Tense	Epidemiology	Past/Now/Not Yet	-
30) Does the text contain information that can uniquely identify individual people? e.g. names, email, phone, and credit card numbers, addresses, user names.	Sensitive	Ethics	YES/NO	-

Table 1: Epidemiology questions given by experts in epidemiology.

Data Annotation Process

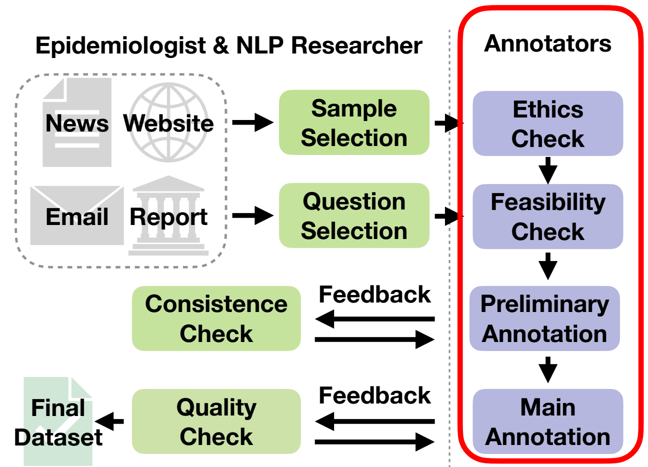
- ☑ Question Selection
- ☐ Samples Selection
  - Raw news alerts from ProMED-mail.
  - Collection of 36,788 raw alerts from Dec 2009 to Dec 2021.
  - Expert scoring system for samples.
  - Keyword prioritization.
- ☐ Annotation
- ☐ Consistency Check
- ☐ Quality Check
- ☐ Ethics Check

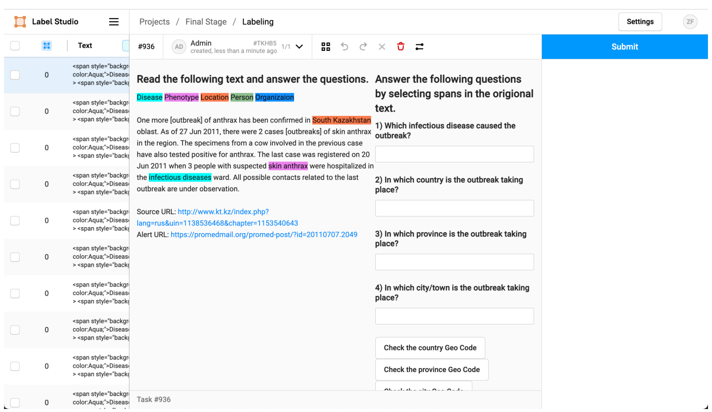




Data Annotation Process

- [x] Question Selection
- [x] Samples Selection
- [ ] Annotation
    - Utilized LabelStudio interface.
    - Professional annotation company employed.
    - 4 batches of annotation: 40, 710, 110, 660 samples.
    - Review and feedback after each stage.
- [ ] Consistency Check
- [ ] Quality Check
- [ ] Ethics Check

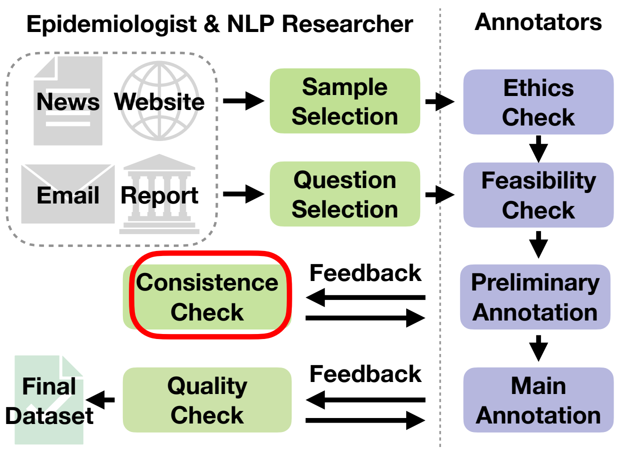




UNIVERSITY OF CAMBRIDGE

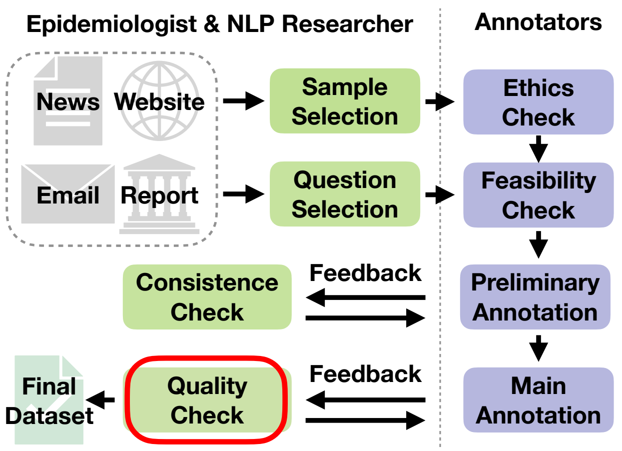
Data Annotation Process

- ☑ Question Selection
- ☑ Samples Selection
- ☑ Annotation
- ☐ Consistency Check
  - 5 annotators for the same 40 samples.
  - Manual review for consistency.
  - High consistency among annotators.
- ☐ Quality Check
- ☐ Ethics Check



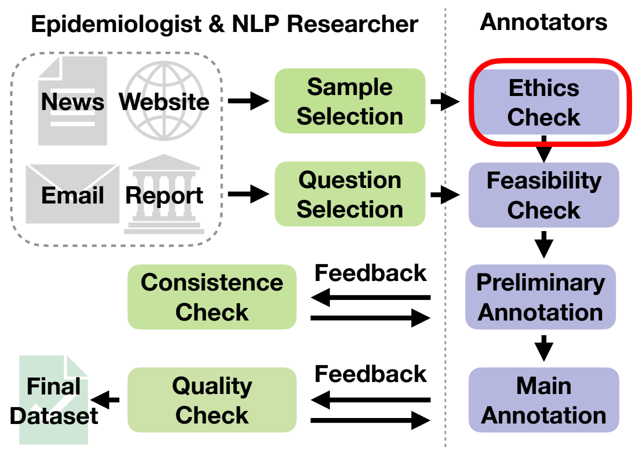
Data Annotation Process

- ☑ Question Selection
- ☑ Samples Selection
- ☑ Annotation
- ☑ Consistency Check
- ☐ Quality Check
  - Manual review of annotations to identify errors.
  - Feedback from experts to rectify misunderstandings.
  - Iterative feedback loop between annotators and experts for refined outcomes.
- ☐ Ethics Check



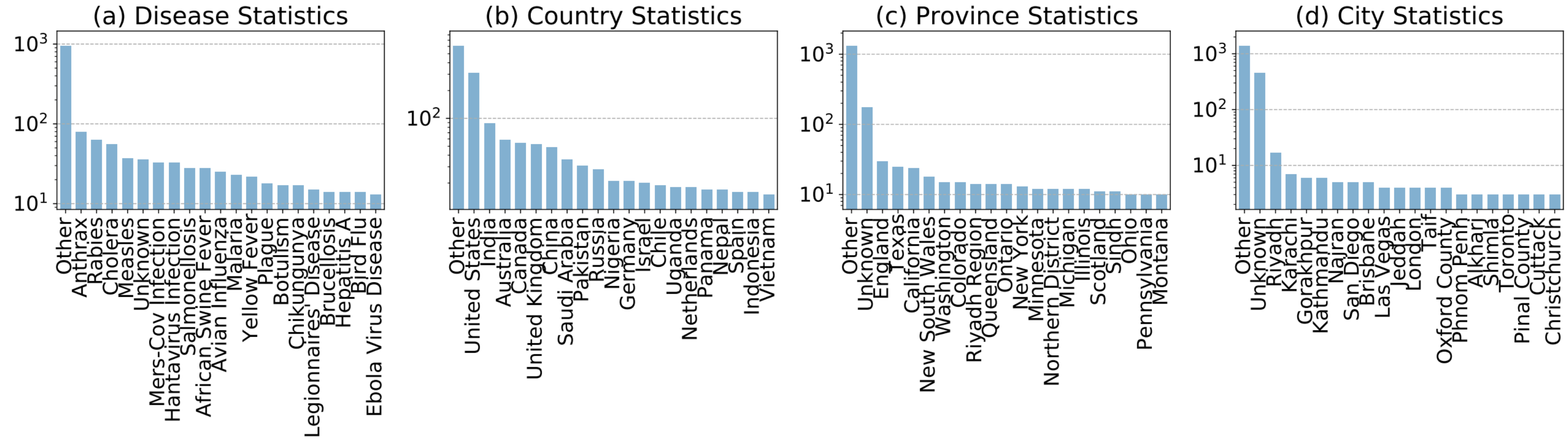
Data Annotation Process

- ☑ Question Selection
- ☑ Samples Selection
- ☑ Annotation
- ☑ Consistency Check
- ☑ Quality Check
- ☐ Ethics Check
    - Initiation of research ethics review.
    - Permission from the faculty's research ethics committee.
    - Annotation phase: Annotators assess samples for ethical rules compliance.
    - Removal of samples violating ethical standards.



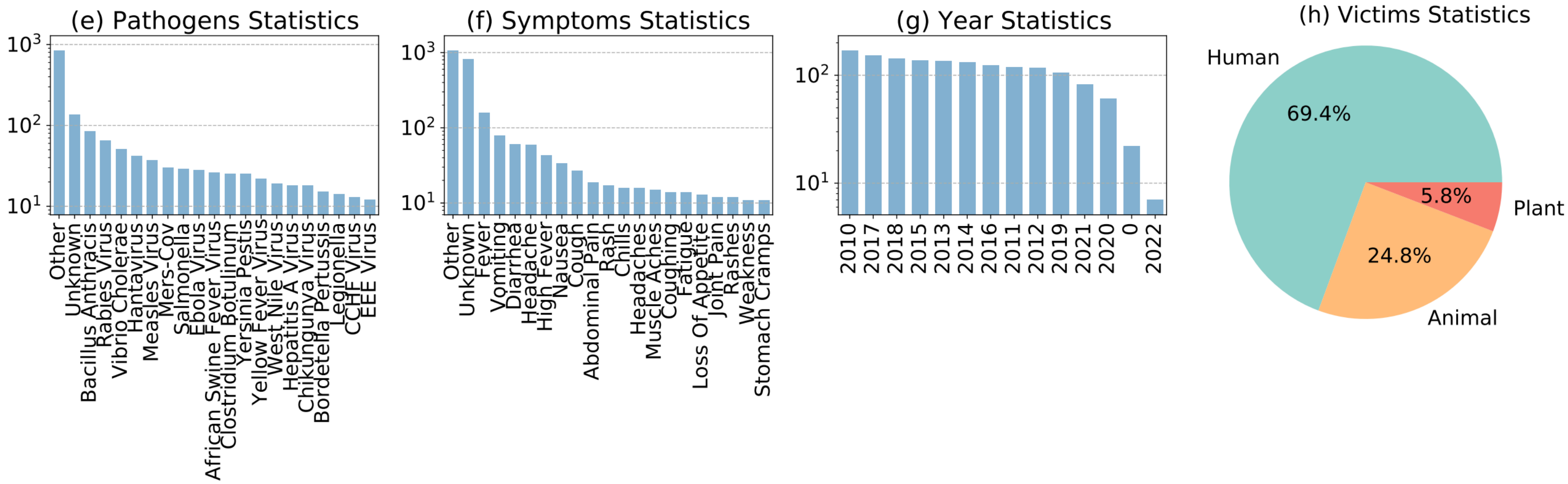
Statistics Overview

- **Disease Distribution:** The BAND dataset provides extensive coverage of a variety of popular infectious diseases such as Anthrax and Cholera.
- **Location Distribution:** The dataset represents a broad range of locations, encompassing diverse countries, provinces, and cities from around the world.



Statistics Overview

- **Pathogen Distribution:** The BAND dataset comprehensively captures mentions of numerous infectious pathogens, including bacteria, fungi, protozoa, and viruses.
- **Victim Distribution:** Focusing mainly on human and animal diseases, the dataset also incorporates data on plant diseases, expanding its application domains.
- **Symptoms Distribution:** The dataset encompasses a wide array of symptoms, making it ideal for training models to recognize various disease indicators.



Data Split

- Two splits: Rand Split & Stratified Split.
  - Rand Split: Random partitioning.
  - Stratified Split: Focus on sparse questions.

	Rand	**Stratified**
train	1,208	1,126
dev	150	149
test	150	233

Table 2: Data split.

Experiments

Tasks and Models

- NER Task (Named Entity Recognition)
    - Objective: Identify disease names, outbreak locations, pathogens, and symptoms.
    - Models: CRFBased, TokenBased, SpanBased, ChatGPT
- QA Task (Question Answering)
    - Objective: Answer questions using extractive and abstractive methods.
    - Models: T5, Bart, GPT2, GPTNEO, OPT, Galactica, BLOOM, ChatGPT
- EE Task (Event Extraction)
    - Objective: Identify and extract relevant information about disease outbreaks.
    - Models: T5, Bart, GPT2, GPTNEO, OPT, Galactica, BLOOM, ChatGPT

NER Task Results

- Supervised models (CRFBased, TokenBased, SpanBased) outperform zero-shot model (ChatGPT) on the BAND corpus.
- ChatGPT's lower performance may be due to:
  - Newness and specialization of our data.
  - ChatGPT's tendency to rephrase entities formally.
- Country, disease, and virus domains have the best NER performance. Provinces and cities see a decline in F1, highlighting a need for better few-shot/zero-shot capabilities.

	Random			Stratified		
Model	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
CRFBased	0.582	0.674	0.625	0.600	0.663	0.630
TokenBased	0.631	0.691	0.660	0.701	0.730	0.715
SpanBased	0.598	0.694	0.642	0.676	0.759	0.715
ChatGPT	0.326	0.353	0.339	0.424	0.318	0.363

Table 3: Named entity recognition results.

	Precision	**Recall**	**F1-score**
City	0.326	0.500	0.395
Country	0.710	0.760	0.734
Disease	0.583	0.758	0.659
Province	0.616	0.517	0.562
Virus	0.696	0.823	0.754

Table 4: NER results for each domain.

QA Task Results

- Decoder-only models (GPT2, Galactica) generally perform better than encoder-decoder models (T5, Bart).
- BLOOM outperforms other models, perhaps due to domain-specific training and controlled output style.
- ChatGPT underperforms due to:
  - Zero-shot nature on new dataset.
  - Inability to perform desired inference despite varied instructions.

Model	Rand	Stratified	Size	Mode
T5	0.674	0.591	220M (base)	Finetune
Bart	0.666	0.510	140M (base)	Finetune
GPT2	0.663	0.647	124M	Finetune
OPT	0.699	0.687	125M	Finetune
GPTNEO	0.695	0.695	125M	Finetune
Galactica	0.717	0.710	125M	Finetune
BLOOM	0.735	0.751	560M	Finetune
ChatGPT	0.497	0.413	-	Zero-Shot

Table 5: Question answering results.

QA Task Results

- Focus on questions with lower accuracy revealed ChatGPT's challenges:
  - Sometimes doesn't infer even when possible.
  - Occasional over-inference compared to human judgment.
  - For example, it identifies a city but fails to infer related country details.

Model		2) Country	3) Province	13) Borders	14) Intentful	15) Fromanimal	18) Animalworkers	23) Child	25) Pregnant	28) Disaster
T5	Accuracy	0.78	0.52	0.7	0.927	0.72	0.787	0.767	0.847	0.92
	Predict	Ukraine	–	Cannot Infer	–	NO	NO	NO	NO	–
	Gold Standard	Russia	Ohio	NO	NO	Cannot Infer	Cannot Infer	Cannot Infer	Cannot Infer	NO
	Error Count	2	1	30	1	14	10	7	9	1
BLOOM	Accuracy	0.794	0.442	0.833	0.97	0.781	0.82	0.824	0.807	0.961
	Predict	DR Congo	nan	Cannot Infer	NO	NO	Cannot Infer	NO	NO	NO
	Gold Standard	Democratic Republic of Congo	Helmand	YES	YES	Cannot Infer	NO	Cannot Infer	Cannot Infer	YES
	Error Count	2	1	25	6	25	22	14	21	8
ChatGPT	Accuracy	0.403	0.236	0.678	0.056	0.176	0.519	0.489	0.528	0.361
	Predict	Cannot Infer	Cannot Infer	No	Cannot Infer	Cannot Infer	Cannot Infer	Cannot Infer	Cannot Infer	Cannot Infer
	Gold Standard	United States	California	Cannot Infer	NO	NO	NO	NO	NO	NO
	Error Count	37	5	50	219	104	62	106	105	143

Table 6: Error analysis for accuracy of each question and top 1 error statistics for T5, BLOOM, and ChatGPT models.

EE Task Results

- Performance varies across categories with "province code" and "city code" often scoring low.
- BART excels in geocode predictions but struggles with other context attributes.
- ChatGPT excels in zero-shot setting, surpassing other decoder-only models.
- Encoder-decoder models (T5, BART) outperform decoder-only models in EE, hinting at decoder-only models' struggle with structured text generation.

Model	Overall F1	Individual F1									
		Disease	Country	Province	City	Country code	Province code	City code	Pathogen	Symptoms	Victim
T5	60.88	76.41	87.87	56.44	58.02	68.63	15.05	2.33	66.17	76.75	97.33
Bart	60.86	68.29	88.16	49.52	53.28	85.15	32.81	6.64	53.58	61.54	98.67
GPT2	45.34	63.92	78.43	38.24	34.48	38.69	0	0	49.82	41.86	93.65
OPT	48.27	66.89	82.51	49.54	43.09	32.24	0.62	0	52.99	54.68	94.31
GPTNEO	34.34	58.42	62.0	31.97	30.49	18.24	0	0	31.73	19.13	74.05
Galactica	49.33	62.35	78.95	50.33	45.97	46.05	0.65	0	56.72	57.61	91.47
Bloom	48.40	62.05	78.29	40.0	49.81	48.68	1.96	0	48.89	53.54	94.28
ChatGPT	47.71	56.16	79.15	47.29	46.15	51.61	7.32	4.4	28.04	45.03	83.5

Table 7: Event extraction results on random split.

Conclusions

- Introduction of Biomedical Alert News Dataset (BAND)
- BAND: 1,508 samples, 30 event & epidemiology related questions
- Tasks: NER, QA, EE
- Models: CRFBased, TokenBased, SpanBased, T5, Bart, GPT2, GPTNEO, OPT, Galactica, BLOOM, ChatGPT

Thanks!

https://github.com/fuzihaofzh/BAND/