Intent Detection and Entity Extraction from BioMedical Literature

Ankan Mullick[1]*, Mukur Gupta[2]*, Pawan Goyal[1]
[1]Computer Science and Engineering Department, IIT Kharagpur, India
[2]Computer Science Department, Columbia University, USA
ankanm@kgpian.iitkgp.ac.in, mukur.gupta@columbia.edu, pawang@cse.iitkgp.ac.in

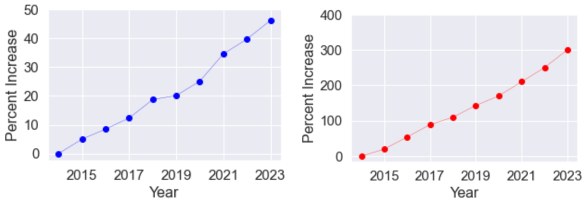
Abstract

Biomedical queries have become increasingly prevalent in web searches, reflecting the growing interest in accessing biomedical literature. Despite recent research on large-language models (LLMs) motivated by endeavors to attain generalized intelligence, their efficacy in replacing task and domain-specific natural language understanding approaches remains questionable. In this paper, we address this question by conducting a comprehensive empirical evaluation of intent detection and named entity recognition (NER) tasks from biomedical text. We show that Supervised Fine Tuned approaches are still relevant and more effective than general-purpose LLMs. Biomedical transformer models such as PubMedBERT can surpass ChatGPT on NER task with only 5 supervised examples.

1. Introduction

Research on large-language models has skyrocketed in the post-ChatGPT era. Researchers are now aiming for generalized intelligence by increasing model size (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022), expanding & rearranging pretraining data (Touvron et al., 2023a,b; Sarkar and Gupta, 2021) and incorporating human feedback (Ouyang et al., 2022; Dubois et al., 2023). It is shown that the adoption of GPT-4 (OpenAI, 2023) can potentially affect up to 80% of the U.S. workforce (Eloundou et al., 2023). These generalization reasoning demonstrations raise an important question for the research community - does this mark an end to the task and domain-specific natural language understanding approaches? While some research places LLMs as "General Purpose Technologies" (Eloundou et al., 2023; Zhang et al., 2023a) for solving a range of complicated tasks, we show that these models struggle to perform well on domain-specific complex tasks and specialized Supervised Fine-tuned (SFT) models are still needed to solve language understanding use-cases.

Over the past two decades, web searches have evolved dramatically transitioning from generic interfaces to more intent-specific and entity-aware systems capable of immediately displaying diverse multi-modal responses. Particularly, biomedical inquiries, spanning topics such as medical treatment, medical diagnosis, disease, etc. have seen a surge in popularity across search engines. Fig. 1 shows the increase in the percentage of Biomedical queries on Bing search and Google trends[1].



(a) Google yearly Trend (b) Bing Query yearly

Figure 1: Biomedical query search Statistics

As large volumes of biomedical data continue to be generated every second on various online platforms the role of information retrieval systems in processing domain-specific texts becomes increasingly important. However, handling biomedical text data presents unique challenges, as the medical queries on search engines and online medical forums are often incomplete, do not follow a specific structure, and contain hard-to-interpret context-specific medical terminologies, as shown in Table 1. While recent research is centered around the development of general-purpose LLMs, that are shown to exhibit exceptional Common Sense Reasoning capabilities (Touvron et al., 2023b), we show that these models face challenges in transferring their performance to intricate biomedical domains. To this end, we focus on two crucial natural language understanding tasks of intent detection and named entity recognition from biomedical text.

For the past two decades, different directions of intent detection and corresponding entity extraction have been explored. (Sun et al., 2016; Wang et al., 2020; Mu et al., 2017b,a) demonstrate intent detection in the form of out-of-domain data detection. Other research works explore methods like few shot (Xia et al., 2021), zero-shot (Xia et al., 2018), and clustering frameworks (Mullick et al., 2022b). (Yani et al., 2022; Zhao et al., 2021; Fetahu et al.,

*Authors contributed equally

[1] Google trends data of last 10 years on five topics (Health, Medical Treatment, Medical diagnosis, Disease, Pharmaceutical drug) was gathered from Google Trends (https://trends.google.com/trends/)

Biomedical Text	Intent
Pharmacokinetic properties of abacavir were not altered by the addition of either lamivudine or zidovudine or the combination of lamivudine and zidovudine .	Drug
Canavan disease, or spongy degeneration of the brain, is a severe leukodystrophy caused by deficiency of aspartoacylase (ASPA).	Disease

Table 1: Intent & corresponding Entity (highlighted) examples from DDI and NCBI Disease datasets.

2022) explore entity recognition task in various settings. In the medical domain, Zhou et al. (2021) focuses on smart healthcare and (Galea et al., 2018; Giorgi and Bader, 2019; Lee et al., 2019) inspect transformer based models for biomedical literature. Mullick et al. (2023, 2022a); Mullick (2023b,a) aims at intent detection and entity extraction and Zhang et al. (2017) explore medical query intents by applying graph-based frameworks. (Mullick et al., 2024; Guha et al., 2021) work on domain specific entity and corresponding relation extraction. (Mullick et al., 2017b, 2016, 2018a,b, 2019, 2017a) aim at opinion-fact entity extraction.

There is no unified and exhaustive comparison of existing approaches with the recent LLMs for intent detection and entity extraction tasks across various datasets in biomedical literature. Our work differs from the prior research in two ways: we present a thorough empirical evaluation of the intent detection on three datasets and corresponding named entity extraction (NER) approaches on 27 unique entities covered in 5 biomedical datasets spanning across domains like drugs, diseases, chemicals, genetics and, human anatomy. We evaluate various supervised approaches (transformer-based, handcrafted features, etc.) and benchmark them against two widely used large language models in the biomedical domain. Our experiments reveal that the biomedical transformer-based PubMedBERT model outperforms few-shot prompted ChatGPT (Turbo 3.5) on 4 biomedical NER benchmarks with just 5 supervised examples. We make our code publicly available.[2]

2. Datasets

We show our comparative study on a variety of datasets, which are widely used as benchmarks in the biomedical domain. We use five different Named Entity Recognition datasets: JNLPBA (Collier and Kim, 2004), DDI (combining DDI-Drugbank and DDI-Medline) (Segura-Bedmar et al., 2013), BC5CDR (Smith et al., 2008), NCBI-Disease (Li et al., 2016) and AnatEM (Ohta et al., 2012). Dataset statistics including the entity types, count, and train-test splits are outlined in Table 2. We use the pre-defined train-test divisions from the respective manuscripts.

Dataset	Entity Type	# Entities	#Train	#Test
JNLPBA	Gene & Protein	5	2000	404
DDI	Drug	4	714	112
BC5CDR	Chem & Diesease	2	1000	500
NCBI-Disease	Disease	4	693	100
AnatEM	Anatomy	12	300	200

Table 2: Statistics of the NER datasets. We use the pre-defined train-test split as mentioned in the papers.

Along with the two popular intent detection datasets - CMID (Chen et al., 2020) and KUAKE-QIC (part of the CBLUE (Zhang et al., 2021) benchmark), we combine the three of the above five NER datasets (JNLPBA, DDI, and NCBI-Disease) with respective intent labels (DDI for drugs, NCBI-Disease for disease and JNLPBA for Genetics) for intent classification task - termed as “Intent-Merged” dataset. Dataset statistics are summarized in Table 3.

CMID and KUAKE-QIC datasets, which are originally in Chinese, are translated to English using Google Translation API. For translation validation, a random sample of 400 translated (to English) examples of each dataset are validated manually by two Chinese experts (ALA Language Center Company) with HSK Level-3 proficiency. The human-validation shows 91.75% and 97.0% translation accuracy for CMID and KUAKE-QIC respectively. Hence, we use the translated English data along with their pre-defined intent labels for our experiments. The inter-annotator agreement is 0.89.

Dataset	#Train	#Test Size	#Intents
CMID	9558	2696	4
KUAKE-QIC	6931	1955	11
Intent-Merged	3905	909	3

Table 3: Statistics of Intent Detection datasets.

3. Experimental Settings

3.1. Intent Detection

Intent detection is a multi-class classification task where we evaluate the accuracy of instruction-tuned ChatGPT (gpt-3.5-turbo-instruct) against various SFT models on three English datasets: CMID, KUAKE-QIC, and Intent-Merged.

1. Large Language Models: To ensure consistency with prior works, we employ a k-shot prompt design, wherein k examples per class from the training set are used in the prompt. Given the

[2] https://github.com/bioNLU-coling2024/biomed-NER-intent_detection

larger text sizes of the Intent-Merged dataset and the limited context window of LLMs, we use $k = 1$ for all datasets. We note no significant performance improvement with increasing k for CMID and KUAKE-QIC datasets. Further details on the prompt template are included in the GitHub repository.

2. Supervised Fined-Tuned Models: For SFT, we finetune - BERT (bert-base-uncased) (Devlin et al., 2018), RoBERTa (roberta-base) (Liu et al., 2019), PubMedBERT (Gu et al., 2021), FastText (Chen et al., 2020) and TextCNN (Kim, 2014). The empirical evaluation is shown in Table 4.

Model	CMID	KUAKE-QIC	Intent-Merged	Mean
BERT	72.26	75.91	96.37	81.51
RoBERTa	**72.88**	**78.16**	**99.11**	**83.38**
PubMedBERT	72.70	76.88	97.90	82.49
Llama-2	51.11	42.50	39.54	44.38
ChatGPT	42.36	44.04	64.44	50.28
Fasttext	68.43	72.48	96.80	79.24
TextCNN	70.69	75.19	96.15	80.68

Table 4: Accuracy (in %) of intent classification tasks on three datasets.

All the SFT approaches consistently outperform the instruction-tuned ChatGPT. The poor performance of LLMs on the Intent-Merged dataset, which is quite easy for all the SFT approaches, reflects their deficiency in domain-specific knowledge within their general-purpose pretraining datasets. This also shows that models like FastText can outperform ChatGPT, given domain-specific finetuning. We note that transformer architectures give better performance on the translated corpus compared to FastText and TextCNN, which are shown to work well on Chinese data (Chen et al., 2020). RoBERTa gives the highest accuracy across overall mean and individual datasets.

3.2. Named Entity Recognition

For NER, we apply a strict match between the predicted entity class and the entity word boundaries and report strict F1-score (as in CoNLL shared task (Tjong Kim Sang and De Meulder, 2003)). We run all models 5 times with different random initialization and report micro-average F1-score along with standard deviations. We also report the overall mean for each approach. For a fair comparison, a maximum sequence length of 512 tokens was used for all models, hence the texts larger the token length were further broken into multiple texts.

3.2.1. Supervised Fine-Tuned Models:

We thoroughly examine five different settings on the five biomedical datasets.

Setting A: Fine-tuned BERT and RoBERTa models are used (pre-trained on general English corpus) without domain pretraining.

Setting B: Transformer systems with continued pretraining on biomedical text. We fine-tune BioBERT, PubMedBERT, BioMed RoBERTa, and ClinicalBERT.

Setting C: LSTM and Convolutional Neural Networks (with/without CRF) are used to generate the word embeddings and softmax classifiers for tag prediction.

Setting D: Hand-crafted word level features with ML classifier: (i) POS tag (ii) shallow parsing features like chunk tag (iii) orthographic boolean features like all capital, is alphanumeric, etc. (iv) n-gram features, etc. We use the GENIA tagger[3] for POS and Chunk tag extraction. We apply XGBoost and a multi-label logistic regression model for NER tag prediction.

Setting E: We use state-of-the-art NER model BINDER (Zhang et al., 2023b) along with domain-specific (PubMedBERT and BioBERT) and RoBERTa encoders.

3.2.2. Large Language Models:

We use instruction-tuned ChatGPT (gpt-3.5-turbo-instruct).

Setting F: We modify (Wang et al., 2023) for reconditioning NER as a Tag generation problem. In addition to the prompt design proposed by (Wang et al., 2023), following (Zhang et al., 2023b) we also add a short description for the entity. Each prompt infers only a single entity tag. Hence, each text instance is passed multiple times for tagging all the entity types. We provide two examples from the train set in each prompt. To motivate both high recall and prevent hallucination in Entity identification, we specifically pick examples with the median number of entity tags in the training dataset.

The evaluation outcomes are shown in Table 5. We find:

a) SFTs outperform LLMs: We observe that all SFT approaches surpass ChatGPT by a big margin. Further, from Table 6, it's evident that PubMedBERT can easily outperform ChatGPT on most benchmarks with just five supervised examples.

b) Transformer SFT Models: i) PubMedBERT learns good embedding vectors due to the largest pretraining corpus. BINDER combined with PubMedBERT gives the best F1 score as it is able to leverage high-quality embeddings along with entity descriptions which pushes the similar entity tokens closer in the embedding space with a contrastive loss objective. (ii) LSTM/CNN-based neural embedding and traditional ML-based models - XGBoost and Logistic Regression perform poorly because they fail to capture contexts and do not leverage domain-specific pretraining.

[3] http://www.nactem.ac.uk/GENIA/tagger/

Model/Dataset	DDI	JNLPBA	BC5CDR	NCBI Disease	AnatEM	Mean
BERT (A)	83.94 ± 0.17	72.60 ± 0.13	87.13 ± 0.24	77.44 ± 0.75	78.66 ± 0.35	79.95
RoBERTa (A)	87.13 ± 0.44	74.91 ± 0.11	89.50 ± 0.09	81.67 ± 0.51	81.90 ± 0.60	83.06
BioBERT (B)	88.06 ± 0.08	74.02 ± 0.40	90.19 ± 0.13	81.91 ± 0.80	83.43 ± 0.36	83.52
PubMedBERT (B)	88.84 ± 0.12	75.15 ± 0.06	90.77 ± 0.08	82.34 ± 0.16	84.21 ± 0.21	84.26
BioMed RoBERTa (B)	88.76 ± 0.31	75.14 ± 0.25	90.24 ± 0.18	82.13 ± 0.92	82.70 ± 0.17	83.80
Clinical BERT (B)	83.79 ± 0.22	72.54 ± 0.07	87.90 ± 0.17	76.34 ± 0.64	73.47 ± 0.53	78.80
LSTM (C)	73.00 ± 0.01	67.00 ± 0.01	79.01 ± 0.01	70.00 ± 0.01	74.01 ± 0.01	72.60
LSTM + CRF (C)	74.75 ± 0.04	70.67 ± 0.01	80.47 ± 0.02	73.13 ± 0.02	77.39 ± 0.05	75.23
CNN (C)	73.07 ± 0.01	68.04 ± 0.01	80.00 ± 0.01	67.09 ± 0.01	72.08 ± 0.01	72.06
CNN + CRF (C)	73.29 ± 0.08	70.84 ± 0.11	81.27 ± 0.14	73.59 ± 0.09	75.15 ± 0.13	74.83
Logistic Regression (D)	78.63 ± 0.01	57.03 ± 0.01	78.20 ± 0.01	56.36 ± 0.01	65.48 ± 0.01	67.14
XGBoost (D)	73.55 ± 0.01	53.06 ± 0.01	67.86 ± 0.01	52.62 ± 0.01	59.91 ± 0.01	61.40
BINDER-BioBERT (E)	89.01 ± 0.01	76.63 ± 0.19	91.59 ± 0.09	**85.47** ± 0.36	86.71 ± 0.25	85.88
BINDER-PubMedBERT (E)	**89.12** ± 0.01	77.01 ± 0.01	**91.88** ± 0.01	85.25 ± 0.02	**86.95** ± 0.02	**86.04**
BINDER-RoBERTa (E)	87.98 ± 0.01	**77.08** ± 0.01	90.48 ± 0.03	84.62 ± 0.06	83.91 ± 0.05	84.81
ChatGPT (F)	42.94 ± 3.10	24.5 ± 1.89	44.68 ± 2.78	19.65 ± 1.21	2.92 ± 0.07	26.94

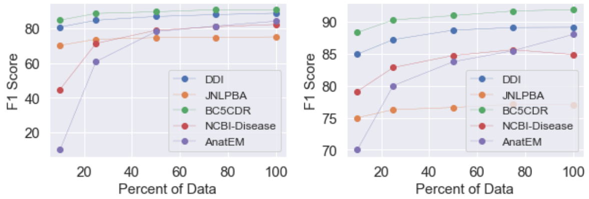
Table 5: Experiment results (Macro average F1-Scores and corresponding standard deviations) on different NER systems trained/finetuned and tested on 5 biomedical Datasets.

c) Feature-based SFT Models: (i) ML-based model performs better than CNN/LSTM embedding systems on the DDI dataset, implying that it might be possible to beat the performances on other datasets if the right feature set is selected, which is usually an expensive process. (ii) The range of performance (best F1 - worst F1) for NCBI-Diease corpus is highest, showing that there is a big difference between the selected feature set and the features captured by the neural models. (iii) The addition of a CRF prediction layer on CNN/LSTM improves the performance significantly.

d) Dataset Quality: In most of the cases, low F1 is observed on the entities having fewer examples in the training set. For example, entities "CompositeMention" and "Disease Class" show poor performance due to less number of samples in training data. We note that the tag generation problem is difficult for instruction-tuned LLMs. We also experiment with Llama-2 (7b) model[4] and observe that vanilla Llama-2-7b does not achieve good results as it was unable to follow the specified output structure and most of times, ended up hallucinating text. So, we omit vanilla Llama-2 results and will explore further in future.

3.2.3. Experimental Setup:

We experiment on Tesla T4 16GB GPU, 6 Gbps clock cycle and GDDR5 memory. All experiments (entity extraction and intent detection) took ∼60 minutes for training. We fine-tune the models for a maximum of 20 epochs with a learning rate of 5e-5 with AdamW optimizer and 10% warm-up steps. The batch size is 16. Additional details are included in the GitHub Repository.



(a) PubMedBERT (b) Binder-PubMedBERT

Figure 2: Ablation: Varying Training Size

4. Ablations

We study the relationship of SFT models with domain-specific finetuning data:

Varying Training Data Size: We vary the size of training data (10%, 25%, 50%, 75% and 100%), while keeping the test set constant and show the performances of PubMedBERT, Binder (PubMedBERT) models in Figure 2. We observe that, unlike raw PubMedBERT, Binder (PubMedBERT) attains a high performance with only 10% of training data. Transformer-based models can learn with very little training data and performance does not decrease much even with 25% training data. Due to domain pre-training, PubMedBERT learns with much fewer samples and saturates faster. This quick-learning behavior seems to be originating from transfer learning. However, LSTM and CNN models suffer from poor performances in low-data settings due to no pretraining (details in GitHub).

# Shots	DDI	JNLPBA	BC5CDR	NCBI	AnatEM
5	2.8 / 60.05	2.0 / 39.09	1.1 / 64.53	3.2 / 27.56	2.6 / 1.5
10	5.8 / 65.54	5.6 /50.56	2.1 /69.23	7.05 /34.32	6.1 / 3.5
30	45.49 /73.71	61.07 / 58.01	54.07 / 77.97	60.63 / 48.16	21.74 / 10.9
50	81.39 / 76.85	71.04 / 62.23	83.3 / 82.78	75.83 / 49.51	40.69 / 30.96
100	83.56 / 80.94	74.24 / 68.02	88.58 / 85.76	82.38 / 72.78	73.34 / 52.73

Table 6: BINDER-PubMedBERT / PubMedBERT F1-score in K-shot setting

[4] https://ai.meta.com/llama/

Few Shot: In Table 6, we show the performances (F1 Score) of the Binder (PubMedBERT) and PubMedBERT models on a few shot settings with different numbers of training samples. PubMedBERT embeddings perform better in very low-resource setups (5, 10 shots). However, when training examples increase further (30 shots onwards), BINDER (PubMedBERT) outperforms PubMedBERT because of the Bi-Encoder architecture trained on contrastive learning objectives.

Error Type	Entity Text	Label	Prediction
Boundary	3-[(...)ethynyl] pyridine	B-D,I-D	B-D,B-D
Entity type	heparinase III	B-D,I-D	B-G,I-G
Entity Miss	Hyaluronan lyase	B-D,I-D	O,O

Table 7: Errors by BINDER-PubMedBERT on entity "drug_n" of DDI dataset. Following abbreviations are used - B-D: B-DRUG_N, I-D: I-DRUG_N, B-G:B-Group, I-G: I-GROUP

5. Error Analysis

We do a detailed analysis on errors as following:

A) Some errors are due to model failure like RoBERTa's failure to classify 52% of the "other" intents from the KUAKE-QIC dataset. For example, a query such as "I have a *cyst* in the *corner of my right eye* and it grows bigger and bigger." is classified wrongly as "diagnosis" intent but it is of "other" category.

B) Three types of errors are observed for entity extraction (examples from the DDI dataset are shown in Table 7).

C) Some models fail to identify the entity "drug_n" which represents new or unapproved drugs so a periodic model update is required.

D) Relaxing entity-type error by considering exact F1-score instead of strict F1, we observe an uplift of 4.57% in mean F1.

6. Conclusion

The biomedical sector has matured significantly in the past few years. We show instead of relying on general-purpose LLMs, it is important to design an intent detection and entity extraction task for processing domain-specific texts. In this work, we show that fine-tuned RoBERTa and BINDER (PubMedBERT) can work efficiently to detect intents and extract named entities across various benchmark datasets in biomedical literature. In the future, we aim to extract intent and entity jointly as a relation tuple and inspect the performances of various cross-domain scenarios.

Limitations

Our dataset needs to be scaled up in terms of different languages, sizes, and intent labels which we aim to do in the near future. The approach needs to be updated as a single model for jointly extracting intents and entities for multilingual scenarios which we aim to do as a part of future work.

Ethical Concerns

We propose to release the algorithmic details and work on public datasets that neither reveal any personal sensitive information nor any toxic statement. So there are no ethical concerns in this work.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Nan Chen, Xiangdong Su, Tongyang Liu, Qizhi Hao, and Ming Wei. 2020. A benchmark dataset and case study for chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, 20.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy

Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790, Seattle, United States. Association for Computational Linguistics.

Dieter Galea, Ivan Laponogov, and Kirill Veselkov. 2018. Exploiting and assessing multi-source data for supervised biomedical named entity recognition. *Bioinformatics*, 34(14):2474–2482.

John M Giorgi and Gary D Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Souradip Guha, Ankan Mullick, Jatin Agrawal, Swetarekha Ram, Samir Ghui, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2021. Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature. *Computational Materials Science (Comput. Mater. Sci.)*, 192:110325.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016:baw068.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xin Mu, Kai Ming Ting, and Zhi-Hua Zhou. 2017a. Classification under streaming emerging new classes: A solution using completely-random trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1605–1618.

Xin Mu, Feida Zhu, Juan Du, Ee-Peng Lim, and Zhi-Hua Zhou. 2017b. Streaming classification with emerging new class by class matrix sketching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Ankan Mullick. 2023a. Exploring multilingual intent dynamics and applications. *IJCAI Doctoral Consortium*.

Ankan Mullick. 2023b. Novel intent detection and active learning based classification (student abstract). *arXiv e-prints*, pages arXiv–2304.

Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.

Ankan Mullick, Surjodoy Ghosh D, Shivam Maheswari, Srotaswini Sahoo, Suman Kalyan Maity, and Pawan Goyal. 2018a. Identifying opinion and fact subcategories from the social web. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*, pages 145–149.

Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 40–49.

Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2017a. Extracting social lists from twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 391–394.

Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2018b. Harnessing twitter for answering opinion list queries. *IEEE Transactions on Computational Social Systems*, 5(4):1083–1095.

Ankan Mullick, Shivam Maheshwari, Pawan Goyal, and Niloy Ganguly. 2017b. A generic opinion-fact classifier with application in understanding opinionatedness in various news section. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 827–828.

Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022a. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.

Ankan Mullick, Sourav Pal, Projjal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh, and Tanmoy Dam. 2019. D-fj: Deep neural network based factuality judgment. *Technology*, 50:173.

Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022b. A framework to generate high-quality datapoints for multiple novel intent detection. *arXiv preprint arXiv:2205.02005*.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 27–36, Jeju Island, Korea. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Dipankar Sarkar and Mukur Gupta. 2021. Tso: Curriculum generation using continuous optimization. *CoRR*.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Larry Smith, Lorraine K Tanabe, Rie Johnson Nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A Struble, Richard J Povinelli, Andreas Vlachos, William A Baumgartner, Jr, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W John Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome Biol.*, 9 Suppl 2(S2):S2.

Yu Sun, Ke Tang, Leandro L Minku, Shuo Wang, and Xin Yao. 2016. Online ensemble learning of data streams with gradually evolved classes. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1532–1545.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Min Wang, Ke Fu, Fan Min, and Xiuyi Jia. 2020. Active learning through label error statistical methods. *Knowledge-Based Systems*, 189:105140.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. *arXiv preprint arXiv:2104.11882*.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2018. Zero-shot user intent detection via capsule neural networks. *arXiv preprint arXiv:1809.00385*.

Mohammad Yani, Adila Alfa Krisnadhi, and Indra Budi. 2022. A better entity detection of question for knowledge graph question answering through extracting position-based patterns. *Journal of Big Data*, 9(1):1–26.

Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, Gyeong-Moon Park, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. 2023a. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era.

Chenwei Zhang, Nan Du, Wei Fan, Yaliang Li, Chun-Ta Lu, and S Yu Philip. 2017. Bringing semantic structures to user intent detection in online medical queries. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1019–1026. IEEE.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023b. Optimizing bi-encoder for named entity recognition via contrastive learning.

Lingyun Zhao, Lin Li, Xinhao Zheng, and Jianwei Zhang. 2021. A bert based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238. IEEE.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2021. Natural language processing for smart healthcare. *arXiv preprint arXiv:2110.15803*.