Fine-Tuned Transformers Beat Large Language Models for Entity Recognition in Complex Eligibility Criteria for Clinical Trials

Supplement

Klaudia Kantor, Mikołaj Morzy

This document serves as a supplement to the paper "*Fine-Tuned Transformers Beat Large Language Models for Entity Recognition in Complex Eligibility Criteria for Clinical Trials*" presented at the 32nd International Conference on Information Systems Development (ISD 2024). It contains detailed results of experiments described in the paper.

Dataset characteristics

Label	Count	Label	Count
treatment	30972	gender	3661
chronic_disease	26212	pregnancy	2773
upper_bound	13967	age	2616
lower_bound	13633	allergy_name	1887
clinical_variable	13255	contraception_consent	1603
cancer	9344	bmi	287
technology_access	132	ethnicity	82

Table 1: Entity distribution the *CTP* data set

Label	Count
treatment	13541
chronic_disease	11362
clinical_variable	7205
cancer	5166
allergy_name	588

Table 2: Entity distribution in the evaluation data set

Comparison of models

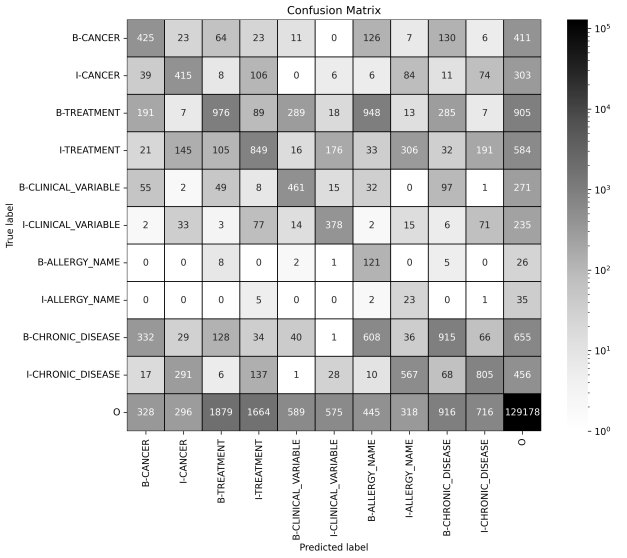
Table 4 presents the comparison of all models included in the study in a low-resource setting. Very interestingly, `CODER` is the only BERT-based model that outperforms `gpt-4-turbo` in this setting. Other BERT-based models cannot operate in the low-resource setting, and their pre-trained abilities are insufficient to perform any meaningful work. Table 5 presents the comparison of the same models trained in a high-resource setting. As can be clearly seen, only when sufficient annotated data for fine-tuning are available, Transformer-based models can correctly mark medical entities.

Confusion matrices for `gpt-4-turbo` and `CODER` models

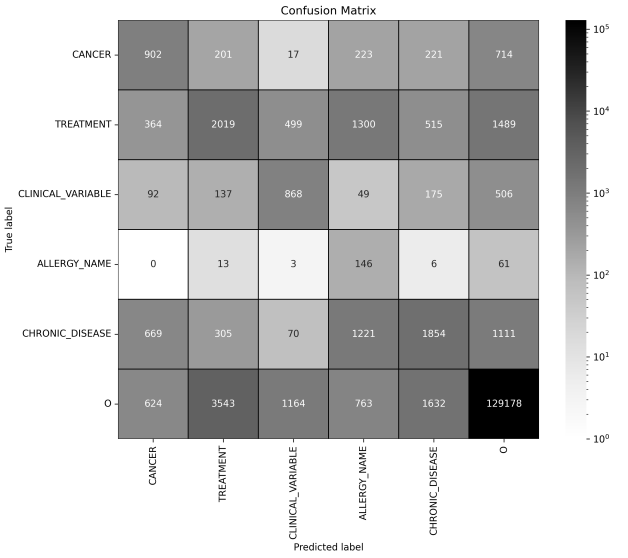
Figure 1 presents the comparison of the confusion matrices for the models. An important thing to note is that the nature of errors of the two models is different. While `gpt-4-turbo` has the tendency to confuse entities (in particular when using the BIO evaluation scheme), the `CODER` makes most of its mistakes with the `O` tag. In other words, `CODER` does not confuse entity spans, and its errors are mostly the omissions of entities.

Table 3: Examples of entity span annotations in the *CTP* data set

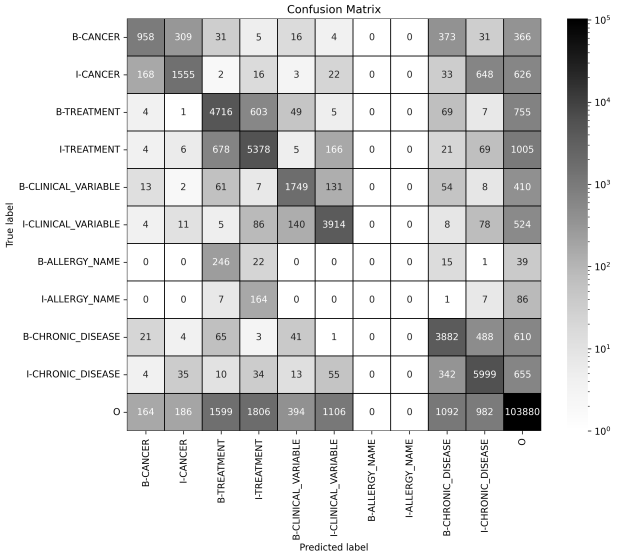
Entity class	Example
allergy_name	Dipeptidyl peptidase-4 (DDP-4) inhibitors, Linagliptin, propofol, glycerol
clinical_variable	total bilirubin, CIWA-Ar score, ALT, AST, Cockcroft-Gault formula
chronic_disease	liver dysfunction, glaucoma, kidney dysfunction, psychiatric disorder
treatment	flutamide, nilutamide, bicalutamide, Prior androgen deprivation therapy
cancer	melanoma, bone marrow plasmacytosis, Philadelphia (Ph)+ ALL



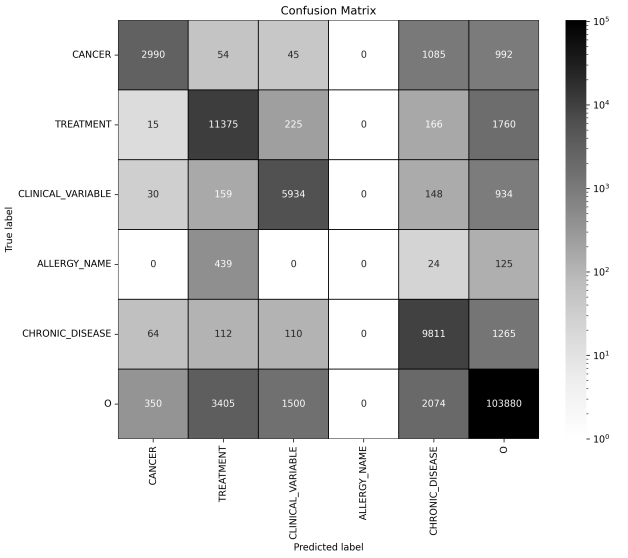
(a) `gpt-4-turbo` BIO



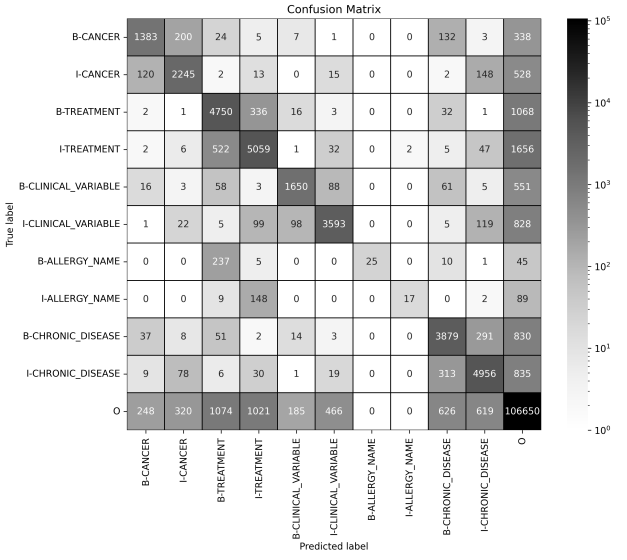
(b) `gpt-4-turbo` IO



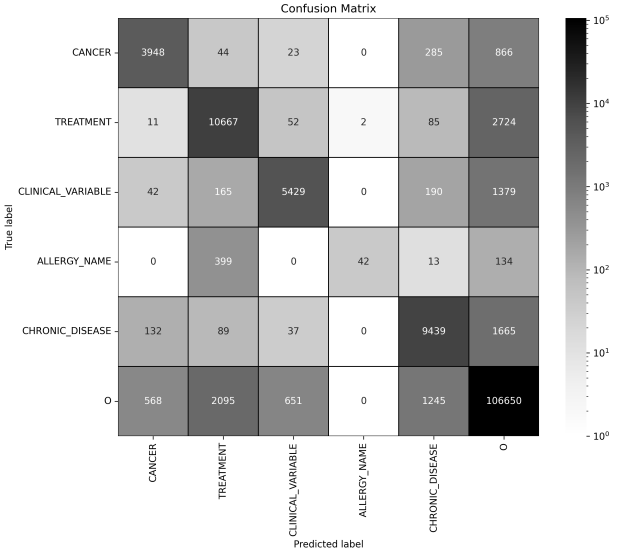
(c) `CODER-27` BIO



(d) `CODER-27` IO



(e) `CODER-100` BIO



(f) `CODER-100` IO

Figure 1: Confusion matrices for `gpt-4-turbo` and `CODER` models

Table 4: Comparison of models in low-resource settings (precision, recall, and F_1 score computed for BIO and IO labeling schemes

	BIO p	BIO r	BIO f	IO p	IO r	IO f
BERT	0,5948	0,0003	0,0006	0,6441	0,0003	0,0007
BioBERT	0,6501	0,0004	0,0007	0,7818	0,0004	0,0007
Biomedical NER	0,5418	0,2005	0,2907	0,6092	0,2226	0,3232
BlueBERT	0,6547	0,0006	0,0013	0,7079	0,0007	0,0014
ClinicalBERT	0,6429	0,0003	0,0007	0,6820	0,0004	0,0008
CODER	0,6722	0,7197	0,6896	0,7458	0,7953	0,7635
GPT4	0,3294	0,3408	0,3279	0,3556	0,3676	0,3550
PubMedBERT	0,4565	0,1401	0,2104	0,5310	0,1659	0,2492
SciBERT	0,6764	0,0210	0,0407	0,7427	0,0230	0,0447

Table 5: Comparison of models in high-resource settings (precision, recall, and F_1 score computed for BIO and IO labeling schemes

	BIO p	BIO r	BIO f	IO p	IO r	IO f
BERT	0,7852	0,0004	0,0008	0,7828	0,0004	0,0008
BioBERT	0,4426	0,0003	0,0007	0,7732	0,0003	0,0007
Biomedical NER	0,6645	0,0162	0,0317	0,7244	0,0177	0,0345
BlueBERT	0,3882	0,0004	0,0007	0,3933	0,0004	0,0007
ClinicalBERT	0,5777	0,0004	0,0007	0,7732	0,0004	0,0007
CODER	0,7780	0,7278	0,7454	0,8313	0,7798	0,7985
GPT4	0,3294	0,3408	0,3279	0,3556	0,3676	0,3550
PubMedBERT	0,7413	0,7468	0,7414	0,7958	0,8036	0,7983
SciBERT	0,7820	0,7515	0,7648	0,8324	0,8013	0,8153

Prompt template

Table 6: Comparison of `gpt-4-turbo` and `CODER` models on IO NER (p-precision, r-recall, f-F_1 score

	gpt-4-turbo			CODER-27			CODER-100			
	p	r	f	p	r	f	p	r	f	support
CLINICAL_VARIABLE	0,33	0,48	0,39	0,76	0,82	0,79	0,88	0,75	0,81	7205
CHRONIC_DISEASE	0,42	0,35	0,38	0,74	0,86	0,80	0,84	0,83	0,83	11362
TREATMENT	0,32	0,33	0,33	0,73	0,84	0,78	0,79	0,79	0,79	13541
CANCER	0,34	0,40	0,37	0,87	0,58	0,69	0,84	0,76	0,80	5166
ALLERGY_NAME	0,04	0,64	0,07	0,00	0,00	0,00	0,95	0,07	0,13	588
micro avg	0,30	0,37	0,33	0,75	0,80	0,77	0,83	0,78	0,80	37862
macro avg	0,29	0,44	0,31	0,62	0,62	0,61	0,86	0,64	0,67	37862
weighted avg	0,36	0,37	0,35	0,75	0,80	0,76	0,83	0,78	0,80	37862

```
Find examples of cancer in the following criterion.
Your response should be a list of comma separated values,  eg: `foo, bar, baz`
If no examples are found, type 'None'. Return only the entities found in the criterion.

criterion: Participant has received no prior radiotherapy or chemotherapy for rhabdomyosarcoma (excluding steroids) unless an
emergency situation requires local tumor treatment
entities: tumor

criterion: Aspartate aminotransferase (AST) (serum glutamic oxaloacetic transaminase [SGOT]) and alanine aminotransferase (ALT)
(serum glutamate pyruvate transaminase [SGPT]) =< 2.5 x ULN (or =< 5 x ULN if liver metastases [mets])
entities: liver metastases [mets]

criterion: Patients must have histologically confirmed, BRAF-mutant (V600E/K) melanoma (molecularly confirmed using validated,
commercially available assay performed in a Clinical Laboratory Improvement Act [CLIA]-approved laboratory) that is metastatic or
unresectable and for which standard curative measures do not exist or are no longer effective
entities: melanoma

criterion: loop recorder explanted within the past 12 months
entities: None

criterion: Other medical or psychiatric disorder placing the subject at undue risk for treatment complications
entities: None

criterion: All patients treated at doses > 120 mg per day must have medullary thyroid cancer (MTC), or a RET-altered solid tumor
per local assessment of tumor tissue and/or blood
entities:
```

Figure 2: Example of a prompt template.