# Leveraging Large Language Models for Analyzing Blood Pressure Variations Across Biological Sex from Scientific Literature

Yuting Guo, Seyedeh Somayyeh Mousavi, Reza Sameni, and Abeed Sarker

Emory University, Atlanta GA 30322, USA
{yuting.guo, seyedeh.somayyeh.mousavi}@emory.edu
{rsameni, abeed}@dbmi.emory.edu

**Abstract.** Hypertension, defined as blood pressure (BP) that is above normal, holds paramount significance in the realm of public health, as it serves as a critical precursor to various cardiovascular diseases (CVDs) and significantly contributes to elevated mortality rates worldwide. However, many existing BP measurement technologies and standards might be biased because they do not consider clinical outcomes, comorbidities, or demographic factors, making them inconclusive for diagnostic purposes. There is limited data-driven research focused on studying the variance in BP measurements across these variables. In this work, we employed GPT-35-turbo, a large language model (LLM), to automatically extract the mean and standard deviation values of BP for both males and females from a dataset comprising 25 million abstracts sourced from PubMed. 993 article abstracts met our predefined inclusion criteria (*i.e.*, presence of references to *blood pressure*, *units of blood pressure such as mmHg*, and mention of biological sex). Based on the automatically-extracted information from these articles, we conducted an analysis of the variations of BP values across biological sex. Our results showed the viability of utilizing LLMs to study the BP variations across different demographic factors.

**Keywords:** Blood Pressure · Natural Language Processing · Large Language Model.

## 1 Introduction

Hypertension is defined as blood pressure (BP) that is above normal. It is of paramount significance for public health, as it serves as a critical precursor to various cardiovascular diseases (CVDs) and significantly contributes to elevated mortality rates worldwide [1]. Its pervasive nature is underscored by the staggering estimate of 1.28 billion individuals living with hypertension, a condition often untreated in 700 million cases as of 2019 [2]. High and unstable BP levels are even more critical for vulnerable populations, including the elderly, pregnant women, and patients with comorbid diseases [1]. BP monitoring tools are widely accessible and BP measurement involves obtaining just two numeric values: systolic (SBP) and diastolic (DBP) [3]. Despite its accessibility and importance, BP readings exhibit notable biases concerning individual conditions and nonstandard measurement setups [4]. Many existing BP technologies and standards were established decades ago, among predominantly White and genetically homogeneous populations from the Global North. These clinical standards persist today and are globally applied without adjusting BP readings for individual factors such as biological sex, age, race, height, weight, medical history, or comorbidities like diabetes. Meanwhile, mounting evidence highlights the influence of race and genetics on BP [5,6,7]. The conventional ranges designating normal and abnormal BP (e.g., $< 120/80$ mmHg for normal) merely indicate a patient's percentile within the stochastically distributed BP of a population [3]. While these ranges offer guidelines and aid in identifying at-risk patients, they do not consider clinical outcomes, comorbidities, or demographic factors (*e.g.,* biological sex), making them inconclusive for diagnostic purposes.

In contemporary times, the scientific literature pertaining to BP is abundant, with thousands of research publications accessible online. These studies collectively present compiled BP data derived from the vital signs of hundreds of millions of individuals, encompassing diverse comorbid and demographic factors within various cohorts. As of now, PubMed alone records 687,401 instances of the keyword 'blood pressure' in its repository of published research.[1] It's important to note that not all of these instances contain quantitative data. Nevertheless, in clinical cohorts, it is customary to include details such as the number of subjects, inclusion/exclusion criteria, averages, and standard deviations of BPs (as well as other vital signs). Our hypothesis posits that the amalgamation of these reports could serve to deduce population-wide distributions of BP across different demographics and individualized factors. Analyzing the published literature on a case-by-case basis would be laborious or infeasible, lack generalizability, and be prone to human error. However, recent advancements in large language models (LLMs) and natural language processing (NLP) offer the opportunity to systematically search and parse the text within these publications. This allows for the retrieval of population-wide BP statistics, including the number of subjects, mean and standard deviations of systolic and diastolic BP, as well as information on comorbid factors and demographics. This addresses the imperative need for heterogeneous BP datasets essential for making clinical decisions based on blood pressure.

In this research, we conducted an investigation into the viability of employing NLP for the automated extraction of information from scientific literature related to BP. Our primary focus was on utilizing a LLM to extract the mean and standard deviation of BP values with a biological sex-based distinction, as variations in patterns across different biological sexes were anticipated. We also wanted to assess how well a zero-shot LLM performs at automatically extracting BP measurements and sex from texts, and linking them. Our results indicate that leveraging an LLM enables the analysis of a substantial volume of scientific literature to extract BP-related information based on various demographic

[1] Accessed on January 29, 2024.

factors. The performance of the LLM in zero-shot setting, however, can be improved. To the best of our knowledge, this is the first study to employ LLMs for extracting BP-related information from large-scale biomedical literature data.

## 2 Related Work

In 2023, inspired by the emergence and success of LLMs, numerous studies have leveraged them for NLP tasks related to biomedical and healthcare topics. Kung et al. [8] evaluated a LLM named ChatGPT on the United States Medical Licensing Exam and demonstrated that the model showed a high level of concordance and insight in its explanations. Chen at al. [9] examined the performance of ChatGPT on various neurological exam grading scales, where ChatGPT demonstrated ability in evaluating neuroexams using established assessment scales. Similarly, Dehghani et al. [10] evaluated the performance of ChatGPT on a radiology board-style examination, and ChatGPT correctly answered 69% of questions. These findings indicate that LLMs might possess the capability to support medical education and potentially assist clinical decision-making. Huang et al. [11] applied ChatGPT to automate diagnosis of dental conditions and demonstrated the model's ability to handle multi-source data and apply sophisticated natural language reasoning for the execution of intricate clinical operations. In the study conducted by Russe et al. [12], GPT-3.5-turbo and GPT-4 were employed to recognize Arbeitsgemeinschaft Osteosynthesefragen (AO) codes in radiology reports. The outcomes indicated that the models extracted AO codes at a significantly faster rate than humans, with commendable accuracy. These studies collectively showcase the potential of employing LLMs to advance medical research. In our work, we employ GPT-3.5-turbo to extract selected variables from medical literature abstracts.

## 3 Methods

### 3.1 Data Collection

The dataset was obtained from PubMed Central (PMC), a freely accessible archive of biomedical and life sciences journal literature located at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) through the PMC File Transfer Protocol (FTP) Service.[2] The information was stored in XML format, with each publication organized in a structured manner, encompassing the title, author, DOI, and plain text of the abstract. We isolated the plain text abstracts and refined the content using the keyword '*blood pressure*' to identify studies related to BP. Subsequently, we further refined the dataset by searching for instances containing the keywords '*mmHg*' or '*mm Hg*' to specifically target studies that reported blood pressure values. Only article abstracts, not full texts, were included in this study.

[2] The data were downloaded from https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/.

```
You are an AI assistant that helps people find information in scientific publications.

Read the abstract carefully:

{{abstract}}

Please answer the mean and deviation of blood pressure values based on sex.
N is the number of patients, and summaries of SBP and DBP include mean ± standard deviation.
Your answer should be formatted as
'''
N_male=[number]
N_female=[number]
SBP_in_male=[mean ± standard deviation]
SBP_in_female=[mean ± standard deviation]
DBP_in_male=[mean ± standard deviation]
DBP_in_female=[mean ± standard deviation]
'''
Answer "NA" if the information is not found.
```

**Fig. 1.** The prompt used for utilizing the LLM to extract the mean and standard deviation based on biological sex, where 'abstract' was the placeholder for the abstract content.

### 3.2 Information Extraction

We employed the GPT-35-turbo LLM for information extraction to derive the mean and standard deviation based on biological sex. The model's input, referred to as the '*prompt*', consisted of a plain text description instructing the model to identify values for ten variables: the number of males and females, the mean and standard deviation of SBP for males and females, and the mean and standard deviation of DBP for males and females. The prompt included explicit instructions for organizing the answer in a specific format to facilitate post-generation parsing and retrieval of variable values. The prompt was refined via trial and error. The final prompt details can be found in Figure 1. This experimental setup falls under the category of zero-shot learning, a machine learning approach that does not rely on training data. Consequently, we could analyze abstracts automatically without the need for manually annotating the data. Following the generation of answers for each article, a simple parsing script using regular expressions was executed to extract the values for each variable. These variables represented the BP levels for males and females within a particular cohort examined in the respective literature.

To evaluate the model performance, we manually reviewed the cases where the abstracts contained all the variables, and the model predicted values for all the variables. We also reviewed the cases where the abstract did not contain all the variables, but the model predicted all the values.

### 3.3 BP Variation Analysis

Different approaches can be employed to analyze BP variation using gathered data, demonstrating how NLP and LLM enhance BP studies. In a recent research study on bias in BP measurement technologies, authors collected information by manually investigating just 20 papers and analyzing BP data related to biological sex [4]. They represented the distribution of BP information using ellipses, with each ellipse associated with a specific study. The centers are linked to the mean of DBP and SBP, while the horizontal and vertical radii correspond to the standard deviations of SBP and DBP, respectively. In the following analysis, researchers used heatmaps to visualize distributions. They employed Gaussian mixture models (GMM) to explore the data further, assuming that the reported BP values follow a normal distribution. GMMs are a statistical method used for clustering and density estimation. In this study, we employed this approach for representing SBP and DBP associated with males and females, automatically extracted from the abstracts.

## 4 Results

We downloaded about 25 million abstracts from Pubmed. The number of abstracts that contained the keyword 'blood pressure' was 321,226. Among those, 71,067 abstracts contained the keyword 'mmHg' or 'mm Hg'. After running the model and parsing the model output, the number of answers that contained the values for all ten variables (i.e., the number of males and females, the mean and standard deviation of SBP for males and females, and the mean and standard deviation of DBP for males and females) was 993.

Out of the 993 cases, 44 cases were identified in which the model successfully predicted all variables, and these predicted values were found in the abstracts. Upon manual examination of these 44 cases, it was noted that 7 cases featured abstracts containing all the variables, and the model accurately predicted all of them. Thus, the model had 100% accuracy for the cases where all the variables were present in the abstract, but fluctuating accuracy when the abstracts did not contain all the variables. We provide further details about these fluctuations in the discussion section.

Following the automatic extraction, a post-processing procedure was conducted to ensure the reliability and accuracy of our findings. Initially, we examined the mean BP values to confirm they fell within a reasonable range ($DBP\_min = 30$, $DBP\_max = 120$, $SBP\_min = 60$, $SBP\_max = 200$ mmHg). Subsequently, a comprehensive dataset assessment was undertaken to identify and address any missing values. Our data analysis revealed a wide range in the number of BP records, ranging from a minimum of 2 to a maximum of approximately 10 million. Therefore, we focused on studies (582) with a total number of BP records exceeding one hundred for more reliability.

Figure 2 illustrates heatmaps and contour plots associated with BP data for each gender, facilitating a comparison of BP results between males and females.

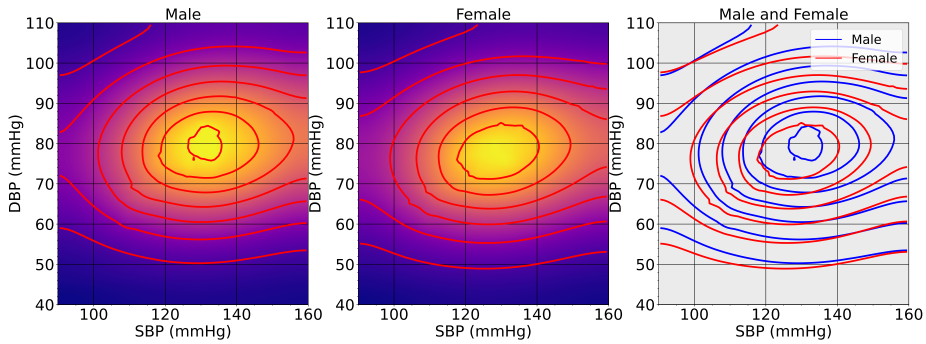


**Fig. 2.** Comparisons of blood pressure distributions between sexes are presented through heatmaps and contour plots, employing Gaussian mixture models for heatmap visualization

GMM is employed for heatmap visualization. As depicted in the figure, according to the peak points on the heatmaps, we can conclude that males tend to exhibit higher BP values than females, aligning with consistent findings from numerous previous studies and our recent work reporting elevated BP levels in males compared to females [13,14,4]. Contour plots for both males and females are presented in a single figure. The contour lines on these plots highlight areas where BP values are more concentrated. Notably, it is interesting to observe that, apart from the variations between males and females, their distributions are distinctly individual, and certain regions exhibit overlaps.

## 5 Discussion

Our findings indicate that when the abstracts included all the required variables, the LLM successfully predicted all cases accurately. This implies the potential to develop a highly accurate information extraction system for analyzing BP values from scientific literature. Importantly, this excellent performance was achieved in a zero-shot setting, demonstrating the superiority of LLMs compared to more traditional low-shot information extraction systems [17].

As mentioned earlier, we also observed that in some cases the model predicted all variables even when the abstracts did not mention those variables. For instance, an abstract detailed the number of males and females but reported BP values based on the mixture of males and females. In such cases, the LLM appeared to assign the BP values randomly for males and females. This suggests that the model has a tendency to generate values for all requested variables, even if these values are not explicitly mentioned. Such errors may perhaps be mitigated in the future by refining the prompts and explicitly instructing the model to ignore such cases. However, this will require substantial trial and error and perhaps more advanced prompting strategies such as chain-of-thought prompting, which we leave as future work.

***Abstract:***
*A health survey of 551 high school students in the coastal town of Busselton, Western Australia, was undertaken in November, 1977. Physical measurements showed that approximately 17% of the boys and 24% of the girls were overweight (115% or more of their standard weight for height and age). Blood pressures ranged from 108/64 mmHg at 13 years to 129/70 mmHg at 17 years for boys, and 109/63 mmHg to 116/66 mmHg for girls of the same age group. Fasting plasma glucose, cholesterol and triglyceride values were also obtained. Twenty-seven per cent of the girls and 17% of the boys were regular smokers, a habit often acquired by the age of 12 years. Most of the children watched television from two to four hours daily during the week, and for longer periods at weekends. Although most had satisfactory levels of physical fitness, there was a trend towards poor performance in overweight children, and there was a significant decrease in participation in regular sports by the older teenagers.*

***LLM'answer:***
*N_male=551*
*N_female=551*
*SBP_in_male=118.5 ± 7.5 mmHg*
*SBP_in_female=112.5 ± 3.5 mmHg*
*DBP_in_male=67 ± 3 mmHg*
*DBP_in_female=64.5 ± 1.5 mmHg*

**Fig. 3.** An example where the LLM averaged the BP values by information within the abstract from [15]. The mean SBP in males was predicted as 118.5 mmHg, which could be computed by averaging the SBP of 13-year-old boys (108 mmHg) and 17-year-old boys (129 mmHg). The same operation was performed for the mean SBP of females, the mean DBP of males, and the mean DBP of females.

An interesting finding was that the model often performed its own calculations based on values present in the abstract, and then presented these calculated values. For example, we observed that when some abstracts contained multiple blood pressure (BP) values for males or females, the model automatically computed the average, as illustrated in Figure 3. This demonstrates the LLM's ability to extrapolate new information from the raw abstract data, which can be useful in some settings. Another possible explanation for these derived values is that they might originate from the full texts of the articles corresponding to the abstracts. Since the pretraining data for GPT-3.5-turbo consisted of all publicly accessible internet data [18], it is possible that the model was pretrained on open-access scientific literature with full-text availability. Consequently, the LLM has the potential to provide information not explicitly present in the abstracts.

## 6 Limitation

One drawback of this study lies in its inadequate assessment of the LLM's responses. The evaluation process demanded substantial human involvement to

***Abstract:***
*Prevalence of hypertension was investigated in Mauritius in 2362 men and 2712 women among Hindu and Muslim Indian, Creole and Chinese ethnic groups aged 25-74 years. The age-standardized prevalence of hypertension varied from 9.4% to 17.3% in men and from 9.5% to 16.9% in women among the four ethnic groups and increased with age. The population mean values of systolic/diastolic blood pressure were 126/79 mmHg in men and 124/75 mmHg in women. The proportion of hypertensive persons aware of their condition was 50.5% in men and 66.6% in women. The proportion of treated and adequately controlled hypertensive patients was only 21.7% in men and 29.3% in women. Creoles had the highest mean value of systolic and diastolic blood pressure and the highest prevalence of hypertension whilst Muslim Asian Indians had the lowest values both in men and women. Further studies are needed to define the determinants of hypertension in the multi-ethnic Mauritian population. Diagnosis and treatment of hypertension need to be improved, especially in men and the Creole ethnic group.*

***LLM'answer:***
*N_male=2362*
*N_female=2712*
*SBP_in_male=126 ± 3 mmHg*
*SBP_in_female=124 ± 3 mmHg*
*DBP_in_male=79 ± 4 mmHg*
*DBP_in_female=75 ± 4 mmHg*

**Fig. 4.** An example where the LLM produced an incorrect answer that was not supported by information within the abstract from [16]. The standard deviations appearing in the LLM's answer (3 mmHg and 4 mmHg) did not appear in the abstract.

scrutinize abstracts and verify the accuracy of the model's answers. Through an examination of various instances, it became apparent that the LLM is susceptible to generating 'hallucinations' in its responses. In simpler terms, the model may produce incorrect answers that are not supported by information within the abstracts. For instance, in Figure 4, the abstract did not specify standard deviations, yet the LLM provided a standard deviation in its response. Hallucinations are now a well-known problem for LLMs, and addressing this issue is an area of active research. Due to resource constraints, a comprehensive manual review of all answers was infeasible. Future research should address the development of effective methods for evaluating the accuracy of LLM responses.

Another constraint of this study is its exclusive concentration on a single demographic factor, neglecting other pertinent factors such as age, race, height, and weight. The experiments were limited to just one factor due to the high cost associated with conducting experiments involving LLMs. Despite this and other limitations, the findings from this study are promising, and we leave the extension to other variables as future work.

## 7 Conclusion

In this research, we investigated the possibility of applying an LLM in a zero-shot setting to extract the mean and standard deviation of BP values with a biological sex-based distinction. Furthermore, we conducted variation analysis on the extracted information. The results suggest that males might have higher BP levels than females. Importantly, this study shows that it is possible to leverage LLMs to perform the analysis of a substantial volume of scientific literature to extract BP-related information based on various demographic factors. Future work will expand this study to investigate the BP variations based on other demographic factors using the scientific literature.

## References


1. Whelton, P.K., Carey, R.M., Aronow, W.S., Casey, D.E., Collins, K.J., Dennison Himmelfarb, C., DePalma, S.M., Gidding, S., Jamerson, K.A., Jones, D.W., MacLaughlin, E.J., Muntner, P., Ovbiagele, B., Smith, S.C., Spencer, C.C., Stafford, R.S., Taler, S.J., Thomas, R.J., Williams, K.A., Williamson, J.D., Wright, J.T.: 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation **138**(17), e484–e594 (2018). https://doi.org/10.1161/CIR.0000000000000596
2. Collaboration, N.R.F.: Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. The Lancet **398**(10304), 957–980 (2021). https://doi.org/10.1016/S0140-6736(21)01330-1
3. James, P.A., Oparil, S., Carter, B.L.: 2014 evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the eighth joint national committee (JNC 8). Journal of Cardiopulmonary Rehabilitation and Prevention **34**(3), 218–219 (2014). https://doi.org/10.1097/HCR.0000000000000071
4. Mousavi, S.S., Reyna, M.A., Clifford, G.D., Sameni, R.: A Survey on Blood Pressure Measurement Technologies: Addressing Potential Sources of Bias (2023). https://doi.org/10.48550/ARXIV.2306.08451
5. Zilbermint, M., Hannah-Shmouni, F., Stratakis, C.A.: Genetics of hypertension in African Americans and others of African descent. International Journal of Molecular Sciences **20**(5) (2019). https://doi.org/10.3390/ijms20051081
6. Fuchs, F.D.: Why do black Americans have higher prevalence of hypertension?: An enigma still unsolved. Hypertension **57**(3), 379–380 (2011). https://doi.org/10.1161/HYPERTENSIONAHA.110.163196
7. Anum Saeed, D.L.D.E.Y.: Racial Disparities in Hypertension Prevalence and Management: A Crisis Control. American College of Cardiology **18**, 2020 (2020)
8. Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, Victor Tseng: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Plos Digital Health **2**(2) (2023)

9. Chen, T.C., Kaminski, E., Koduri, L., Singer, A., Singer, J., Couldwell, M., Delashaw, J., Dumont, A., Wang, A.: Chat GPT as a Neuro-Score Calculator: Analysis of a Large Language Model's Performance on Various Neurological Exam Grading Scales. World Neurosurgery **179**, e342–e347 (2023). https://doi.org/10.1016/j.wneu.2023.08.088
10. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., van Steenkiste, S., Elsayed, G.F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M.P., Gritsenko, A.A., Birodkar, V., Vasconcelos, C., Tay, Y., Mensink, T., Kolesnikov, A., Pavetić, F., Tran, D., Kipf, T., Lučić, M., Zhai, X., Keysers, D., Harmsen, J., Houlsby, N.: Scaling Vision Transformers to 22 Billion Parameters. Proceedings of Machine Learning Research **202**, 7480–7512 (2023). https://doi.org/10.48550/arXiv.2302.05442
11. Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., Yin, H., Xu, C., Yang, R., Zheng, Q., Shi, B.: ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. International Journal of Oral Science **15**(1) (2023). https://doi.org/10.1038/s41368-023-00239-y
12. Russe, M.F., Fink, A., Ngo, H., Tran, H., Bamberg, F., Reisert, M., Rau, A.: Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Scientific Reports **13**(1) (2023). https://doi.org/10.1038/s41598-023-41512-8
13. Sandberg, K., Ji, H.: Sex differences in primary hypertension. Biology of Sex Differences **3**(1), 7 (2012). https://doi.org/10.1186/2042-6410-3-7
14. Reckelhoff, J.F.: Gender differences in the regulation of blood pressure. Hypertension **37**(5), 1199–1208 (May 2001). https://doi.org/10.1161/01.HYP.37.5.1199
15. Gracey, M., Hitchcock, N.E., Wearne, K.L., Garcia-Webb, P., Lewis, R.: The 1977 Busselton children's survey. The Medical journal of Australia **2**(5), 265–267 (1979). https://doi.org/10.5694/j.1326-5377.1979.tb127134.x
16. Nan, L., Tuomilehto, J., Dowse, G., Zimmet, P., Gareeboo, H., Chitson, P., Korhonen, H.J., Alberti, K.G.M., Fareed, D.: Prevalence and medical care of hypertension in four ethnic groups in the newly-industrialized nation of Mauritius. Journal of Hypertension **9**(9), 859–866 (1991). https://doi.org/10.1097/00004872-199109000-00014
17. Ge, Y., Guo, Y., Das, S., Al-Garadi, M.A., Sarker, A.: Few-shot learning for medical text: A review of advances, trends, and opportunities. Journal of Biomedical Informatics **144**, 104458 (2023). https://doi.org/https://doi.org/10.1016/j.jbi.2023.104458
18. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. Advances in Neural Information Processing Systems **2020-Decem**(NeurIPS) (2020)