Large Language Models Lack Understanding of Character Composition of Words

Andrew Shin [1] **Kunitake Kaneko** [1]

Abstract

Large language models (LLMs) have demonstrated remarkable performances on a wide range of natural language tasks. Yet, LLMs' successes have been largely restricted to tasks concerning words, sentences, or documents, and it remains questionable how much they understand the minimal units of text, namely characters. In this paper, we examine contemporary LLMs regarding their ability to understand character composition of words, and show that most of them fail to reliably carry out even the simple tasks that can be handled by humans with perfection. We analyze their behaviors with comparison to token level performances, and discuss the potential directions for future research.

1. Introduction

Large language models (LLMs) (Achiam et al., 2023; Chowdhery et al., 2022; Touvron et al., 2023; Reid et al., 2024; OpenAI, 2022; Jiang et al., 2023) have exhibited outstanding performance across a diverse array of natural language tasks. It has largely outperformed pre-LLM approaches on benchmark tasks, such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), often surpassing humans on a number of tasks (Chowdhery et al., 2022). It is noteworthy that most of the tasks upon which LLMs have been tested revolve around words, sentences, or passages, but hardly involve character-level understanding. Intuitively, character-level tasks should be much easier to tackle, as they rarely deal with complex semantics, grammatical structures, or background knowledge, while only requiring highly elementary understanding of characters and, depending on the task, simple counting. Indeed, humans are able to perform basic character-level tasks very easily as we will see in Sec 3.2. It has also been known that LLMs hardly make spelling errors and can be used for spelling correction of human-written passages (Whittaker & Kitagishi, 2024). Surprisingly, however, our examination shows that LLMs struggle with very simple tasks involving character composition, severely underperforming humans, making a striking contrast with their performance on more complex tasks at token level.

Humans are able to instantly recognize which characters constitute a given word. However, large language models, most of which are trained at token-level, struggle to grasp the nuances of character composition within words. This difficulty arises from the fact that LLMs primarily learn at the token level, where words are treated as indivisible units separated by spaces or punctuation marks. Consequently, LLMs lack the fine-grained understanding of character-level relationships and morphology that humans possess. Understanding character composition is crucial for various linguistic tasks, including morphological analysis, semantic interpretation, and language generation. As such, addressing the challenge of character composition is essential for enhancing the reliability of LLMs across a diverse range of languages and writing systems.

In this paper, we examine LLMs with a number of simple tasks designed to test the understanding of character composition. None of the tasks requires any advanced knowledge of grammar or semantics, and can be easily tackled with elementary understanding of characters. Yet, our results show a surprisingly poor performance, suggesting that there may be a fundamental drawback with regards to how LLMs are trained and how they perceive the language. We compare LLMs' performances at character level tasks with those at token level tasks of the same types, and investigate the implications of the large discrepancies. We further discuss potential future research directions to enhance LLMs' understanding of character composition, such as incorporating character embedding and visual features into language representation of LLMs.

2. Related Works

Although a majority of language models have relied on token-level embeddings, there have been a number of notable endeavors to incorporate character composition or sub-

[1]Faculty of Science and Technology, Keio University, Kanagawa, Japan. Correspondence to: Andrew Shin <shin@inl.ics.keio.ac.jp>.

word tokenization into language models, some of which have demonstrated improved performance on relevant tasks. (Kim et al., 2015) introduced character-aware neural language models, which utilize character-level embeddings alongside word embeddings to capture morphological and orthographic features of words. Similarly, (Wieting et al., 2016) proposed Charagram, a character-level language model that generates word representations based on character n-grams, enabling better handling of out-of-vocabulary words. (Bojanowski et al., 2016) presented FastText, a fast and efficient word embedding technique that leverages sub-word information to enhance word representations, particularly for morphologically rich languages. While these approaches demonstrate the effectiveness of integrating character information into language models, paving the way for improved performance in various natural language processing tasks, they have mostly been tested on natural language generation tasks, such as Penn Treebank (Marcus et al., 1993), and have not explicitly been tested for understanding of character composition.

Subsequent works in language modeling have further explored the integration of character-level information. For instance, (Peters et al., 2018) introduced deep contextualized word representations (ELMo), which enhance word embeddings by considering the internal structure of words through character-level convolutions. This method significantly improved the performance of various NLP tasks by capturing complex word morphologies. (Akbik et al., 2019) proposed Flair embeddings, which combine character-level embeddings with contextual string embeddings to provide a more comprehensive representation of words in their context. (Clark et al., 2020) introduced ELECTRA, a pre-training method that includes a discriminative component to identify corruptions at the token level, which indirectly benefits from finer-grained text representations. For most of these works, however, the primary focus has been on token-level tasks rather than specifically addressing character composition understanding.

With regards to more recent LLMs, there have been a number of works that highlight their downsides from various angles. For example, (Qian et al., 2022) claims that LLMs struggle with arithmetic and symbolic manipulations, while (Lee & Lim, 2024) shows that LLMs fail to learn physical manifestation of language, such as the visuals and sounds of the language. (Truong et al., 2023) also shows that LLMs' performances degrade when negation is involved. With regards to the character composition, there have been a few attempts to benchmark the performances of LLMs (Srivastava et al., 2022; Efrat et al., 2022), although the scope of evaluating character composition was highly restricted, with stronger emphasis on evaluation of word-level understanding.

3. Experiments

3.1. Setting

We perform simple tasks that are designed to assess the LLM's understanding of character composition of words. Nearly all tasks are simple and straightforward with hardly any component for complexity or confusion. It would be fair to state that even humans with very little educational background of up to elementary school can solve most of these tasks without difficulty.

Word retrieval: We provide the LLM with input text and ask it to retrieve all words containing a certain character. For example, "Find all words that contain the character *h* in the following text: *She is home.*" should output "*She*" and "*home*". The task may be examined in variations by specifying the position or the number of occurrences of the characters within a word.

Character insertion / deletion / replacement: We ask LLM to insert a character to words in the input text at a specified position, or delete a specified character or any character at a specified position from the input text, or replace a character with another character. For example, "Insert the character *a* to the beginning of all words in the following text: *I am well*" should output "*aI aam awell*," and similarly for deletion and replacement.

Character reordering: We provide the LLM with words and ask it to reorder the characters within each word to form a new word, in a similar manner to anagram, e.g., generate "*epics*" from the input word "*spice*." The output is deemed correct if it contains all characters in the input word with the same number of occurrences. Note that there is no restriction as to whether new word should be an existing word, as long as all characters have been used.

Character counting: We provide the LLM with input text and ask it to count the number of certain characters or a category of characters, such as vowels and consonants. For example, "How many occurrences of the character *s* are in the following word: *obsessed*?" should return 3.

We experimented with 4 publicly available LLMs, namely GPT4 (Achiam et al., 2023), Claude (Claude, 2023), Gemini 1.5 (Reid et al., 2024), and Mistral 7B (Jiang et al., 2023). We randomly sampled words, phrases, or sentences from Wikipedia corpus. Note that, while it is possible that such publicly available text was used during the pre-training of target LLMs, the character-based nature of our experiments prevents the models from taking advantage of it, and the results in Sec 3.2 seem to reinforce the claim. For each task, 100 prompts were used, where each prompt may contain multiple answers. In order to compare the LLM's understanding of character composition with that of humans, we also asked human annotators to perform exactly the same

Table 1. Precision, recall, and F-score for each model on evaluation tasks at character level. For reordering and counting, accuracy is reported in precision column.

Task	Human			GPT4			Claude			Gemini			Mistral		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Word Retrieval	**1.0**	**.989**	**.994**	.523	.691	.595	.406	.534	.461	.549	.602	.574	.614	.671	.641
Character Insertion	**1.0**	**1.0**	**1.0**	.286	.514	.368	.214	.357	.268	.203	.414	.272	.429	.443	.436
Character Deletion	**1.0**	**1.0**	**1.0**	.236	.336	.277	.372	.439	.403	.270	.342	.302	.353	.362	.357
Character Replacement	**1.0**	**.943**	**.971**	.725	.453	.558	.815	.435	.567	.823	.725	.771	.488	.328	.392
Character Reordering	**1.0**	–	–	.91	–	–	.93	–	–	.92	–	–	.88	–	–
Character Counting	**.98**	–	–	.59	–	–	.51	–	–	.63	–	–	.60	–	–

Table 2. F-score for each model on evaluation tasks at token level. For reordering and counting, accuracy is reported.

Task	Human	GPT4	Claude	Gemini	Mistral
Sentence Retrieval	**1.0**	.926	.893	.921	.953
Word Insertion	**1.0**	.625	.643	.701	.632
Word Deletion	**1.0**	.578	.542	.602	.529
Word Replacement	**1.0**	.991	.994	.993	.981
Word Reordering	**.99**	.95	.97	.97	.96
Word Counting	**1.0**	.98	.93	.97	.91

tasks, providing identical prompts and passages.

In order to compare LLMs' performances at character level and token level tasks, we also extend each task described above to token level tasks. Word retrieval is extended to sentence retrieval, where the model is given 5-sentence passage and is asked to return all sentences containing a target word. Insertion and deletion work similarly by providing target word and position within sentence, whereas we provide target word and another input word for replacement task. Reordering and counting are extended similarly. For reordering, as with character-level reordering, we compute accuracy from whether the final answer is correct, without computing precision and recall for each reordered word.

3.2. Results

Table 1 summarizes the results of our experiments with precision, recall, and F-score for each task at character level. For token level, we only report F-score for brevity in Table 2. It is clearly shown that, for most tasks, all target LLMs display severely degraded performance at character level when compared to token level. While discrepancies exist among respective models' performances, none rises to the level of demonstrating a clear superiority over other models. It is also out of scope of this paper to determine which LLM is better, as our focus is on assessing LLMs in terms of understanding character composition in general.

Humans, not surprisingly, demonstrated near-perfect performance throughout all tasks. There was hardly any mistake in precision, while defects in recall tended to occur mostly around characters that are placed in the middle of the word, rather than beginning or the end, suggesting attention to saliency in human perception of character composition. Considering that humans have been surpassed by LLMs in many NLP tasks that are supposedly more complex, our results suggest an unsettling dichotomy between LLM's capability at token-level and character-level tasks.

Table 3 shows some of the failure cases for each model at character level. It is notable that the tasks for which LLMs struggled the most frequently involved specifying positions of the characters, mostly using numbers, as in insertion or deletion tasks. It should be noted that a similar performance decline was observed even at token level, as illustrated in Table 2. Table 4 shows example failure cases at token level. This suggests that some of the limitations in understanding character composition may not simply be attributed to the fact that LLMs are trained at token level, but to a more fundamental drawback in their training approach. Notably, all LLMs performed far better on character reordering task than on other tasks, closely trailing the performance of humans. We conjecture that this is due to abundant resources available online about anagram, which are likely to have been used in pre-training of the models. Even when the newly formed words are non-existing words, many of them are likely to have appeared in the training corpora as possible anagrams of an existing word . It is therefore only natural that all models struggled with character reordering as the word gets longer, or with an unknown word, as shown in Table 4.

A clearer contrast between LLMs' performances on token level and character level tasks is made on the tasks that do not involve numerical elements, such as replacement. As illustrated in an example in Table 5, LLMs rarely have any trouble with replacement task at token level, indicating that token-based embeddings are functioning in a desired manner. Word reordering task also turned out to be reliable, even for fairly long sentences. Such clear contrast between LLMs' performances on token level and character level tasks highlights a fundamental discrepancy in how these models process linguistic information, which suggests that, while LLMs have been effectively optimized for tasks involving tokens, their handling of finer-grained character-level tasks remains inadequate.

4. Discussion

As shown throughout the paper, much of limitation in terms of understanding character composition derives from the very nature of LLMs where they are almost invariably trained at token levels, regardless of the pre-training ob-

Table 3. Example failure cases at character level tasks. Bold letters indicate correct answers by the model.

Prompt	GPT4	Claude	Gemini	Mistral
Find all words with character *o*: *People enjoy music.* (*answer: People, enjoy*)	People, enjoy, music	People, enjoy, music	People, enjoy, music	**People, enjoy**
Insert character *i* after the first character for all words: *The Great Wall stretches far.* (*answer: Tihe Gireat Wiall sitretches fiar.*)	Tihie Giireat Wiiall strietches fari.	iThe iGreat iWall istretches ifar.	Thie Great Wall stretches far.	Tihe Griat Wiall striatches fiar.
Remove the third character from the end for all words: *Gravity affects falling objects.* (*answer: Gravty affets fallng objets.*)	Gravie affect fallin object.	Grav affects fallin objec.	Gravit affect fallin object.	Gravty affecs faling obects.
Replace all occurrences of *h* with *x*: *He has three children.* (*answer: Xe xas txree cxildren.*)	xe xas three cxildren.	He xas three children.	**Xe xas txree cxildren**.	Ex has three children
Reorder the characters in the following word to form a new word: *supercalifragilistic* (*answer: any valid anagram apart from input word itself*)	upercalifra-gilistic	supercalifr-agilistic	lapsticalifr-agiceorous	cilisuparegalf-itisticxedocious
How many vowels are in the following word: *supercalifragilistic* (*answer: 8*)	9	11	**8**	7

Table 4. Failure cases at token level tasks. Note that they frequently involve numerical elements such as position. Bold letters indicate the correct answer by the model.

Model \ Prompt	Remove the third word from the following sentence: *The Renaissance was a period of cultural and artistic rebirth.* (*ans: The Renaissance a period of cultural and artistic rebirth.*	What is the seventh word from the end: *The Great Pyramid of Giza is one of the Seven Wonders.* (*ans: Giza*)
GPT4	*The was a period of cultural and artistic rebirth.*	*one*
Claude	*The Renaissance was a cultural and artistic rebirth.*	*Wonders*
Gemini	*The Renaissance was a period of artistic rebirth.*	*of*
Mistral	***The Renaissance a period of cultural and artistic rebirth.***	*Pyramid*

Table 5. Example of LLMs' performances at token level in tasks that do not involve numerical elements. Bold letters indicate the correct answer by the model.

Model \ Prompt	Replace all occurrences of "the" with "X": *The history of the city is influenced by the river.* (*ans: X history of X city is influenced by X river.*)
GPT4	***X history of X city is influenced by X river.***
Claude	***X history of X city is influenced by X river.***
Gemini	***X history of X city is influenced by X river.***
Mistral	***X history of X city is influenced by X river.***

jectives. By operating primarily at the token level, LLMs overlook the intrinsic characteristics and nuances of individual characters within words. This oversight hinders their ability to capture the rich semantic and syntactic information encoded at the character level, leading to sub-optimal performance in tasks requiring fine-grained understanding of language structure.

A promising direction to address this limitation involves embedding character-level information directly into word embeddings, enabling models to capture the intricate relationships and structures within individual characters. For example, BERT (Devlin et al., 2019) represents input tokens not only with token embedding, but also with segment embedding, which indicates the sentence that the token belongs to, and position embedding, which shows the position of the token within the sentence. A similar structural approach can be made with respect to character, where character is embedded also with information of the word it belongs to, and its position within the word. Such multi-level embedding strategy could significantly enhance the model's ability to understand and manipulate text at a finer granularity, and can help ensure that the model obtains a robust understanding of word composition while being sensitive to the arrangement of characters within words. Another potential line of approach involves harnessing visual recognition techniques to simulate human-like character perception. In scene text recognition literature, there has been a number of endeavors to integrate computer vision methodologies to visually identify characters, replicating the cognitive processes humans employ when reading and comprehending text (Du et al., 2022; Bartz et al., 2017). By leveraging the complementary strengths of both domains, these approaches may potentially offer novel opportunities for improving robustness for character-level comprehension within large language models.

5. Conclusion

We examined LLMs' ability to understand character composition of words. Our experiments suggest that LLMs utterly fail to demonstrate the ability to understand character composition even at highly simple tasks that can be easily solved by humans with elementary knowledge of language, making a stark contrast with their performances at token level. We further discussed potential future directions, such as incorporating character-embedding and visual features.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J. R., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A. A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D. P., Mu, T., Murati, M., Murk, O., M'ely, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Long, O., O'Keefe, C., Pachocki, J. W., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Pokorny, M., Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M. D., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B. D., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N. A., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C. L., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report. 2023.

Akbik, A., Bergmann, T., Blythe, D. A. J., Rasul, K., Schweter, S., and Vollgraf, R. Flair: An easy-to-use framework for state-of-the-art nlp. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Bartz, C., Yang, H., and Meinel, C. See: Towards semi-supervised end-to-end scene text recognition. In *AAAI Conference on Artificial Intelligence*, 2017.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2016.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N. M., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B. C., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., García, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Díaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K. S., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2022.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020.

Claude. Claude.ai. https://claude.ai/, 2023. [Accessed 17-05-2024].

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., and Jiang, Y.-G. Svtr: Scene text recognition with a single visual model. In *International Joint Conference on Artificial Intelligence*, 2022.

Efrat, A., Honovich, O., and Levy, O. Lmentry: A language model benchmark of elementary language tasks. *ArXiv*, abs/2211.02069, 2022. URL https://api.semanticscholar.org/CorpusID:253370569.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

Kim, Y., Jernite, Y., Sontag, D. A., and Rush, A. M. Character-aware neural language models. In *AAAI Conference on Artificial Intelligence*, 2015.

Lee, B. W. and Lim, J. Language models don't learn the physical manifestation of language. 2024. URL https://api.semanticscholar.org/CorpusID:267750661.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19:313–330, 1993.

OpenAI. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt, 2022.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *ArXiv*, abs/1802.05365, 2018.

Qian, J., Wang, H., Li, Z., LI, S., and Yan, X. Limitations of language models in arithmetic and symbolic induction. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:251467816.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J.-B., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A. M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K. W., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., Vilnis, L., Chang, O., Morioka, N., Tucker, G., Zheng, C., Woodman, O., Attaluri, N., Kocisky, T., Eltyshev, E., Chen, X., Chung, T., Selo, V., Brahma, S., Georgiev, P., Slone, A., Zhu, Z., Lottes, J., Qiao, S., Caine, B., Riedel, S., Tomala, A., Chadwick, M., Love, J. C., Choy, P., Mittal, S., Houlsby, N., Tang, Y., Lamm, M., Bai, L., Zhang, Q., He, L., Cheng, Y., Humphreys, P., Li, Y., Brin, S., Cassirer, A., Miao, Y.-Q., Zilka, L., Tobin, T., Xu, K., Proleev, L., Sohn, D., Magni, A., Hendricks, L. A., Gao, I., Ontan'on, S., Bunyan, O., Byrd, N., Sharma, A., Zhang, B., Pinto, M., Sinha, R., Mehta, H., Jia, D., Caelles, S., Webson, A., Morris, A., Roelofs, B., Ding, Y., Strudel, R., Xiong, X., Ritter, M., Dehghani, M., Chaabouni, R., Karmarkar, A., Lai, G., Mentzer, F., Xu, B., Li, Y., Zhang, Y., Paine, T. L., Goldin, A., Neyshabur, B., Baumli, K., Levskaya, A., Laskin, M., Jia, W., Rae, J. W., Xiao, K., He, A., Giordano, S., Yagati, L., Lespiau, J.-B., Natsev, P., Ganapathy, S., Liu, F., Martins, D., Chen, N., Xu, Y., Barnes, M., May, R., Vezer, A., Oh, J., Franko, K., Bridgers, S., Zhao, R., Wu, B., Mustafa, B., Sechrist, S., Parisotto, E., Pillai, T. S., Larkin, C., Gu, C., Sorokin, C., Krikun, M., Guseynov, A., Landon, J., Datta, R., Pritzel, A., Thacker, P., Yang, F., Hui, K., Hauth, A., Yeh, C.-K., Barker, D., Mao-Jones, J., Austin, S., Sheahan, H., Schuh, P., Svensson, J., Jain, R., Ramasesh, V. V., Briukhov, A., Chung, D.-W., von Glehn, T., Butterfield, C., Jhakra, P., Wiethoff, M., Frye, J., Grimstad, J., Changpinyo, B., Lan, C. L., Bortsova, A., Wu, Y., Voigtlaender, P., Sainath, T. N., Smith, C., Hawkins, W., Cao, K., Besley, J., Srinivasan, S., Omernick, M., Gaffney, C., de Castro Surita, G., Burnell, R., Damoc, B., Ahn, J., Brock, A., Pajarskas, M., Petrushkina, A., Noury, S., Blanco, L., Swersky, K., Ahuja, A., Avrahami, T., Misra, V., de Liedekerke, R., Iinuma, M., Polozov, A., York, S., van den Driessche, G., Michel, P., Chiu, J., Blevins, R., Gleicher, Z., Recasens, A., Rrustemi, A., Gribovskaya, E., Roy, A., Gworek, W., Arnold, S. M. R., Lee, L., Lee-Thorp, J., Maggioni, M., Piqueras, E., Badola, K., Vikram, S., Gonzalez, L., Baddepudi, A., Senter, E., Devlin, J., Qin, J., Azzam, M., Trebacz, M., Polacek, M., Krishnakumar, K., yiin Chang, S., Tung, M., Penchev, I., Joshi, R., Olszewska, K., Muir, C., Wirth, M., Hartman, A. J., Newlan, J., Kashem, S., Bolina, V., Dabir, E., van Amersfoort, J. R., Ahmed, Z., Cobon-Kerr, J., Kamath, A. B., Hrafnkelsson, A. M., Hou, L., Mackinnon, I., Frechette, A., Noland, E., Si, X., Taropa, E., Li, D., Crone, P., Gulati, A., Cevey, S., Adler, J., Ma, A., Silver, D., Tokumine, S., Powell, R., Lee, S., Chang, M. B., Hassan, S., Mincu, D., Yang, A., Levine, N., Brennan, J., Wang, M., Hodkinson, S., Zhao, J., Lipschultz, J., Pope, A., Chang, M. B., Li, C., Shafey, L. E.,

Paganini, M., Douglas, S., Bohnet, B., Pardo, F., Odoom, S., Rosca, M., dos Santos, C. N., Soparkar, K., Guez, A., Hudson, T., Hansen, S., Asawaroengchai, C., Addanki, R., Yu, T., Stokowiec, W., Khan, M., Gilmer, J., Lee, J., Bostock, C. G., Rong, K., Caton, J., Pejman, P., Pavetic, F., Brown, G., Sharma, V., Luvci'c, M., Samuel, R., Djolonga, J., Mandhane, A., Sjosund, L. L., Buchatskaya, E., White, E., Clay, N., Jiang, J., Lim, H., Hemsley, R., Labanowski, J., Cao, N. D., Steiner, D., Hashemi, S. H., Austin, J., Gergely, A., Blyth, T., Stanton, J., Shivakumar, K., Siddhant, A., Andreassen, A., Araya, C. L., Sethi, N., Shivanna, R., Hand, S., Bapna, A., Khodaei, A., Miech, A., Tanzer, G., Swing, A., Thakoor, S., Pan, Z., Nado, Z., Winkler, S., Yu, D., Saleh, M., Maggiore, L., Barr, I., Giang, M., Kagohara, T., Danihelka, I., Marathe, A., Feinberg, V., Elhawaty, M., Ghelani, N., Horgan, D., Miller, H., Walker, L., Tanburn, R., Tariq, M., Shrivastava, D., Xia, F., Chiu, C.-C., Ashwood, Z. C., Baatarsukh, K., Samangooei, S., Alcober, F., Stjerngren, A., Komarek, P., Tsihlas, K., Boral, A., Comanescu, R., Chen, J., Liu, R., Bloxwich, D., Chen, C., Sun, Y., Feng, F., Mauger, M., Dotiwalla, X., Hellendoorn, V., Sharman, M., Zheng, I., Haridasan, K., Barth-Maron, G., Swanson, C., Rogozi'nska, D., Andreev, A., Rubenstein, P. K., Sang, R., Hurt, D., Elsayed, G., Wang, R., Lacey, D., Ili'c, A., Zhao, Y., Aroyo, L., Iwuanyanwu, C., Nikolaev, V., Lakshminarayanan, B., Jazayeri, S., Kaufman, R. L., Varadarajan, M., Tekur, C., Fritz, D., Khalman, M., Reitter, D., Dasgupta, K., Sarcar, S., Ornduff, T., Snaider, J., Huot, F., Jia, J., Kemp, R., Trdin, N., Vijayakumar, A., Kim, L., Angermueller, C., Lao, L., Liu, T., Zhang, H., Engel, D., Greene, S., White, A., Austin, J., Taylor, L., Ashraf, S., Liu, D., Georgaki, M., Cai, I., Kulizhskaya, Y., Goenka, S., Saeta, B., Vodrahalli, K., Frank, C., de Cesare, D., Robenek, B., Richardson, H., Alnahlawi, M., Yew, C., Ponnapalli, P., Tagliasacchi, M., Korchemniy, A., Kim, Y., Li, D., Rosgen, B., Levin, K., Wiesner, J., Banzal, P., Srinivasan, P., Yu, H., cCauglar Unlu, Reid, D., Tung, Z., Finchelstein, D. F., Kumar, R., Elisseeff, A., Huang, J., Zhang, M., Zhu, R., Aguilar, R., Gim'enez, M., Xia, J., Dousse, O., Gierke, W., Yeganeh, S. H., Yates, D., Jalan, K., Li, L., Latorre-Chimoto, E., Nguyen, D. D., Durden, K., Kallakuri, P., Liu, Y., Johnson, M., Tsai, T., Talbert, A., Liu, J., Neitz, A., Elkind, C., Selvi, M., Jasarevic, M., Soares, L. B., Cui, A., Wang, P., Wang, A. W., Ye, X., Kallarackal, K., Loher, L., Lam, H., Broder, J., Holtmann-Rice, D. N., Martin, N., Ramadhana, B., Toyama, D., Shukla, M., Basu, S., Mohan, A., Fernando, N., Fiedel, N., Paterson, K., Li, H., Garg, A., Park, J., Choi, D., Wu, D., Singh, S., Zhang, Z., Globerson, A., Yu, L., Carpenter, J., de Chaumont Quitry, F., Radebaugh, C., Lin, C.-C., Tudor, A., Shroff, P., Garmon, D., Du, D., Vats, N., Lu, H., Iqbal, S., Yakubovich, A., Tripuraneni, N., Manyika, J., Qureshi, H., Hua, N., Ngani, C., Raad, M. A., Forbes, H., Bulanova, A., Stanway, J., Sundararajan, M., Ungureanu, V., Bishop, C., Li, Y., Venkatraman, B., Li, B., Thornton, C., Scellato, S., Gupta, N., Wang, Y., Tenney, I., Wu, X., Shenoy, A., Carvajal, G., Wright, D. G., Bariach, B., Xiao, Z., Hawkins, P., Dalmia, S., Farabet, C., Valenzuela, P., Yuan, Q., Welty, C. A., Agarwal, A., Chen, M., Kim, W., Hulse, B., Dukkipati, N., Paszke, A., Bolt, A., Davoodi, E., Choo, K., Beattie, J., Prendki, J., Vashisht, H., Santamaria-Fernandez, R., Cobo, L. C., Wilkiewicz, J., Madras, D., Elqursh, A., Uy, G., Ramirez, K., Harvey, M., Liechty, T., Zen, H., Seibert, J., Hu, C. H., Khorlin, A. Y., Le, M., Aharoni, A., Li, M., Wang, L., Kumar, S., Lince, A., Casagrande, N., Hoover, J., Badawy, D. E., Soergel, D., Vnukov, D., Miecnikowski, M., Simsa, J., Koop, A., Kumar, P., Sellam, T., Vlasic, D., Daruki, S., Shabat, N., Zhang, J., Su, G., Zhang, J., Liu, J., Sun, Y., Palmer, E., Ghaffarkhah, A., Xiong, X., Cotruta, V., Fink, M., Dixon, L., Sreevatsa, A., Goedeckemeyer, A., Dimitriev, A., Jafari, M., Crocker, R., Fitzgerald, N. A., Kumar, A., Ghemawat, S., Philips, I., Liu, F., Liang, Y., Sterneck, R., Repina, A., Wu, M., Knight, L., Georgiev, M., Lee, H., Askham, H., Chakladar, A., Louis, A., Crous, C., Cate, H., Petrova, D., Quinn, M., Owusu-Afriyie, D., Singhal, A., Wei, N., Kim, S., Vincent, D., Nasr, M., Choquette-Choo, C. A., Tojo, R., Lu, S., de Las Casas, D., Cheng, Y., Bolukbasi, T., Lee, K., Fatehi, S., Ananthanarayanan, R., Patel, M., Kaed, C. E., Li, J., Sygnowski, J., Belle, S. R., Chen, Z., Konzelmann, J., Poder, S., Garg, R., Koverkathu, V., Brown, A., Dyer, C., Liu, R., Nova, A., Xu, J., Petrov, S., Hassabis, D., Kavukcuoglu, K., Dean, J., and Vinyals, O. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A. A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmuller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakacs, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Ozyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B. S., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ram'irez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C., Siro, C., Raffel, C., Ashcraft,

C., Garbacea, C., Sileo, D., Garrette, D. H., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., Gonz'alez, D. M., Perszyk, D. R., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E. P., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Mart'inez-Plumed, F., Happ'e, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-L'opez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schutze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Koco'n, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J. N., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J. O., Xu, J., Song, J., Tang, J., Waweru, J. W., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernández-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Col'on, L. O., Metz, L., cSenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M., Arnaud, M., McElrath, M. A., Yee, M., Cohen, M., Gu, M., Ivanitskiy, M. I., Starritt, M., Strube, M., Swkedrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., MukundVarma, T., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P.-B., Milkowski, P., Patil, P. S., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Milliere, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S. S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615, 2022. URL https://api.semanticscholar.org/CorpusID:263625818.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

Truong, T. H., Baldwin, T., Verspoor, K. M., and Cohn, T. Language models are not naysayers: an analysis of language models on negation benchmarks. *ArXiv*, abs/2306.08189, 2023. URL https:

//api.semanticscholar.org/CorpusID:259164714.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, 2018.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Superglue: A stickier benchmark for general-purpose language understanding systems. *ArXiv*, abs/1905.00537, 2019.

Whittaker, E. and Kitagishi, I. Large language models for simultaneous named entity extraction and spelling correction. *ArXiv*, abs/2403.00528, 2024.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. Charagram: Embedding words and sentences via character n-grams. In Su, J., Duh, K., and Carreras, X. (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1504–1515, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1157.