Sample size considerations for fine-tuning Large Language Models for Named Entity Recognition Tasks: A methodological study

Zoltan P. Majdik, S. Scott Graham, Sabrina N. Rodriguez, Martha S. Karnes, Jared T. Jensen, Joshua B. Barbour, Justin F. Rousseau

Submitted to: JMIR AI
on: August 22, 2023

Table of Contents

Sample size considerations for fine-tuning Large Language Models for Named Entity Recognition Tasks: A methodological study

Zoltan P. Majdik[1]; S. Scott Graham[2] BA, MA, PhD; Sabrina N. Rodriguez[3]; Martha S. Karnes[4]; Jared T. Jensen[2]; Joshua B. Barbour[5]; Justin F. Rousseau[6]

[1]North Dakota State University Fargo US
[2]The University of Texas at Austin Austin US
[3]The Dell Medical School at The University of Texas at Austin Austin US
[4]Curry College Milton US
[5]The University of Illinois at Urbana-Champaign Urbana US
[6]The University of Texas Southwestern Medical Center Dallas US

Corresponding Author:
S. Scott Graham BA, MA, PhD
The University of Texas at Austin
Parlin Hall 29
Mail Code: B5500
Austin
US

Abstract

Background: Large language models (LLM) have the potential to support promising new applications in health informatics. However, there is a lack of practical data available on sample size considerations for fine-tuning LLMs to perform specific tasks in biomedical and health policy contexts.

Objective: To evaluate sample size and sample selection techniques for fine-tuning LLMs to support improved named-entity recognition (NER) for a custom dataset of conflict of interest (COI) disclosure statements.

Methods: A random sample of 200 disclosure statements was prepared for annotation. All PERSON and ORG entities were identified by each of the two raters and once appropriate agreement was established, the annotators independently annotated an additional 290 disclosure statements. From the 490 annotated documents, 2500 stratified random samples in different size ranges were drawn. The 2500 training set subsamples were used to fine-tune RoBERTa models for improved NER, and multiple regression was used to assess the relationship between sample size (sentences), entity density (entities per sentence or EPS) and trained model performance (F1). Additionally, single-predictor threshold regression models were used to evaluate the possibility of diminishing marginal returns from increased sample size or entity density.

Results: Fine-tuned models ranged in overall NER performance from F1 = 0.433 to F1 = 0.936, with an average model performance of F1 = 0.836 (SD = 0.135). The two-predictor multiple linear regression model was statistically significant, F(2,2497) = 2034, P<.001; multiple R2 = 0.6197. The estimates for each independent variable were also statistically significant with ?EPS = 0.04 (95% CI: 0.02 to 0.06) and ?sent = 0.0004 (95% CI: 0.00034 to 0.00036).The threshold model for total sentences estimates that a diminishing margin of return occurs at 448 sentences (95% CI: 437 to 456), p = 0. The threshold model for EPS likewise indicates a diminishing margin of return at a token density of 1.36 (95% CI: 1.35 to 1.37), P>.001.

Conclusions: Relatively modest sample sizes can be used to fine-tune LLMs for NER tasks applied to biomedical text, and training data entity density should representatively approximate entity density in production data.

(JMIR Preprints 22/08/2023:52095)
DOI: https://doi.org/10.2196/preprints.52095

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Sample size considerations for fine-tuning Large Language Models for Named Entity Recognition Tasks: A methodological study

Abstract

Background: Large language models (LLMs) have the potential to support promising new applications in health informatics. However, there is a lack of practical data available on sample size considerations for fine-tuning LLMs to perform specific tasks in biomedical and health policy contexts.

Objective: To evaluate sample size and sample selection techniques for fine-tuning LLMs to support improved named-entity recognition (NER) for a custom dataset of conflict of interest (COI) disclosure statements.

Methods: A random sample of 200 disclosure statements was prepared for annotation. All PERSON and ORG entities were identified by each of the two raters and once appropriate agreement was established, the annotators independently annotated an additional 290 disclosure statements. From the 490 annotated documents, 2500 stratified random samples in different size ranges were drawn. The 2500 training set subsamples were used to fine-tune a selection of language models across two model architectures (BERT and GPT) for improved NER, and multiple regression was used to assess the relationship between sample size (sentences), entity density (entities per sentence or EPS) and trained model performance (F1). Additionally, single-predictor threshold regression models were used to evaluate the possibility of diminishing marginal returns from increased sample size or entity density.

Results: Fine-tuned models ranged in topline NER performance from F1 = 0.79 to F1 = 0.96 across architectures. Two-predictor multiple linear regression models all were statistically significant with multiple R^2 ranging from 0.6057 to 0.7896. EPS and number of sentences were significant predictors of F1 scores in all cases except for the GPT-2 model where EPS was not significant. Model thresholds indicate points of diminishing marginal return from increased training dataset sample size measured by number of sentences, with point estimates ranging from 439 sentences for RoBERTa_large to 527 sentences for GPT-2. Likewise, the threshold regression models indicate a diminishing marginal return for EPS with point estimates between 1.36 and 1.38.

Conclusion: Relatively modest sample sizes can be used to fine-tune LLMs for NER tasks applied to biomedical text, and training data entity density should representatively approximate entity density in production data. Training data quality and a model architecture's intended use (text generation vs. text processing/classification) may be as, or more, important as training data volume and model parameter size.

Keywords: Named-Entity Recognition; Large Language Models; Fine-Tuning; Transfer

Learning; Expert Annotation; Sample Size; Sample

Introduction

Named entity recognition (NER) has many applications in biomedical and clinical natural language processing (cNLP). As its core function, NER identifies and categorizes specific terms or phrases representing people, places, organizations, and other entities. It has been used to identify or extract named entities in free text clinical notes and reports in the secondary analysis of electronic health records [1,2]. NER also has been used alone or as part of an NLP pipeline to detect protected health information in order to de-identify clinical text for secondary analysis [3,4]. Additionally, NER has been used to identify and classify medications[5,6], specific disease and clinical condition entities [7], and laboratory tests [8] into existing taxonomies for purposes of secondary research, cohort generation, or clinical decision support [9–12]. While NER solutions have a long history for applications in NLP and cNLP domains, their effectiveness has recently been enhanced through the addition of Large Language Models (LLMs) in relevant data parsing pipelines. LLMs, broadly, have become an integral part of research pipelines in fields as diverse as digital humanities [13], computational social science [14], bioinformatics, applied ethics, and finance.

LLMs like GPT-3 have demonstrated remarkable performance across a variety of tasks. For instance, the GPT-3.5-powered LLM application ChatGPT performed close to or at the passing threshold of 60% accuracy on the United States Medical Licensing Exam (USMLE) without the specialized input of human trainers [15]. Widely available models like Google's BERT or OpenAI's GPT series are trained, bi-directionally or uni-directionally, on large volumes of generic textual data, designed to represent a wide array of common language use contexts and scenarios [16]. In specialized use-contexts, these generic models often fail to accurately classify information because the language structures that require classification – their words, syntax, semantic context, and other textual or lexical signatures – are sparsely represented in the data that were used to train the generic model [17,18]. Some language models, like ElutherAI's GPT-J-6B, are trained on open-source language modeling datasets curated from a mix of smaller open web crawl data sets alongside more technical papers from PubMedCentral and arXiv, and can offer improved classification accuracy for technical applications [19]. Nevertheless, specialized tasks often require fine-tuning of general purpose LLMs. Fine-tuning provides a way of overcoming the limitations of generic LLMs by augmenting their training data with data selected to more accurately reflect the target domains toward which a model is fine-tuned. The fine-tuning process updates the model's parameters – the weights that affect which connections between the nodes and layers of a neural network become activated – and so helps a model permanently learn. Unlike practices like prompt engineering that leave the underlying language model untouched, fine-tuning changes the model itself, yielding a new model optimized for the specific use case.

Fine-turning LLMs to perform technical, specialized tasks is expensive, however. Because the target domain of a fine-tuned model is usually complex and technical – otherwise, fine-tuning would not be necessary – it requires annotators with some degree of domain-level expertise, which come with potentially significant financial and time costs. Indeed, one study of NER annotation speed found it can take between 10 and 30 seconds per sentence for experts to annotate named entities [8]. The gold-standard annotated BioSemantics corpus is composed of 163,219 sentences which implies an optimal annotation time of over 11 weeks at 40-hours per week (453.39 hours) [20]. This estimate, of course, excludes time required for annotator training and inter-annotator reliability assessments. And because fine-tuning adjusts many or all of the model's parameters, it also consumes computational resources. Compute time and power consumption for fine-tuning scales with training data size [21,22] and with the size of the underlying model. As of the date of writing, for example, it would be

unrealistic to fine-tune very large models like GPT-4.

These limitations notwithstanding, it is increasingly recognized that longstanding presumptions about sufficiently large training datasets are likely substantially inflated [23]. We suspect this comes from a research and development environment dominated by a significant focus on promulgating new models that can claim to be state-of-the-art (SOTA) based on some pre-identified benchmark. In a research environment dominated by so-called "SOTA chasing," ever larger datasets are often required to eke out minor performance improvements over the previous benchmarks. Notably, development teams from disciplines with generally small research budgets have found that fine-tuning can result in substantial performance improvements from relatively small amounts of expert-annotated data [13,24] or from a combination of pre-/transfer-learning followed by a brief fine-tuning phase [25]. In one case, significant improvements over baseline were derived from training samples as small as 50 lemmas [13]. Despite the growing recognition that smaller gold-standard training sets can provide substantial performance improvements, there is little in the way of actionable guidance for sample size and sample curation.

The primary goal of this paper is to establish some initial baselines for sample size considerations in terms of training set size and relevant entity density for NER applications in specialized technical domains. To that end, we have conducted a fine-tuning experiment that compares the performance improvements resulting from 2,500 randomly selected training datasets stratified by size. These training sets were used to fine-tune four distinct language models to perform NER in a highly specific language domain: identification of two internal components (conflict sources and conflict targets) in conflict of interest (COI) disclosures. The results presented below indicate that only relatively small samples are required for substantial improvement. They also demonstrate a rapidly diminishing marginal return for larger sample sizes. In other words, while larger and larger sample sizes may be useful for "SOTA-chasing," their value for fine-tuning LLMs shrinks beyond a certain threshold, which we estimate below. These findings provide actionable guidance about how to select and generate fine-tuning samples by attending to issues of relevant token density. As such, they should have great value for NER applications that rely on them.

Literature Review

During our initial review of the literature, we were unable to locate any widely accepted, evidence-based guidance on appropriate sample sizes for training data in NER fine-tuning experiments. Therefore, to evaluate the state of the field, we conducted a literature search focused on identifying existing practices. We searched PubMed for prior relevant work to determine current sample size conventions in NER fine-tuning. We used a simple search strategy "("named entity recognition" OR "entity extraction") AND (fine tuning OR transfer learning) AND (annotat*)" which returned 138 relevant articles. We reviewed each of these articles and extracted information related to human-annotated NER training sets. Specifically, for each article, we assessed if a human-annotated training set was used, and if so, we extracted data on sample units, sample size, and any available sample size justification. In cases where authors described the size of human-annotated training sets on multiple levels (e.g., N of documents, N of sentences, N of entities), we prioritized units that would most effectively guide prospective sampling. That emphasis meant we prioritized sentences (as comparable across document types and identifiable without annotation) over documents (which vary widely in length) or entities (which cannot be assessed until after annotation). In cases where multiple human-annotated samples were used, we noted the largest reported sample as indicative of the researchers' sense of the sample necessary to conduct the research in its entirety. Additionally, for each article that made use of a human-annotated training set, we sought to identify any possible justifications for the chosen sample size. We anticipated that common justifications might include (1)

collecting a sample sufficient to achieve target performance, (2) collecting a sample consistent with or larger than prior work, or (3) collecting a sample appropriate given relevant power calculations.

Of the articles surveyed, the majority (n = 93/138, 67.4%) reported use of human-annotated NER training data. The remaining 45 (32.6%) articles used only computational approaches to curate training data sets. Notably, many articles reported using a mix of human-annotated and computationally-annotated training sets and/or performing multiple experiments with different training sets. As long as any given article used at least one human-annotated training set, it was included in the tally. Reported sample units varied quite widely across articles with many reporting only the number of documents used. Document types were similarly variable and specific to research contexts. For example, several articles reported training sample sizes as the number of clinical notes, number of published abstracts, or number of scraped tweets. In contrast, some articles reported sample size using non-context specific measures such as sentences, entities, or tokens. Given this variety, we classified sample units as belonging to one of six common categories: clinical notes/reports, sentences, abstracts/articles, entities, tokens, or other. The most commonly used sample unit was clinical notes/reports (n = 35) followed by sentences and articles/abstracts (both n = 21). Sample size ranges also varied widely by unit type, as would be expected. The smallest clinical notes or reports sample used a scant 17 documents [26], but this was likely a larger sample than the smallest reported sentence sample size of 100 [27]. Among the articles reporting non-document type specific sample units, human annotated data sets ranged from 1480 tokens to 79,401 tokens (M = 42424 tokens); 100 entities to 39,876 entities (M = 15,597 entities); and 100 sentences to 36,938 sentences (M = 26,678 sentences). Details on sample size ranges by sample type are available in Table 1. Complete details on each article's approach to sample size are available in Multimedia Appendix 2.

Of the 93 articles that used human-annotated NER training data, only 3 (3.2%) provided an explicit justification for the chosen sample size. In each case, the justification for the sample size was based on reference to prior relevant work and determined to be as large or larger than a sample used in the previously published scholarship [28–30]. Ultimately, the wide range of sample reporting practices and the broad lack of attention to sample size justification indicate a strong need for explicit sample selection guidance for fine-tuning NER models. The current article contributes to addressing this need.

Table 1. Unit types, N of articles by type, and sample size ranges

Unit Type	N of Articles	Min	Mean	Max
Clinical notes/reports	34	17	709	5098
Abstracts/articles	21	20	1966	7000
Sentences	21	100	26,678	360,938
Other	9	47	5979	25,678
Entities	5	100	15,957	39,876
Tokens	3	1840	42,121	79,401

Methods

The primary aim of this study was to evaluate sample size considerations for fine-tuning LLMs for domain- and context-specific NER tasks. Specifically, the goal was to evaluate how changes in re-training dataset sizes and token density impact overall NER performance. To accomplish this task, we used stratified random samples of training sets to create 2500 fine-tuned instances of RoBERTa_base, GatorTron_base, RoBERTa_large, and GPT-2_large. In what follows, we describe

(1) the data and target NER task, (2) the gold standard annotation protocol, (3) the fine-tuning approach, and (4) our sample feature analysis.

Data description and context

We selected COI disclosures in biomedical literature as a highly domain-specific, technical language context suitable for the goals of this paper. In recent years, significant research efforts have been devoted to studying the effects of financial conflicts of interest (COI) on the biomedical research enterprise[31–33], finding that COI are associated with favorable findings for sponsors [31], increased rates of "spin" in published reports [34], increased likelihood of trial discontinuation or non-publication [35], editorial and peer reviewer biases [36], and increased adverse events rates for developed products [37]. Unfortunately, as compelling as this body of evidence is, a recent methodological review of research in this area indicates that most studies treat COI as a binary variable (present or absent) rather than quantifying COI rates or disaggregating COI types [32]. This limitation in the available evidence is, no doubt, driven in part by the data structures of COI reporting. When COI are reported, they are generally reported in unstructured or semi-structured text. COI disclosure statements can also be quite long, as individual authors frequently receive and report multiple lines of funding from a wide variety of granting agencies and corporate sponsors. Ultimately, the lack of tabular data structures for COI makes it difficult to extract appropriate information [38], such as the sources and recipients of funding, the precise links between COI sources and recipients, or the quantity and degree of COI in a given disclosure statement.

These limitations notwithstanding, there has been some recent research leveraging informatics techniques, including NER, to transform text disclosure statements into tabular data [18,37]. Recently developed systems leverage NER to identify authors and sponsors as PERSONs and ORGs, respectively. Secondary processing makes use of regular expressions to parse the types of relationships reported between each NER-identified PERSON and ORG. Since NER-tagging in this context is focused on identifying canonical entity types, applying these tools to COI disclosure statements may seem relatively straightforward at the outset. However, variances in reporting formats and the lack of specific training data on relevant entities presents a number of challenges. In the first case, author identification is stymied by different journal guidelines for rendering author names. For example, a disclosure statement for Rudolf Virchow might be rendered as "Rudolf Virchow," "Virchow," "Dr. Virchow," or "RLCV." Likewise, pre-trained NER models have not been found to offer high-quality out-of-the-box performance for pharmaceutical company names [18]. Variations in incorporation type (Inc, Llc, GmbH, etc.) typically induce entity boundary issues and multi-national companies often report national entity names (e.g., Pfizer India) leading standard NER models to assign inappropriate geopolitical entity tags. Finally, effective NER on COI disclosure statements is also challenged by atypical distribution of relevant tokens. It is not uncommon for a single sentence in a disclosure to have a dozen author names or a dozen company names, for example, when a disclosure statement lists all authors who have the same COI (e.g., "such-and-such authors are employed at MSD"). These atypical sentence structures also occur when a single author has many COI to disclose, as in, "RLCV receives consulting fees from MSD, Pfizer, GSK, Novartis, and Sanofi."

To more clearly demonstrate these limitations, we provide the following authentic example from a COI disclosure statement published in a 2018 issue of the *World Journal of Gastrointestinal Oncology* [39]. Here we show the NER tagging performance of RoBERTa_base without fine-tuning:

> Sunakawa Y[ORG] has received honoraria from Taiho Pharmaceutical[ORG], Chugai Pharma[ORG], Yakult Honsha[ORG], Takeda[ORG], Merck Serono[ORG], Bayer

> Yakuhin[ORG], Eli Lilly Japan[ORG], and Sanofi[ORG]; Satake H[ORG] has received honoraria from Bayer[ORG], Chugai Pharma[ORG], Eli Lilly Japan[ORG], Merck Serono[ORG], Takeda[ORG], Taiho Pharmaceutical[ORG] and Yakult Honsha[ORG]; Ichikawa W[ORG] has received honoraria from Chugai Pharma[ORG], Merck Serono[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG]; research funding from Chugai Pharma[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG].

And here, we show the NER tags provided by the human annotation team.

> Sunakawa Y[PERSON] has received honoraria from Taiho Pharmaceutical[ORG], Chugai Pharma[ORG], Yakult Honsha[ORG], Takeda[ORG], Merck Serono[ORG], Bayer Yakuhin[ORG], Eli Lilly Japan[ORG], and Sanofi[ORG]; Satake H[PERSON] has received honoraria from Bayer[ORG], Chugai Pharma[ORG], Eli Lilly Japan[ORG], Merck Serono[ORG], Takeda[ORG], Taiho Pharmaceutical[ORG] and Yakult Honsha[ORG]; Ichikawa W[PERSON] has received honoraria from Chugai Pharma [ORG], Merck Serono[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG]; research funding from Chugai Pharma[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG].

It is evident that the base LLM classifier makes critical errors that make mapping COI relationships between researchers, funding streams, and funding sources impossible. In the above example, a base-trained classifier mistakenly tags PERSONS as ORGs; elsewhere, we have seen the opposite, where non-fine-tuned classifiers mistakenly identify companies like Novartis or Eli Lilly as PERSON. General purpose language models (such as BERT, GPT-3, etc.) are not well-suited to the NER task of classifying and linking named authors and disclosed payors (pharmaceutical companies, nonprofit foundations, federal funders, etc.) because of challenges that arise from the aforementioned lack of standardized disclosure conventions for author names. Likewise, another challenge arises because these models are not well-trained on biomedical companies, nonprofit entities, and federal funders. In this study as well as earlier research, we found that pharmaceutical companies – frequently named after founding families – are often tagged as PERSON rather than ORGANIZATION. Finally, the linguistic signature of COI disclosure statements is distinctive: COI statements deploy semicolons in non-standard ways. For large research teams, a single disclosure sentence can cover the length of a long paragraph; and grammatical conventions that govern the relationship between subjects, direct objects, and indirect objects are often elided or circumvented in favor of brevity, which makes linking authors to payors and payors to type of payment challenging. At the same time, the linguistic conventions used for disclosure statements vary between and even within journals, rendering rule-based NER approaches unfeasible. As such, the task of identifying and linking authors to payors and payment types in COI statements is an ideal use case for fine-tuning parameter-dense language models based on gold standard human annotated COI statements.

Data sources and pre-processing

The data used for fine-tuning COI-relevant NER tags in this study come from COI disclosure statements drawn from 490 articles published in a diverse range of biomedical journals. The selected disclosure statements were randomly sampled from a preexisting dataset of 15,374 statements with A.I.-identified COI [40]. The original data set was created by extracting all PubMed-indexed COI statements in 2018. At the time of download there were 274,246 articles with a COI-statement field in the PubMed XML file. The substantial majority of these are statements of no conflict disclose, and so collected statements were analyzed using a custom machine-learning enhanced NER system that can reliably identify relationships between funding entities and named authors [18,37]. The sample used in this study was drawn from the population of COI statements with AI-confirmed conflict

disclosures.

Two annotators independently tagged named entities in the collected COI statements as either people (PERSON) or organizations (ORG). The PERSON tag was applied to all named authors, regardless of the format of the name. This included initials with and without punctuation, e.g. “JAD” or “J.A.D” as well as full names “Jane A. Doe” or names with titles “Dr. Doe”. ORG tags were applied to named pharmaceutical companies, nonprofit organizations, and funding agencies. To assure that NER tagging was consistent, a random sample of 200 COI statements was tagged by both annotators and assessed for inter-annotator agreement using inter-class correlation coefficient for unit boundaries and Cohen’s kappa for entity type agreement. The raters had 98.3% agreement on unit boundaries (ICC=0.87, 95% CI: 0.864-0.876). For named entities with identical unit boundaries, classification (PERSON or ORG) agreement was 99.6% (k=0.989). After this high degree of inter-rater reliability was established, the annotators independently annotated the remaining COI statements. Prior to training the language model, a third rater reconciled the few annotation disagreements in the initial IRR sample.

Model fine-tuning and analysis

A subset of 147 (30%) of the annotated disclosure statements were reserved to serve as an evaluation set. The remaining 343 statements were used to generate 2500 training sets for subsequent experimentation. Each set was created by randomly selecting an N size in five pre-identified strata of 40 possible sample sizes, at the statement level. The strata included size ranges of 1-40, 41-80, 81-120, 121-160, and 161-200. Once each N size was selected, a random sample of COI statements at that N size was derived. We created 500 random samples within each stratum.

We fine-tuned four commonly used language models using the open-source spaCy natural language processing library (version 3.2.1, running on python version 3.9.7). To ensure repeatability of results and to make the fine-tuning process as accessible as possible to research teams, we used spaCy’s default configuration settings for Named Entity Recognition. Selected models included RoBERTa_base, GatorTron_base, RoBERTa_large, and GPT-2_large; for the latter three, we used the `spacy-transformers` package to access these models through Hugging Face’s `transformers` library. These models were selected to provide a range of parameter sizes (125M to 744M) and to allow for a comparison between language models trained on general-use as well as on biomedical texts specifically. Fine-tuning was performed on spaCy’s pre-trained transformer pipeline, with only the ‘transformer’ and ‘NER’ pipeline components enabled in the configuration file. All fine-tuning processes were run on a high-performance computing cluster at North Dakota State University’s Center for Computationally Assisted Science and Technology (CCAST), using AMD EPYC CPUs and NVIDIA GPUs. Pre-processing and tokenization were done using spaCy’s built-in tokenizer; training runs were optimized with the Adam algorithm, with decay rates of 0.9 (beta1) and 0.999 (beta2) and a learning rate of 0.01. For each training run, spaCy was set to check NER classifications against the test set after every 200 iterations within an epoch, to generate language models at regular intervals during the training process, and to stop whenever additional training steps failed to improve the classification metrics. We then extracted the highest-scoring language model from each set, for a total of 2500 fine-tuned language models.

Each of the 2500 re-training sets was subsequently categorized by sample size (measured in number of sentences) and relevant entity density (entities per sentence or EPS). Sentence boundaries were determined using the sentencizer in the R tidytext (0.3.4) library [41]. Sentences were used to provide a more regularized comparator as disclosure statements vary widely in length. We also focus on sentences as opposed to tokens since the number of sentences in a sample can be identified

prospectively (i.e., prior to annotation). Multiple regression was used to assess the linear relationship between sample size (N of sentences), entity density (EPS) and trained model F1. Additionally, we used single-predictor threshold regression models for N of sentences and EPS to evaluate the possibility of diminishing marginal returns from increased sample size or taken density[42]. Threshold regression offers an effective way to model and evaluate non-linear relationships, and as the term suggests, to identify any threshold effects. Multiple threshold models are available, and our approach relies on a hinge model that can be expressed as follows:

$$\eta = \alpha_1 + \alpha_2^T z + \beta_1 (x - e)_{+¿¿}$$

All statistical tests were performed in R 4.2.2 and the threshold modeling was performed using the R chngpt package.[43]

Results

The 2500 sets ranged from 1 to 200 disclosure statements with an average of 100 (SD =57.42). The number of sentences in each fine-tuning set ranged from 5 to 1031, with an average of 712.9 (SD = 405.9). The tagged entity density ranged from 0.771 to 0.936 EPS, with an average of 0.836 (SD = 0.135). Fine-tuned model performance on NER tasks ranged from F1=0.3 to F1=0.96. The top F1 score for each architecture was 0.72 for GPT-2, 0.92 for GatorTron, 0.94 for RoBERTa_base, and 0.96 for RoBERTa_large. Dataset and model descriptive statistics are available in Table 2.

Table 2: Descriptive statistics of training sets and model performance

	Min	**Mean (SD)**	**Max**
N Docs	1	100.0 (57.42)	200
N Tokens	4	712.9 (405.94)	1402
N Sentences	5	525.2 (294.13)	1031
EPS	0.771	1.34 (0.14)	1.72
RoBERTa_base F1	0.43	0.81(0.13)	0.94
GatorTron_base F1	0.37	0.84 (0.13)	0.92
RoBERTa_large F1	0.44	0.84 (0.14)	0.96
GPT-2_large F1	0.30	0.58 (0.12)	0.72

Multiple linear regressions were used to assess and compare the relationship between the independent variables (number of sentences and EPS) and the overall model performances (measured by F1) for each architecture. EPS and number of sentences predictors correlate weakly (Pearson's r = 0.28, $p < 0.01$), and diagnostic tests for multicollinearity indicate that the variables do not violate Klien's rule of thumb and have a low variance inflation score (VIF = 1.11) and high tolerance (TOL = 0.9) [44]. All models were statistically significant with multiple R^2 ranging from 0.6057 to 0.7896. EPS and number of sentences were significant predictors of F1 scores in all cases except for the GPT-2 model where EPS was not a significant predictor. Standardized regression coefficients and full model results are available in Table 3.

Table 3: Standardized Multiple Linear Regression Results by Architecture

Model (Params)	**β_{EPS}**	**β_{sent}**	**F-statistic, model p-value , multiple R^2**
RoBERTa (125M)	0.04*	0.78*	$F(2,2497) = 2034$, $p < 0.001$, $R^2 = 0.6197$
GatorTron (345M)	0.05*	0.79*	$F(2,2497) = 2236$, $p < 0.001$, $R^2 = 0.6417$
RoBERTa (355M)	0.05*	0.76*	$F(2,2497) = 1918$, $p < 0.001$, $R^2 = 0.6057$
GPT-2 (774M)	-0.01	0.89*	$F(2,2497) = 4685$, $p < 0.001$, $R^2 = 0.7896$

* Predictor results significant at the $p < 0.01$ level.

In this study, we focus primarily on total sentences as our measure of data size. This is because the

number of sentences can be identified prospectively (prior to annotation) and is comparable across data sets with different document lengths. However, it should be noted that other measures of sample size are similarly predictive of F1 scores. The total number of relevant entities per training data set correlates very closely with the number of sentences (Pearson's r = 0.998, p < 0.01). This high collinearity makes it inadvisable to fit regression models with both predictors. We did, however, fit a series of models with EPS and number of relevant entities as predictors. In all cases, the results were quite similar to those reported in Table 3. Specific values are available in Appendix 2. It is notable that, in all cases, the multiple R^2 for models with EPS and number of relevant entities as predictors are lower than the counterpart models with EPS and number of sentences. Subsequent pairwise analyses of variance (ANOVA), however, indicate that there are no significant differences in model fit. ANOVA p-values were p = 0.85 for RoBERTa, p = 0.74 for GatorTron, p = 0.93 for RoBERTa_large, and p = 0.53 for GPT-2.

Threshold regression models were also used to assess the possibility of diminishing marginal returns on training data sizes and EPS for each model and model architecture. All threshold models indicate that there was a diminishing marginal return from increased training dataset sample size measured by number of sentences. Point estimates ranged from 439 for RoBERTa_large to 527 for GPT-2. Likewise, the threshold models indicate a diminishing marginal return for EPS with point estimates between 1.36 and 1.38. Complete threshold regression results are available in Table 4. Single predictor plots are available in Fig 1, with technical threshold model plots in Multimedia Appendix 2.

Table 4: Threshold Regression Results, Mean and Max F1 Scores by Architecture

Model (Params)	N Sent Threshold (95% CI)	EPS Threshold (95% CI)
RoBERTa (125M)	448 (437 – 456)	1.36 (1.35 – 1.37)
GatorTron (345M)	448 (409 – 456)	1.36 (1.36 - 1.38)
RoBERTa (355M)	439 (409 – 451)	1.36 (1.35 - 1.38)
GPT-2 (774M)	527 (511-540)	1.38 (1.36-1.38)

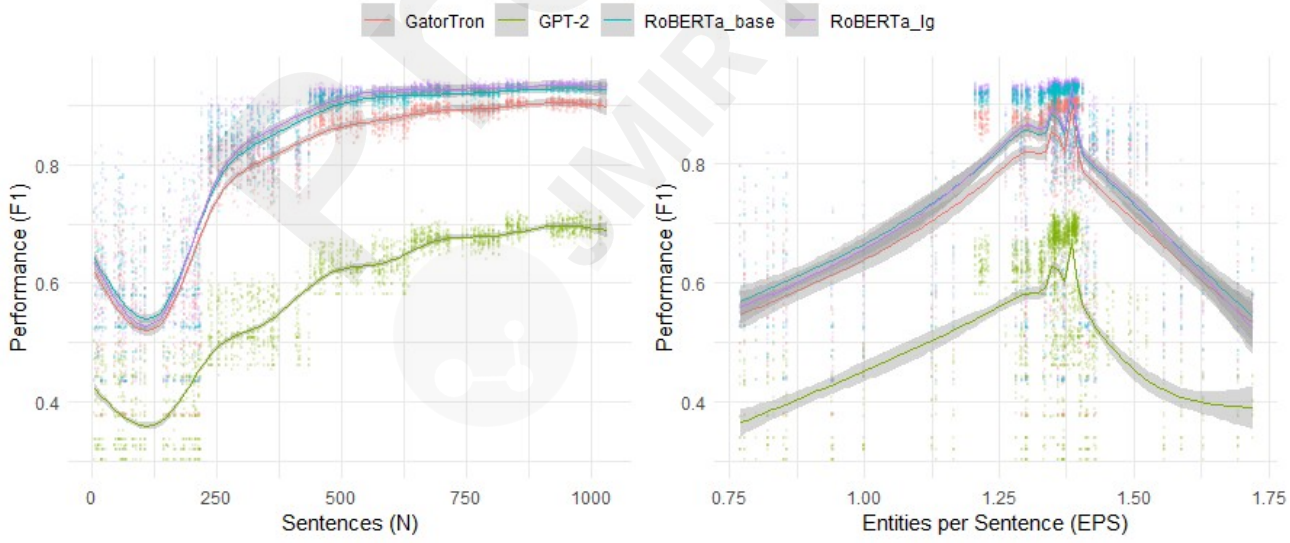


Fig. 1: Single predictor plots for N of sentence (left) and EPS (right). Fit with a generalized additive model.

Discussion

Our review of the available literature on human-annotated training data for NER fine-tuning indicates that there is a strong need for useful guidance on requisite sample sizes. Reported sample

units and sizes vary widely, providing little foundation for prospective approaches to sample curation. Given the significant time and costs associated with gold-standard annotation, it is critical that researchers and practitioners can effectively determine appropriate samples before fine-tuning neural network language models. The results of the experiment presented here provide initial actionable guidance for the development of gold-standard annotated training sets for NER fine-tuning in highly specific, specialized domains. Specifically, they indicate that contrary to common assumptions, transformer-based language models can be optimized for new tasks using relatively small amounts of training data. Furthermore, the results presented here indicate that NER fine-tuning is subject to threshold effects whereby there are diminishing marginal returns from increased sample sizes. Our data revealed that a scant 439 sentences were sufficient to reach that threshold with RoBERTa_large. While smaller datasets may not be as helpful for SOTA-chasing, these data indicate that they may be sufficient for efficient development of production-line models. These findings are consistent with the growing multidisciplinary body of literature demonstrating the efficacy of smaller sample sizes for fine-tuning [13,23,24]. Additionally, we note that given prior estimates for NER annotation rates, a sample of approximately 450 sentences would take between 74 and 225 minutes to annotate [8].

Importantly, the data provided here also indicate that neither model size nor content-area specific foundational training data may be essential for maximizing performance, but that model architecture is. RoBERTa_base, GatorTron, and RoBERTa_large all achieved comparable performance levels in terms of max F1 with similarly low training sample sizes. GPT-2, despite being the largest model tested, showed the worst performance on our NER tasks. On the one hand, neither finding is surprising. Devlin et al.'s foundational paper on the BERT transformer architecture suggests that BERT's capacity for fine-tuning for NLP tasks like classification is better compared to GPT-based models' [16]. And a recent Microsoft Research paper argues that general-language models like GPT-4 can perform as well or better on domain-specific language tasks – specifically as they relate to medicine – than models trained on language specific to that domain [45]. But where the latter study focused on a very large language model built with RLHF and designed to be responsive to prompting, we found that for smaller – and therefore more tunable – models, fine-tuning with domain-specific texts yields significant performance improvements. For domain-specific NER tasks, then, architecture differences may matter most: decoder-based unidirectional architectures may be better suited for sentence generation, while encoder/decoder-based bi-directional architectures better capture sentence-level contexts that are essential to NER tasks.

The results presented here also indicate that there are similar threshold effects for token density. That is, selecting or synthetically creating specifically token-rich samples may not improve model performance. Unlike the sample size data that indicate a diminishing marginal return, the hinge model for token density shows a substantial decrease in overall performance after the EPS threshold is achieved. We note that these threshold point estimates and narrow 95% CIs converge on the average EPS (1.34) of the 2500 training sets, and this suggests that the relevant entity density of training data needs to approximate the relevant entity density of testing and production-line data.

This finding is especially relevant given the increasing interest in artificial training data generated by large language models. While the insights presented here indicate that fine-tuning training data can be much smaller than generally anticipated, high-quality small training data sets still require adequate funding and time to pay, train, and deploy human annotators. In response, some research seeks to leverage LLMs as sources of training data for subsequent fine-tuning of smaller neural network models [46]. This is an intriguing line of research worthy of further scrutiny. However, it is notable that our findings about relevant token density suggest that artificially generated data must mirror real data in terms of token density. If token density is too low or too high, we can expect to

see reduced model performance when compared to naturally derived training data and high-quality expert annotation.

While these findings provide an important initial foundation for fine-tuning sample size considerations in NER applications, the specifically identified thresholds may not apply to markedly different NER use cases. This study focused on fine-tuning of PERSON and ORG tags, entity types that are well-represented across the heterogeneous data sources that are used to train LLMs. Bioinformatics use cases that focus on entity types that are more unique to biomedical contexts (e.g., symptoms, chemicals, diseases, genes, proteins, etc.) or that require generating new entity categories may require larger training samples to optimize LLM performance. Additionally, this study focuses on semi-structured natural language (disclosure statements). While we would expect similar guidelines to apply for NER in other semi-structured biomedical contexts (e.g., research articles, clinical notes, abstracts, figure or image annotations, etc.), the threshold guidance here may not apply well to less formalized linguistic contexts.

Conclusion

The emergence of LLMs offers significant potential for improving NLP applications in biomedical informatics, with research demonstrating the advantages of fine-tuned, domain-specific language models for healthcare applications [47] and environmental cost [22]. However, given the novelty of these solutions, there is a general dearth of actionable guidelines on how to efficiently fine-tune language models. In the context of NER applications, this study demonstrates that there is a general lack of consensus and actionable guidance on sample size selection concerns for fine-tuning LLMs. Training sets reporting units and sample size varied widely in the published literature, with samples ranging from N=100 sentences to N = 35,938 sentences for training sets. Additionally, human-annotated training set sample sizes are seldom justified or explained. In the rare cases where sample size is discussed explicitly, justifications focus narrowly on simple size comparisons to previously published efforts in a similar domain. In this context, biomedical informatics researchers could benefit from actionable guidelines about sample size considerations for fine-tuning LLMs.

The data presented here provide sample size guidance for fine-tuning LLMs drawn from an experiment on 2500 gold-standard human annotated fine-tuning samples. Specifically, the data demonstrate the importance of both sample size as measured in number of sentences and relevant token density for training data curation. Furthermore, the findings indicate that both sample size and token density can be subject to threshold limitations where increased sample size or token density do not confer additional performance benefits. In the current study, sample sizes of greater than 439-527 sentences failed to produce meaningful accuracy improvements. This suggests that researchers interested in levering LLMs for NER applications can save considerable time, effort, and funding, which has been historically devoted to producing gold-standard annotations. The data presented here also indicate that relevant token density of training samples should reliably approximate the relevant token density of real-world cases. This finding has important ramifications for the production of synthetic data which may or may not effectively approximate real-world cases. The findings presented here can directly inform future research in health policy informatics and may also be applicable for a wider range of health and biomedical informatics tasks.

Acknowledgements

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM141476. The funder did not participate in study design, conduct, or preparation of findings.

This work used resources of the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University, which were made possible in part by NSF MRI Award No. 2019077.

Authors' Contributions

SSG and ZPM designed the study. ZPM implemented the fine-tuning pipelines. MSK and JTJ provided ground-truth annotations. JSE and SNR conducted the review of prior findings. SSG conducted the statistical analyses. All authors participated in interpretation of findings, drafting, and revision.

Conflicts of Interest

SSG reports grant funding from NIGMS and the Texas Health and Human Services Commission. ZPM reports grant funding from NIGMS, NSF, and the Summer Institute in Computational Social Science. SNR reports grant funding from National Institute of Neurological Disorders and Stroke. JBB reports grant funding from NIGMS, NSF, and Blue Cross Blue Shield/Health Care Service Corporation. JRF reports grant funding from NIGMS, NIMH, NIAID, NLM, Health Care Cost Institute, Austin Public Health, Texas Child Mental Health Care Consortium, Texas Alzheimer's Research and Care Consortium, the Michael & Susan Dell Foundation that includes: funding grants. JFR also reports receiving a grant from the NIH Division of Loan Repayment that includes. All other authors report no conflicts of interest.

Multimedia Appendix 1: Review of Sample Sizes and Justifications

Multimedia Appendix 2: Detailed Statistical Results and Threshold Model Plots

References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011 Sep 1;18(5):544–551. doi: 10.1136/amiajnl-2011-000464

2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001;17–21. PMID:11825149

3. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. J Am Med Inform Assoc JAMIA 2020 Jan 1;27(1):65–72. PMID:31504605

4. Ahmed A, Abbasi A, Eickhoff C. Benchmarking Modern Named Entity Recognition Techniques for Free-text Health Record Deidentification. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2021;2021:102–111. PMID:34457124

5. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc JAMIA 2010;17(5):514–518. PMID:20819854

6. Alfattni G, Belousov M, Peek N, Nenadic G. Extracting Drug Names and Associated

Attributes From Discharge Summaries: Text Mining Study. JMIR Med Inform 2021 May 5;9(5):e24678. PMID:33949962

7. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, Savova G. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc JAMIA 2015 Jan;22(1):143–154. PMID:25147248

8. Chen Y, Lask TA, Mei Q, Chen Q, Moon S, Wang J, Nguyen K, Dawodu T, Cohen T, Denny JC, Xu H. An active learning-enabled annotation system for clinical named entity recognition. BMC Med Inform Decis Mak 2017 Jul 5;17(2):82. doi: 10.1186/s12911-017-0466-9

9. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. J Am Med Inform Assoc 2022 Oct 1;29(10):1810–1817. doi: 10.1093/jamia/ocac121

10. Idnay B, Dreisbach C, Weng C, Schnall R. A systematic review on natural language processing systems for eligibility prescreening in clinical research. J Am Med Inform Assoc 2022 Jan 1;29(1):197–206. doi: 10.1093/jamia/ocab228

11. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017 Sep 1;73:14–29. doi: 10.1016/j.jbi.2017.07.012

12. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760–772. PMID:19683066

13. Manjavacas Arevalo E, Fonteyn L. Non-Parametric Word Sense Disambiguation for Historical Languages. Proc 2nd Int Workshop Nat Lang Process Digit Humanit Taipei, Taiwan: Association for Computational Linguistics; 2022. p. 123–134. Available from: https://aclanthology.org/2022.nlp4dh-1.16 [accessed May 31, 2023]

14. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can Large Language Models Transform Computational Social Science? arXiv; 2023. doi: 10.48550/arXiv.2305.03514

15. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health Public Library of Science; 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198

16. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. doi: 10.48550/arXiv.1810.04805

17. Liu X, Hersch GL, Khalil I, Devarakonda M. Clinical Trial Information Extraction with BERT. 2021 IEEE 9th Int Conf Healthc Inform ICHI 2021. p. 505–506. doi: 10.1109/ICHI52183.2021.00092

18. Graham SS, Majdik ZP, Clark D, Kessler MM, Hooker TB. Relationships among commercial practices and author conflicts of interest in biomedical publishing. PLOS ONE Public

Library of Science; 2020 Jul 24;15(7):e0236166. doi: 10.1371/journal.pone.0236166

19. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, Leahy C. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv; 2020. Available from: http://arxiv.org/abs/2101.00027 [accessed Dec 30, 2022]

20. Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SARP, Sayle R, Kors JA, Muresan S. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLOS ONE Public Library of Science; 2014 Sep 30;9(9):e107477. doi: 10.1371/journal.pone.0107477

21. Ciosici MR, Derczynski L. Training a T5 Using Lab-sized Resources. arXiv; 2022. doi: 10.48550/arXiv.2208.12097

22. Luccioni AS, Viguier S, Ligozat A-L. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv; 2022. doi: 10.48550/arXiv.2211.02001

23. Widner K, Virmani S, Krause J, Nayar J, Tiwari R, Pedersen ER, Jeji D, Hammel N, Matias Y, Corrado GS, Liu Y, Peng L, Webster DR. Lessons learned from translating AI from development to deployment in healthcare. Nat Med Nature Publishing Group; 2023 May 29;1–3. doi: 10.1038/s41591-023-02293-9

24. Majdik ZP, Wynn J. Building Better Machine Learning Models for Rhetorical Analyses: The Use of Rhetorical Feature Sets for Training Artificial Neural Network Models. Tech Commun Q Routledge; 2022 May 13;0(0):1–16. doi: 10.1080/10572252.2022.2077452

25. Weber L, Münchmeyer J, Rocktäschel T, Habibi M, Leser U. HUNER: improving biomedical NER with pretraining. Bioinformatics 2020 Jan 1;36(1):295–302. doi: 10.1093/bioinformatics/btz528

26. Doan S, Xu H. Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine. Proc COLING Int Conf Comput Linguist 2010 Aug;2010:259–266. PMID:26848286

27. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, Xu H. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc JAMIA 2018 Mar 1;25(3):331–336. PMID:29186491

28. Furrer L, Jancso A, Colic N, Rinaldi F. OGER++: hybrid multi-type entity recognition. J Cheminformatics 2019 Jan 21;11(1):7. PMID:30666476

29. Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. Patterns 2021 Jul 9;2(7):100289. doi: 10.1016/j.patter.2021.100289

30. Chun H-W, Tsuruoka Y, Kim J-D, Shiba R, Nagata N, Hishiki T, Tsujii J. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC Bioinformatics 2006 Nov 24;7 Suppl 3(Suppl 3):S4. PMID:17134477

31. Lundh A, Lexchin J, Mintzes B, Schroll J, Bero L. Industry sponsorship and research

outcome. Cochrane Database Syst Rev John Wiley & Sons, Ltd; 2017;(2). doi: 10.1002/14651858.MR000033.pub3

32. Graham SS, Karnes MS, Jensen JT, Sharma N, Barbour JB, Majdik ZP, Rousseau JF. Evidence for stratified conflicts of interest policies in research contexts: a methodological review. BMJ Open British Medical Journal Publishing Group; 2022 Sep 1;12(9):e063501. PMID:36123074

33. Grundy Q, Dunn AG, Bourgeois FT, Coiera E, Bero L. Prevalence of Disclosed Conflicts of Interest in Biomedical Research and Associations With Journal Impact Factors and Altmetric Scores. JAMA 2018 Jan 23;319(4):408–409. doi: 10.1001/jama.2017.20738

34. Lieb K, Osten-Sacken J von der, Stoffers-Winterling J, Reiss N, Barth J. Conflicts of interest and spin in reviews of psychological therapies: a systematic review. BMJ Open British Medical Journal Publishing Group; 2016 Apr 1;6(4):e010606. PMID:27118287

35. Roddick AJ, Chan FTS, Stefaniak JD, Zheng SL. Discontinuation and non-publication of clinical trials in cardiovascular medicine. Int J Cardiol 2017 Oct 1;244:309–315. PMID:28622947

36. van Lent M, Overbeke J, Out HJ. Role of Editorial and Peer Review Processes in Publication Bias: Analysis of Drug Trials Submitted to Eight Medical Journals. PLoS ONE 2014 Aug 12;9(8):e104846. PMID:25118182

37. Graham SS, Majdik ZP, Barbour JB, Rousseau JF. Associations Between Aggregate NLP-Extracted Conflicts of Interest and Adverse Events by Drug Product. Stud Health Technol Inform 2022 Jun 6;290:405–409. PMID:35673045

38. Grundy Q, Dunn AG, Bero L. Improving researchers' conflict of interest declarations. BMJ British Medical Journal Publishing Group; 2020 Mar 11;368:m422. PMID:32161006

39. Sunakawa Y, Satake H, Ichikawa W. Considering FOLFOXIRI plus bevacizumab for metastatic colorectal cancer with left-sided tumors. World J Gastrointest Oncol 2018 Dec 15;10(12):528–531. PMID:30595807

40. Graham SS, Majdik ZP, Clark D. Methods for Extracting Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research. J Open Humanit Data Ubiquity Press; 2020 Nov 19;6(1):8. doi: 10.5334/johd.21

41. Silge J, Robinson D. tidytext: Text mining and analysis using tidy data principles in R. J Open Source Softw 2016;1(3):37.

42. Hastie TJ. Generalized additive models. Stat Models S Routledge; 2017. p. 249–307.

43. Fong Y, Huang Y, Gilbert PB, Permar SR. chngpt: threshold regression model estimation and inference. BMC Bioinformatics 2017 Oct 16;18(1):454. doi: 10.1186/s12859-017-1863-x

44. Ullah MI, Aslam M, Altaf S, Ahmed M. Some New Diagnostics of Multicollinearity in Linear Regression Model. Sains Malays 2019 Sep 30;48(9):2051–2060. doi: 10.17576/jsm-2019-4809-26

45. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, King N, Larson J, Li Y, Liu W, Luo R, McKinney SM, Ness RO, Poon H, Qin T, Usuyama N, White C, Horvitz E. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. 2023 Nov 28; Available from: https://www.microsoft.com/en-us/research/publication/can-generalist-foundation-models-outcompete-special-purpose-tuning-case-study-in-medicine/ [accessed Dec 2, 2023]

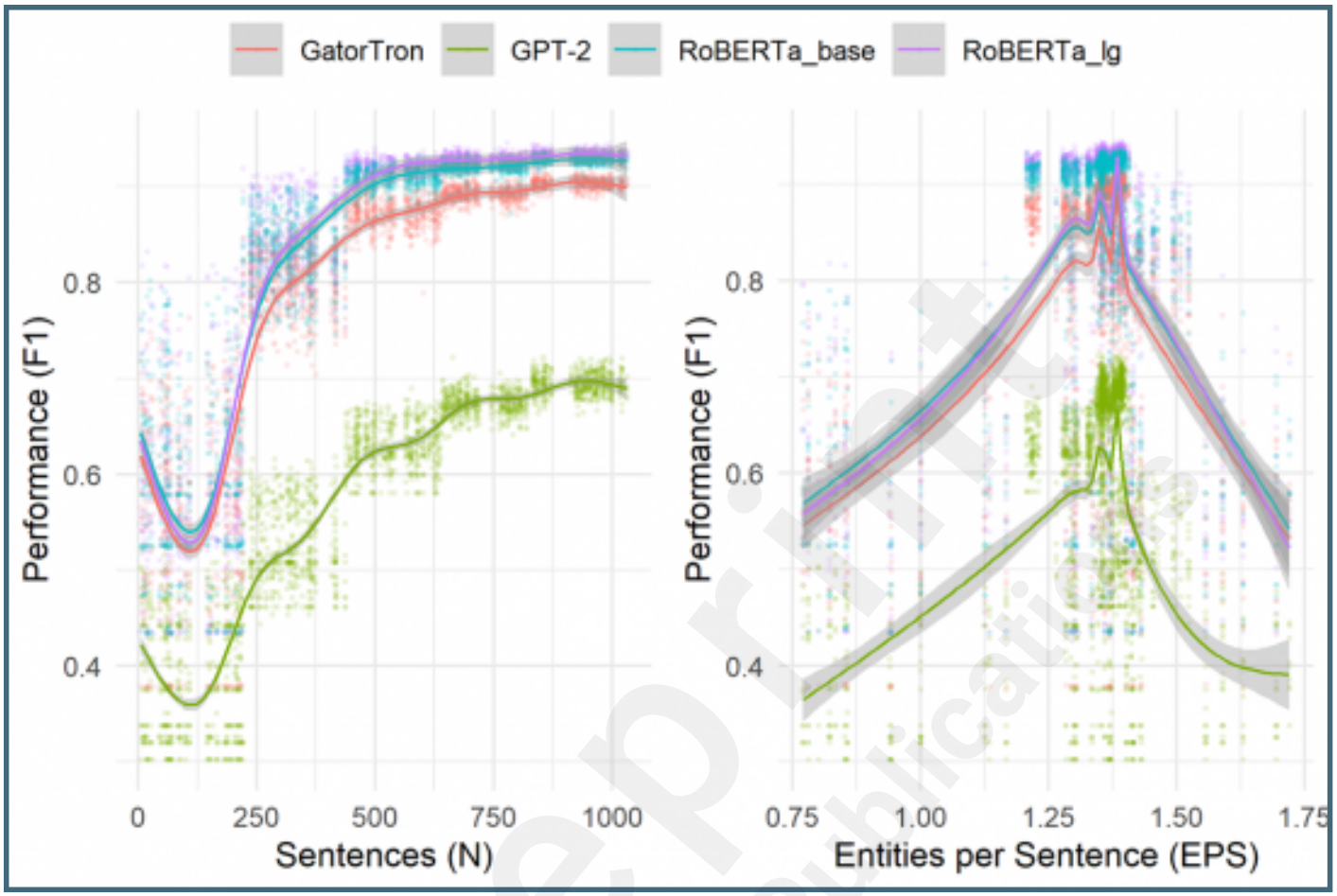
46. Hsieh C-Y, Li C-L, Yeh C-K, Nakhost H, Fujii Y, Ratner A, Krishna R, Lee C-Y, Pfister T. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. arXiv; 2023. doi: 10.48550/arXiv.2305.02301

47. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Sci 2023;2(4):255–263. doi: 10.1002/hcs2.61

Supplementary Files

Figures

Single predictor plots for N of sentence (left) and EPS (right). Fit with a generalized additive model.



Multimedia Appendixes

Review of Sample Sizes and Justifications.
URL: http://asset.jmir.pub/assets/1059a42d735b82f9d143edf0b7a842f0.docx

Detailed Statistical Results and Threshold Model Plots.
URL: http://asset.jmir.pub/assets/e53d59e3cab2c740ae3940cd95e23c2b.docx