MiDRED: An Annotated Corpus for Microbiome Knowledge Base Construction

William Hogan[1], **Andrew Bartko**[2,3,4], **Jingbo Shang**[1], **Chun-Nan Hsu**[5]
[1]Department of Computer Science & Engineering,
[2]Center for Microbiome Innovation, [3]Department of Bioengineering,
[4]Department of Pediatrics, [5]Department of Neurosciences
University of California, San Diego, La Jolla, CA 92093
whogan@ucsd.edu

Abstract

The interplay between microbiota and diseases has emerged as a significant area of research facilitated by the proliferation of cost-effective and precise sequencing technologies. To keep track of the many findings, domain experts manually review publications to extract reported microbe-disease associations and compile them into knowledge bases. However, manual curation efforts struggle to keep up with the pace of publications. Relation extraction has demonstrated remarkable success in other domains, yet the availability of datasets supporting such methods within the domain of microbiome research remains limited. To bridge this gap, we introduce the Microbe-Disease Relation Extraction Dataset (MiDRED); a human-annotated dataset containing 3,116 annotations of fine-grained relationships between microbes and diseases. We hope this dataset will help address the scarcity of data in this crucial domain and facilitate the development of advanced text-mining solutions to automate the creation and maintenance of microbiome knowledge bases.

1 Introduction

Microbiota play a pivotal role in human health in diverse environments such as the gut, skin, and oral cavity, influencing various physiological processes and disease mechanisms (Cho and Blaser, 2012; Lynch and Pedersen, 2016; Singh et al., 2017). The significance of microbiome research is underscored by its immense potential to unlock new understandings and treatments for various health conditions (Stefano et al., 2022; Yu et al., 2022; Kustrimovic et al., 2023). For example, perturbations in gut microbiota composition, exemplified by fluctuations in Bacteroidetes and Firmicutes populations, have been linked to obesity and type 2 diabetes, respectively, providing valuable insights into the pathophysiology of these conditions (Baek et al., 2023; Kusnadi et al., 2023). The growth of microbiome research introduces significant challenges in knowledge consolidation and utilization (Badal et al., 2019; Huang et al., 2022). Current efforts often involve domain experts spending countless hours manually curating experimentally validated associations between diverse microbiota and diseases to form knowledge bases (KBs) (Li et al., 2021; Dai et al., 2021; Qi et al., 2022; Zhang et al., 2022). These KBs are invaluable for researchers and practitioners, providing a consolidated view of current findings, yet their maintenance is becoming unsustainable due to the rapid pace of publication. Advanced text-mining methods designed to extract knowledge from biomedical texts are a well-established area of research (Wei et al., 2016; Zhang et al., 2018; Hogan et al., 2021; Xu et al., 2022; Li, 2022; Lai et al., 2023; Liu et al., 2023). Methods often leverage human-annotated data to train and validate a model's performance; however, robust datasets annotating microbe-disease associations are lacking.

To address these challenges, we introduce MiDRED; a comprehensive text-mining dataset designed to automate the construction and maintenance of microbiome KBs. MiDRED consists of 3,116 annotated relationships between microbe-disease pairs extracted from 1,655 scholarly articles. We specifically craft relation classes to align with classes used in major microbe-disease KBs to ensure MiDRED's compatibility with existing databases. Importantly, MiDRED annotates negative instances (e.g., a "no relation" class) to mitigate positive bias from trained models (Zhang et al., 2017). MiDRED also includes span-level annotations of entities, which are crucial for training in Named Entity Recognition (NER) and Named Entity Normalization (NEN) tasks. See Table 2 for statistics on the complete dataset. We conducted experiments on MiDRED using a variety of generative and discriminative large language models to obtain robust baselines to serve as a foundation for future research. We openly release the MiDRED

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 398–408

dataset on Hugging Face.[1]

2 Related Work

MiDRED is designed as a text-mining dataset and draws inspiration from numerous biomedical (Khettari et al., 2023; Bossy et al., 2019; Luo et al., 2022; Li et al., 2016; Taboureau et al., 2010; Janssens et al., 2018) and general domain text-mining datasets (Zhang et al., 2017; Stoica et al., 2021; Yao et al., 2019). Text-mining datasets typically consist of manually annotated texts which can be used to train and evaluate automated NER, NEN, and relation extraction algorithms (Zhang et al., 2017; Yao et al., 2019). Works such as Herrero-Zazo et al. (2013), Luo et al. (2022), González et al. (2019) are similar in task but differ either in entity types, association types, or both.

The Human Microbe-disease Dataset (HMDAD) (Ma et al., 2016) is a database of associations between human microbes and diseases. However, the dataset does not provide span-level information denoting entity pairs which limits the dataset's use in training NER and NEN algorithms. Microbes in HMDAD were primarily curated at the genus level due to the sequencing technologies available when the dataset was annotated. MiDRED benefits from advancements in sequencing technologies, allowing for a majority (95.4%) of microbial concepts to be annotated at the species level. Furthermore, HMDAD relation annotations are done at the article-level. Article-level annotation is commonly used in microbiome knowledge bases (Janssens et al., 2018; Cheng et al., 2019; Li et al., 2021; Skoufos et al., 2020) and fails to denote the location of textual evidence supporting an association, making it challenging to train automated text-mining tools. Lastly, MiDRED differs from HMDAD in that it does not limit its annotations based on host type, leading to more diverse associations.

The Species-species Interaction (SSI) dataset (Khettari et al., 2023) is a dataset that annotates binary associations between species of microbes. SSI does not provide human-annotated entities and relies on automated methods for NER. MiDRED differs from SSI in entity types and the number of relation classes—in MiDRED, we annotate four relation classes (see Section 3.2 for more details), moving beyond binary associations. Bacteria Biotope (BB 2019) (Bossy et al., 2019) is an NER/RE dataset featuring microbes, diseases, habitats, and locations. BB 2019 seeks to mine associations of microbes and environments (habitats) to better understand how microbes interact within various environments. MiDRED, in contrast, focuses on how microbes relate to diseases more generally and offers a large number of annotated entities and relations.

3 Methods

3.1 Data Collection and Entity Normalization

We collect an initial set of abstracts from PubMed (Sayers et al., 2020) using the PubTator tool (Wei). To ensure the subset of abstracts are relevant to microbiome studies, we prioritize PMIDs found within the Disbiome database (Janssens et al., 2018). From this subset, we randomly select abstracts and annotate microbes and diseases. Microbial entities were normalized to the List of Prokaryotic Names with Standing in Nomenclature (LPSN) ontology (Parte et al., 2020). Disease entities were normalized to the Comparative Toxicogenomics Database (CTD)(Davis et al., 2020). See Appendix A.1 for details about our entity annotation process.

3.2 Relation Annotation

As stated in Section 1, a primary goal of MiDRED is compatibility with existing microbiome KBs. As such, we align our relation classes to those used by major microbiome KBs and annotate four classes: *connecting*, *contrasting*, *pathogen*, and *no relation*. The *connecting* class aligns with positive classes (e.g., "associated," "increase," and "positive"), signifying a microbe is associated with a disease, while the *contrasting* class signifies a microbe that contrasts with a disease, aligning definitionally to negative classes (e.g., "reduce," "decrease," and "inhibit") (Qi et al., 2022; Janssens et al., 2018; Li et al., 2021; Zhang et al., 2022; Dai et al., 2021). We also include *pathogen*, which is a stronger, more causal relation compared to *connecting*, as well as *no relation* to help prevent positive bias. Each instance is double annotated by different annotators and conflicting annotations are resolved in a third annotation round. With this systematic approach, we achieved a high inter-annotator agreement (Fleiss' Kappa) of 0.710. See Appendix A.1.1 for additional details about the relation annotation process, class definitions, and examples (Table 7).

MiDRED's data splits are constructed by collecting the set of unique fact-triples (e.g., ⟨*head*

[1] https://huggingface.co/datasets/shangdatalab-ucsd/midred

Dataset	Entity Types	Relation Classes	Host Type	Negative Instances	Entity Spans	# Microbes	# Relation Instances
HMDAD (Ma et al., 2016)	Microbes/diseases	2	Human	✗	✗	292	483
SSI (Khettari et al., 2023)	Microbes/microbes	2	Human	✔	✔	N/A*	999
Bacteria Biotope (Bossy et al., 2019)	Microbes/diseases/habitats/locs	2	Varied	✗	✔	1,760	2,639
MiDRED	Microbes/diseases	4	Varied	✔	✔	5,590	3,116

Table 1: A comparison between our proposed dataset, MiDRED, and other microbiome text-mining datasets. MiDRED features a multi-class relation classification task with annotated negatives (the "no relation" class) and span-level entity annotations (*the SSI dataset does not provide manually annotated entities).

entity, *relation*, *tail entity*⟩). Unique triples are divided into train, development, and test splits using 0.8/0.1/0.1 ratios, resulting in no overlapping fact-triples between data splits. See Appendix A.2 for statistics on each data split.

Documents:		1,655
Entities:	All	12,027 (678)
	Microbes	5,590 (197)
	Diseases	6,437 (482)
Relationships:	All	3,116
	Connecting	1,744
	Contrasting	161
	Pathogen	920
	No relation	291

Table 2: Counts of annotated entities and relationships in MiDRED. Parenthesized values denote the number of unique concepts. For detailed statistics on train, development, and test splits, see Appendix A.2.

4 Baseline Experiments

We explore the performance of popular NLP models using MiDRED on Named Entity Recognition (NER) and Relation Extraction (RE) tasks to establish the baseline performance and highlight challenging areas for future development.

4.1 Named Entity Recognition

In our NER experiments, we treat each entity mention span individually. We tested three NER models on our corpus: BiLSTM-CRF (Hochreiter and Schmidhuber, 1997), BioBERT-CRF (Lee et al., 2019), and PubMedBERT-CRF (Gu et al., 2020). Sentences were transformed into hidden state vector sequences by the respective models. Each model was tasked with predicting the labels for each token within these sequences. Subsequently, a fully connected layer was employed to calculate the network score, and a conditional random field (CRF) layer decoded the optimal tag path from all possible paths, utilizing the BIO (Begin, Inside, Outside) tagging scheme to categorize each token accurately. See Appendix A.4 for hyperparameter details.

Model	P	R	F1
BiLSTM-CRF	0.877	0.891	0.884
PubMedBERT-CRF	0.947	0.972	0.959
BioBERT-CRF	**0.957**	**0.981**	**0.969**

Table 3: Precision, recall, and F1-micro scores of various NER models on the MiDRED test set. Results are averages from three runs.

4.2 Relation Extraction

For RE experiments, we explore fine-tuning encoder-only biomedical language models (BioLinkBERT (Yasunaga et al., 2022) and PubMedBERT (Gu et al., 2020)). We send representations for the [CLS] token through a fully connected layer trained with cross-entropy. Additionally, we explore the current in-context learning abilities of frontier LLMs (GPT 3.5 (OpenAI, 2021) and GPT 4 (OpenAI et al., 2024))[2]. For details on the prompt we use, see Appendix A.5.

5 Results and Discussion

Figure 1 displays the top ten microbes and diseases and the distribution of relation classes in MiDRED. We observe a long-tail distribution for both entity types. The distribution of microbes, in particular, features a steep drop-off in mention frequency after the most mentioned microbe, Helicobacter pylori, indicating that current research focuses on a relatively narrow set of microbes.

We observe relatively high scores for both the NER (Table 3) and RE (Table 4) experiments when looking at performance across all test instances using small, fine-tuned biomedical language models ($\text{PubMedBERT}_{\text{base}}$ and $\text{BioLinkBERT}_{\text{large}}$), indicating the effectiveness of modern information ex-

[2]Specifically, we use *gpt-3.5-turbo-16k-0613* and *gpt-4-turbo-preview* via OpenAI's API, accessed on 5/3/2024.

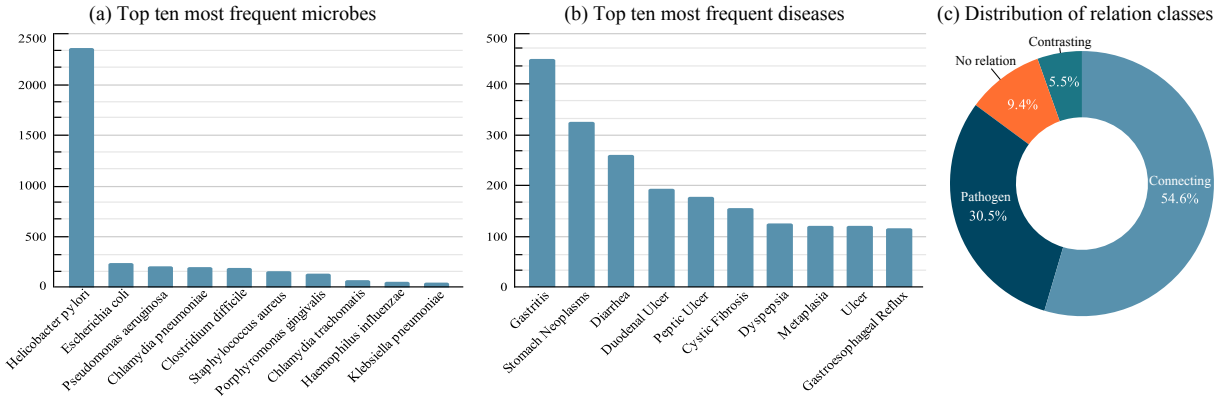

Figure 1: Counts of the top ten most frequent (a) microbial and (b) disease concepts, as well as (c) the distribution of relation classes found in the combined splits of MiDRED.

Model	P	R	F1
$\text{PubMedBERT}_{\text{base}}$	0.867	0.855	0.861
$\text{BioLinkBERT}_{\text{large}}$	**0.907**	**0.904**	**0.905**
GPT 3.5	0.542	0.562	0.552
GPT 4	0.716	0.725	0.721

Table 4: Precision, recall, and F1-micro scores of relation extraction models on the test set.

Model	Q1	Q2	Q3	Q4
$\text{PubMedBERT}_{\text{base}}$	0.895	0.852	0.800	0.571
$\text{BioLinkBERT}_{\text{large}}$	**0.929**	**0.839**	**0.801**	0.601
GPT 3.5	0.512	0.533	0.402	0.600
GPT 4	0.696	0.710	0.606	**0.667**

Table 5: F1-micro scores of RE models on test instances decomposed into quartiles based on microbe frequency, where Q1 is the performance on triples containing the top 25% most frequent microbes across all of MiDRED, followed by Q2, Q3, and finally, the least frequent quartile of microbes in Q4.

traction methods. Large, general domain language models (GPT 3.5 and GPT 4) leveraging in-context learning struggle to identify relations compared to smaller biomedical language models. This aligns with Peng et al. (2024)'s findings, offering additional evidence that large language models have yet to overtake smaller language models in information extraction tasks.

Furthermore, Table 5 shows a steady drop-off in $\text{PubMedBERT}_{\text{base}}$ and $\text{BioLinkBERT}_{\text{large}}$'s performance across quartiles of test triples decomposed based on microbe frequency, while the performance of GPT 3.5 and 4 remains relatively stable. This indicates that the smaller models generalize poorly and signify an area for future development.

Annotation Challenges: Numerous challenges were encountered when annotating microbes, diseases, and their associations. Challenges with acronyms and abbreviations arose due to variations in naming conventions, which sometimes differed from standard classifications. Relation types posed difficulties in accurately describing the links between microbe-disease pairs, particularly in cases involving numerical data or complex biological semantics. We record these and other challenges in Appendix A.3 in hopes of improving future versions of MiDRED and biomedical annotation efforts in general.

6 Conclusion

Microbiota, integral to human health and prevalent in various body environments like the gut, skin, and oral cavity, are at the forefront of promising research avenues that could revolutionize our understanding and treatment of numerous health conditions. However, the manual curation of microbiome knowledge bases, though invaluable, faces scalability challenges in keeping pace with the rapid influx of new research findings. In this paper, we introduce MiDRED, a dataset that aims to bridge this gap by providing a resource to help automate the creation and maintenance of microbiome bases. MiDRED can be used to train and validate state-of-the-art NLP models on various tasks such as named entity recognition, named entity normalization, bacteria-disease relationship extraction, and knowledge graph creation. We hope MiDRED will unlock new applications and innovations within microbiome research.

Limitations

MiDRED is a sentence-level annotated dataset, which inherently limits its scope to capturing relationships expressed within individual sentences. Consequently, the dataset does not encompass inter-sentence relationships, which could provide additional context and depth to understanding microbe-disease interactions. Furthermore, MiDRED maintains a focused thematic scope, exclusively concentrating on relationships between microbes and diseases. While beneficial for depth and specificity in this area, this focus excludes potential relationships involving other biological entities or environmental factors that could influence or be influenced by the microbe-disease dynamics. Such annotations could offer deeper insights into the context and contingencies of the documented relationships. We aim to address these limitations in future versions of the dataset.

Ethics Statement

In the development and release of the MiDRED dataset, we have carefully considered ethical aspects and do not anticipate any major ethical concerns. The dataset is constructed from publicly available academic articles, focusing solely on the relationships between microbes and diseases without involving individual patient data or personal information. By openly releasing the MiDRED dataset, we commit to facilitating transparency in our research process. This open access approach allows for peer review, replication of results, and collaborative improvements to the dataset.

Acknowledgements

Many thanks to Ho-Cheol Kim and Yannis Katsis for their support in all aspects of this project. Thank you to the anonymous reviewers for their thoughtful comments and corrections. This work is supported by IBM Research AI through the AI Horizons Network.

References

pubtator central: automated concept annotation for biomedical full text articles.

Varsha D. Badal, Dustin Wright, Yannis Katsis, Ho-Cheol Kim, Austin D. Swafford, Rob Knight, and Chun-Nan Hsu. 2019. Challenges in the construction of knowledge bases for human microbiome-disease associations. *Microbiome*, 7.

Ga Hyeon Baek, Ki-Myeong Yoo, Seon-Yeong Kim, Da Hee Lee, Hayoung Chung, Suk-Chae Jung, Sung-Kyun Park, and Jun-Seob Kim. 2023. Collagen Peptide Exerts an Anti-Obesity Effect by Influencing the Firmicutes/Bacteroidetes Ratio in the Gut. *Nutrients*, 15.

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. 2019. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In *Conference on Empirical Methods in Natural Language Processing*.

Liang Cheng, Changlu Qi, Zhuang He, Tongze Fu, and Xue Zhang. 2019. gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Research*, 48:D554 – D560.

Ilseung Cho and Martin J. Blaser. 2012. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13:260–270.

Die Dai, Jiaying Zhu, Chuqing Sun, Min Li, Jinxin Liu, Sicheng Wu, Kang Ning, Li-jie He, Xing-Ming Zhao, and Wei-Hua Chen. 2021. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Research*, 50(D1):D777–D784.

Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. 2020. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49:D1138 – D1143.

Janet Piñero González, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura Inés Furlong. 2019. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48:D845 – D855.

Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of biomedical informatics*, 46 5:914–20.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9:1735–1780.

William P Hogan, Molly Huang, Yannis Katsis, Tyler Baldwin, Ho-Cheol Kim, Yoshiki Baeza, Andrew Bartko, and Chun-Nan Hsu. 2021. Abstractified Multi-instance Learning (AMIL) for Biomedical Relation Extraction. In *3rd Conference on Automated Knowledge Base Construction*.

Zhiqiang Huang, Kun C. Liu, Wenwen Ma, Dezhi Li, Tianlu Mo, and Qing Liu. 2022. The gut microbiome in human health and disease—where are we and where are we going? a bibliometric analysis. *Frontiers in Microbiology*, 13.

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. TeamTat: a collaborative text annotation tool.

Yorick Janssens, Joachim Nielandt, Antoon Bronselaer, Nathan Debunne, Frederick Verbeke, Evelien Wynendaele, Filip Van Immerseel, Yves-Paul Vandewynckel, Guy De Tré, and Bart de Spiegeleer. 2018. Disbiome database: linking the microbiome to disease. *BMC Microbiology*, 18.

Oumaima El Khettari, Solen Quiniou, and Samuel Chaffron. 2023. Building a Corpus for Biomedical Relation Extraction of Species Mentions. In *Workshop on Biomedical Natural Language Processing*.

Yulianto Kusnadi, Mgs Irsan Saleh, Zulkhair Ali, Hermansyah Hermansyah, Krisna Murti, Zen Hafy, and Eddy Yuristo. 2023. Firmicutes/Bacteroidetes Ratio of Gut Microbiota and Its Relationships with Clinical Parameters of Type 2 Diabetes Mellitus: A Systematic Review. *Open Access Macedonian Journal of Medical Sciences*.

Natasha Z. Kustrimovic, Raffaella Bombelli, Denisa Baci, and Lorenzo Mortara. 2023. Microbiome and Prostate Cancer: A Novel Target for Prevention and Treatment. *International Journal of Molecular Sciences*, 24.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. 2023. BioREx: Improving Biomedical Relation Extraction by Leveraging Heterogeneous Datasets. *ArXiv*.

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*, 36:1234 – 1240.

Jiacheng Li. 2022. SPOT: Knowledge-Enhanced Language Representations for Information Extraction. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

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: The Journal of Biological Databases and Curation*, 2016.

Longqing Li, Qingxu Jing, Sen Yan, Xuxu Liu, Yuanyuan Sun, Defu Zhu, Dawei Wang, Chenjun Hao, and Dongbo Xue. 2021. Amadis: A Comprehensive Database for Association Between Microbiota and Disease. *Frontiers in Physiology*, 12.

Haiyan Liu, Pingping Bing, Mei jun Zhang, Geng Tian, Jun Ma, Haigang Li, Meihua Bao, Kunhui He, Jianjun He, Binsheng He, and Jialiang Yang. 2023. MNNMDA: Predicting human microbe-disease association via a method to minimize matrix nuclear norm. *Computational and Structural Biotechnology Journal*, 21:1414 – 1423.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia Noemi Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23.

Susan V. Lynch and Oluf Pedersen. 2016. The Human Intestinal Microbiome in Health and Disease. *The New England Journal of Medicine*, 375 24:2369–2379.

Wei Ma, Lu Zhang, Pan Zeng, Chuanbo Huang, Jianwei Li, Bin Geng, Jichun Yang, Wei Kong, Xuezhong Zhou, and Qinghua Cui. 2016. An analysis of human microbe–disease associations. *Briefings in Bioinformatics*, 18(1):85–97.

OpenAI. 2021. ChatGPT-3.5: Optimizing Language Models for Dialogue. Available: https://openai.com/blog/chatgpt-3-5/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo,

Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report.

A.C. Parte, J. Sardà Carbasse, J.P. Meier-Kolthoff, L.C. Reimer, and M. Göker. 2020. List of Prokaryotic names with Standing in Nomenclature (LPSN).

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. MetaIE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks.

Changlu Qi, Yiting Cai, Kai Qian, Xuefeng Li, Jia-Yi Ren, Ping Wang, Tongze Fu, Tianyi Zhao, Liang Cheng, Lei Shi, and Xue Zhang. 2022. gutMDisorder v2.0: a comprehensive database for dysbiosis of gut microbiota in phenotypes and interventions. *Nucleic Acids Research*, 51:D717 – D722.

Eric W. Sayers, Jeff Beck, Evan E. Bolton, Devon Bourexis, James Rodney Brister, Kathi Canese, Donald C. Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa J. Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L. Madden, Nuala A. O'Leary, Lon Phan, Sanjida H. Rangwala, Valerie A. Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W. Trawick, Kim D. Pruitt, and Stephen T. Sherry. 2020. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*.

Rasnik K. Singh, Hsin Wen Chang, Di Yan, Kristina M. Lee, Derya Uçmak, Kirsten Wong, Michael Abrouk, Benjamin Farahnik, Mio Nakamura, Tian Hao Zhu, Tina Bhutani, and Wilson J. Liao. 2017. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, 15.

Giorgos Skoufos, Filippos S. Kardaras, Athanasios Alexiou, Ioannis Kavakiotis, Anastasia Lambropoulou, Vasiliki Kotsira, Spyros Tastsoglou, and Artemis G. Hatzigeorgiou. 2020. Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Research*, 49:D1328 – D1333.

Mattia Di Stefano, Alessandro Polizzi, Simona Santonocito, Alessandra Romano, Teresa Lombardi, and Gaetano Isola. 2022. Impact of Oral Microbiome in Periodontal Health and Periodontitis: A Critical Review on Prevention and Treatment. *International Journal of Molecular Sciences*, 23.

George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. 2021. Re-TACRED: Addressing Shortcomings of the TACRED Dataset. In *AAAI Conference on Artificial Intelligence*.

Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak, and Tudor I. Oprea. 2010. ChemProt: a disease chemical biology database. *Nucleic Acids Research*, 39:D367 – D372.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database: The Journal of Biological Databases and Curation*, 2016.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *ArXiv*, abs/2303.03846.

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2022. Can NLI Provide Proper Indirect Supervision for Low-resource Biomedical Relation Extraction? *ArXiv*, abs/2212.10784.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. *ArXiv*, abs/1906.06127.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining Language Models with Document Links. In *Annual Meeting of the Association for Computational Linguistics*.

Irene S. Yu, Rongrong Wu, Yoshihisa Tokumaru, Krista P. Terracina, and Kazuaki Takabe. 2022. The Role of the Microbiome on the Pathogenesis and Treatment of Colorectal Cancer. *Cancers*, 14.

J Zhang, Xiqian Chen, Jiaxin Zou, Chen Li, Wanying Kang, Yang Guo, Sheng Liu, Wenjing Zhao, Xiangyu Mou, Jiayuan Huang, and Jia Ke. 2022. MADET: a Manually Curated Knowledge Base for Microbiomic Effects on Efficacy and Toxicity of Anticancer Treatments. *Microbiology Spectrum*, 10.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. 2018. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Conference on Empirical Methods in Natural Language Processing*.

A Appendix

A.1 Annotation Details

We developed an in-house annotation tool with functionality similar to open-source annotation tools such as Islamaj et al. (2020) to aid in annotating entities and relationships. When annotating articles, annotators can tag text and select "disease" or "microbe" based on the entity they intend to annotate. Depending on their choice, a select box displays a list of microbes from the List of Prokaryotic names with Standing in Nomenclature (LPSN) (Parte et al., 2020) dictionary or diseases based on the Comparative Toxicogenomics Database (CTD) (Davis et al., 2020) dictionary, allowing for the selection of an ontology concept. The annotation tool presents annotators with a list of potential microbe or disease concept matches sorted based on the mention text's similarity to the concepts and concept synonyms in the corresponding ontology. Individual diseases and microbe concepts can also be searched for using quotes. The selection of either the disease or microbe allows for the normalization of entities.

Additionally, the annotation tool we developed has multiple features to aid the annotating process. It underlines the annotated text based on selected entities, with microbe entities underlined in purple and disease entities underlined in orange, allowing for quick verification by the annotator. Furthermore, annotators can quickly cycle through, delete, and clear annotations using select keys, decreasing the annotating process's time-intensiveness.

Normalization is a classification process that classifies the different named entities of the same disease or microbe into a unique concept. Annotators were instructed to label microbes and diseases, including full names, abbreviations, synonyms, and acronyms. Adjectives and entities beyond LPSN and CTD databases were not annotated.

A.1.1 Annotating Relationships

After our entity annotation process, single sentences containing at least one microbe-disease pair were extracted and split into two subgroups. Sentences that had 80 characters or less and contained a rule-based keyword (Table 6) were placed into *Group A*, while all other sentences were placed into *Group B*. *Group A* sentences were then given pre-labels by rule-based algorithms (Table 6) concluded from observations in pilot annotation trials. Each assigned relation type was later manually verified by human annotators. Sentences in *Group B* were all manually labelled with relation types by human annotators. Each sentence across both groups were doubly-annotated to ensure the accuracy of the annotations. Instances of conflicting annotations were re-visited and relabeled in a third round of annotation. Using this process, we observe an inter-annotator agreement (Fleiss' Kappa) of 0.710, indicating high annotator agreement.

In pilot observation trials, we found that in describing relationships in which the microbial entity favored the development of the disease entity, a positive relation type was insufficient to encompass all associations. Thus, we employed two positive relation types of *pathogen* and *connecting*. *Pathogen* is used for more explicitly defined cases, where the

Relation type	Rule-based Keywords
Connecting	Associated, antibody, initiate, increase, develop, positive, accelerate, triggered, recognized, identify, colonized, diagnose, eradication+decrease, isolate
Contrasting	Reduce, decrease, eradication+increase, inhibit+proliferation, induced+delayed, inhibit
Pathogen	Caused, pathogen, agent, induce, due to
No relation	Not associated with, not present in, no effect against

Table 6: Keywords for pre-labelling rules used for annotating relationships.

microbe is a pathogen or causative agent for the disease or characterizes a particular sub-type of the disease. *Connecting* is used when the microbe is associated with or is a risk factor for the disease. See Table 7 for definitions of each relation class.

A.2 Data Splits

As mentioned in Section A.1.1, MiDRED is split using a holdout set of fact triples. This ensures that trained models cannot simply memorize relationships between head and tail entities. In Table 8, we show the statistics of each data split in MiDRED.

A.3 Challenges

In this section, we openly discuss the challenges we faced in annotating microbe and disease entities and associating relations. We hope these lessons will inform subsequent versions of MiDRED and future biomedical annotation efforts.

A.3.1 Challenges with acronyms and abbreviations:

While microbial and disease entities in this dataset were fully normalized to respective classification standards, challenges and limitations were encountered during the annotation process. As a nomenclature convention, bacterium names are often abbreviated after the first introduction. As a result, bacteria mentions had to be normalized, with its abbreviated form, which could differ from paper to paper. Similar challenges were found in disease acronyms, while compounded and embedded naming involving disease acronyms brings extra complexity. Moreover, while bacteria mentions follow relatively rigid and uniform nomenclature standards, disease mentions are more flexible and versatile according to authors' naming and writing style. With the differing naming techniques of authors, disease and bacteria entities were occasionally not encompassed by LPSN or CTD dictionaries and, therefore, unable to be annotated. Unnormalized entities were excluded from MiDRED and thus could be missed in developing computational models.

A.3.2 Challenges with relation types:

We found that *pathogen*, *connecting*, *contrasting*, and *no relation* relation types could not describe the linking relation between all microbe-disease pairs. As mentioned in the Limitations section, annotation units were annotated in single sentences, which led to lost context and instances where we could not determine a relation type and associations. A similar problem occurred when numbers were involved, for instance:

> ***Helicobacter pylori*** was found in 12 of 13 **AIDP** patients (92%), and in 10 of 20 controls (50%), (P = 0.02). (PMID: 15679702)

Although the four relation types, particularly *connecting* and *contrasting*, could be inferred from cases in which numbers were involved, more often than not, we felt that the numbers were taken out of context, and perceived relation types could be inaccurate from just the sentence. Consequently, we decided not to label all cases in which numbers were needed to determine relation types.

As the proposed four relation types were used to successfully label most annotation units, two commonly encountered complexity issues need to be further addressed in future annotation efforts:

1. **Relations Dependent on Quantitative Semantics:** As *connecting* and *contracting* relation types categorize the directions of associated development, which are often hinted at by keywords, more specific descriptions of experiments are often presented in quantitative data. As a single sentence can only provide limited information, the implication of the quantities is sometimes indefinite, as in the following example:

   > During the study period, a total of 373 blood cultures were obtained from patients in whom **brucellosis** was suspected, and 27 (7.2%) of

Relation type	Definition	Example
Connecting	The microbe is a risk factor for the development of the disease.	BACKGROUND: The presence of ***Mycoplasma pneumoniae*** has been associated with worsening **asthma** in children.
	The microbe is associated with the disease.	The ***Helicobacter pylori*** (H. pylori) bacterium has been classified by the World Health Organization as a type 1 carcinogen with associations to the development of peptic and gastric ulcers, **gastric carcinoma** and primary B-cell lymphoma.
Contrasting	The microbe or substances extracted from it is beneficial for the treatment of the disease.	Bacille Calmette-Gurin (BCG), an attenuated strain of ***Mycobacterium bovis***, is one of the most effective agents in the treatment of **superficial bladder cancer**.
	The microbe is beneficial in the improvement of the disease.	CONCLUSION: ***Lactobacillus reuteri*** effectively reduced the duration of acute **diarrhea** and hospital stays in children hospitalised with acute gastroenteritis.
No relation	No association between disease and microbe.	A high density of ***H. pylori*** colonization in the gastric mucosa was not associated with a higher frequency of **dyspepsia** (P > 0.80).
Pathogen	The microbe is a pathogen/causative agent for the disease.	*Orientia tsutsugamushi* (***O. tsutsugamushi***), the causative agent of **scrub typhus**, is an obligate intracellular pathogen.
	The microbe name is used immediately preceding the disease name to form a specific subtype of the disease	Fifteen children (41%) had ulcers associated with ***H. pylori*** **gastritis**, including all 10 children with a chronic ulcer.

Table 7: Classification standards for microbe-disease relation annotation used when annotating MiDRED. Annotated microbe and disease concepts are in bold.

them, drawn from 21 different patients, were positive for ***B. melitensis***. (PMID: 7989539)

2. **Relations Dependent on Biological Semantics:** The relations between microbe and host are essentially dynamic biological processes that, in many cases, can hardly be interpreted without implementing biological semantics. For instance, concepts such as vaccine, attenuated strains, microbe eradication, and co-infections are sometimes used in sentences, and excluding these semantics in the annotation process often leads to incorrect labels. Below is an example that our annotators found ambiguous without additional context from biological semantics:

   These results demonstrate that ***B. burgdorferi***-specific T lymphocytes primed by vaccination with a whole-cell preparation of inactivated B. burgdorferi sensu stricto isolate C-1-11 in adjuvant are involved in the development of **severe destructive arthritis**. (PMID: 7890402)

   In the current version of MiDRED, such instances

Annotations		Train	Dev	Test	All
Documents		1,521	521	549	1,655
Entities	All	8,985 (613)	1,452 (311)	1,590 (310)	12,027 (678)
	Microbes	4,182 (179)	687 (100)	721 (95)	5,590 (197)
	Diseases	4,803 (435)	765 (212)	869 (216)	6,437 (482)
Relationships	All	2,169	447	500	3,116
	Connecting	1,224	248	272	1,744
	Contrasting	100	29	32	161
	Pathogen	635	132	153	920
	No relation	210	38	43	291

Table 8: Counts of entities and relationships annotated in the MiDRED dataset across the train, development, and test data splits. Parenthesized values denote counts of unique concepts.

are excluded from the dataset as they cast challenges for human annotators and the design of the classification standards. We intend to rectify these issues in future versions of the dataset.

A.3.3 Challenges with relation annotations:

There were some limitations to the rule-based pre-labelling that we employed, as we could not assign rule-based pre-labels to *Group B* sentences. The reasoning behind this was twofold. *Group B* housed all the sentences without rule-based keywords (Table 6), so we could not give pre-labels by rule-based algorithms as we had done with *Group A*. Furthermore, *Group B* sentences were longer, and relations dependent on biological semantics were encountered more often, which required human annotators to interpret individual cases. Based on these challenges, we decided to forego rule-based pre-labeling on *Group B* sentences, resulting in these sentences being subject to more ambiguity.

A.4 Baseline NER Settings

For our NER experiments in 4, we use the following hyperparameter settings: 1,024 embedding dimensions, 512 max sequence length, and 64 batch size. We trained BioLinkBERT-CRF and PubMedBERT-CRF over three epochs and the BiLSTM-CRF for ten epochs.

A.5 GPT 3.5 and GPT 4 Prompts

GPT 3.5 and GPT 4 often perform better on tasks with the help of in-context learning (Wei et al., 2023; Wang et al., 2023). We construct a prompt that lists all relation classes and offers a couple of examples of extracted relationships. The following is the prompt we used for soliciting predictions for our tests:

> You are a relation extraction expert tasked with labeling relationships between head and tail entities in a sentence. Each example below has the head and tail entities appended to the sentence in the form: (head: head entity) (tail: tail entity). Predict if the sentence expresses one of the four following relation classes: "no relation", "connecting", "contrasting", "pathogen". The following are some examples:
>
> ### Sentence: At day 0 , 25 acute ulcers were associated with chronic H. pylori gastritis ; one patient had neither gastritis nor H. pylori infection (head: "H. pylori") (tail: "ulcers")
>
> ### Label: connecting
>
> *. . . [We include 4x examples of each relation class in the prompt.] . . .*
>
> ### Sentence: Significant resistance enhancement of mice pretreated with P. acnes against vaccinia virus or herpes simplex virus type 1 infection was observed. (head: "P. acnes") (tail: "herpes simplex")
>
> ### Label: ?

GPT 3.5 and GPT 4 responses were then aligned to ground truth classes via partial string matching for evaluation.