Matching domain experts by training from scratch on domain knowledge

Xiaoliang Luo [1]   Guangzhi Sun [2]   Bradley C. Love [1] [3]

Abstract

Recently, large language models (LLMs) have outperformed human experts in predicting the results of neuroscience experiments (Luo et al., 2024). What is the basis for this performance? One possibility is that statistical patterns in that specific scientific literature, as opposed to emergent reasoning abilities arising from broader training, underlie LLMs' performance. To evaluate this possibility, we trained (next word prediction) a relatively small 124M-parameter GPT-2 model on 1.3 billion tokens of domain-specific knowledge. Despite being orders of magnitude smaller than larger LLMs trained on trillions of tokens, small models achieved expert-level performance in predicting neuroscience results. Small models trained on the neuroscience literature succeeded when they were trained from scratch using a tokenizer specifically trained on neuroscience text or when the neuroscience literature was used to finetune a pretrained GPT-2. Our results indicate that expert-level performance may be attained by even small LLMs through domain-specific, auto-regressive training approaches.

1. Introduction

Large language models (LLMs) are statistical machines typically designed to predict the next token—whether it's a word, pixel, or protein sequence. Leveraging vast amounts of training data, LLMs have demonstrated impressive capabilities, including passing professional exams, reasoning (though with limitations), translation, solving mathematics problems, and writing computer code (Strack, 2023; Srivastava et al., 2022; Gunasekar et al., 2023).

Traditionally, the human-level performance of large language models (LLMs) has been evaluated using benchmarks that focus on their *backward-looking* capabilities, such as core knowledge retrieval and reasoning within a given context. Notable benchmarks include MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022). However, recent research by Luo et al. (2024) has highlighted LLMs' exceptional forward-looking capabilities, particularly in predicting novel outcomes of neuroscience studies. With the development of BrainBench, a *forward-looking* neuroscience benchmark, Luo et al. (2024) have shown that LLMs can outperform neuroscientists in predicting the results of neuroscientific experiments when provided with the experiment's background and methodologies. These findings raise important questions about the nature of scientific progress, suggesting that many discoveries might largely be iterations of noisy signals from decades of scientific literature. Additionally, they prompt a reevaluation of the extent to which accurate predictions of the future rely more on pattern recognition by auto-regressive models than on traditional scientific reasoning.

In this contribution, we explore the effects of training on domain-specific data by employing a significantly smaller language model, GPT-2 with 124 million parameters (Radford et al., 2019), on a neuroscience-focused dataset containing 1.3 billion tokens. This approach helps assess the effectiveness of auto-regressive training on specialized data in approximating human-level performance. Despite the model size being only about 0.056% to 1%[1] of those evaluated by Luo et al. (2024) and the training data being about 0.065%[2] of those used in Luo et al. (2024), we show both finetuning a pretrained 124M-parameter GPT-2 and training it from scratch with a custom tokenizer for neuroscience yield models that achieve 63.5% and 63% accuracy on BrainBench, matching the performance of human experts (63.4%). Larger GPT-2 variant (774M) pretrained on the same data yield even stronger BrainBench performance (surpassing human experts; Fig 1) with improved confidence calibration beneficial for human-model teaming (see Appendix B, D).

---

[1]Estimated using 7B and 180B LLMs.

[2]Estimated based on reported Llama-2 training data size (2 trillion tokens).

[1]Department of Experimental Psychology, University College London, UK [2]Department of Engineering, University of Cambridge, UK [3]The Alan Turing Institute, UK. Correspondence to: Xiaoliang Luo <xiao.luo.17@ucl.ac.uk>.

2. Method

2.1. BrainBench

BrainBench has curated 200 test cases from abstracts in the *Journal of Neuroscience published in 2023*. These abstracts are categorized into five sections: Behavioral/Cognitive, Systems/Circuits, Neurobiology of Disease, Development/Plasticity/Repair, and Cellular/Molecular.

Each test case includes a published abstract alongside a modified version crafted by neuroscientists. These modifications, though minimal, significantly alter the results—for instance, by changing the roles of brain regions or reversing a result's direction (e.g., from "decreases" to "increases"). Despite these changes, the altered abstracts remain logically coherent.

The test-taker's challenge is to identify the correct study outcome by choosing between the original abstract and its altered counterpart.

2.2. Model evaluation

We presented models with two versions of the abstracts from each test case separately. We prefixed each abstract with the prompt "*You are a neuroscientist with deep knowledge in neuroscience. Here is an abstract from a neuroscience publication:*". We then measured the perplexity of both passages and used perplexity as the indicator of whether models favor one abstract or the other.

Perplexity measures the degree of uncertainty of a model when generating a particular sequence of text and is defined as the exponentiated average negative log-likelihood of a tokenized sequence. If we have a tokenized abstract $X = (x_0, x_1, \ldots, x_t)$, then the perplexity of X, given a model parameterized by θ is,

$$PPL(X) = \exp\left\{ -\frac{1}{t} \sum_i^t \log p_\theta(x_i | x_{<i}) \right\} \quad (1)$$

where $\log p_\theta(x_i | x_{<i})$ is the log-likelihood of the ith token conditioned on the preceding tokens $x_{<i}$ according to the model. Given both the original and the altered abstracts, we used the abstract with lower perplexity as the model's decision and evaluated the overall accuracy across the entire BrainBench dataset accordingly.

2.3. Human evaluation

Previous work (Luo et al., 2024) collected human judgements from 171 neuroscience experts on BrainBench. These data are publicly available[3] and provide a useful comparison to LLM performance.

Table 1. **Model variants.**

Variant	Training	Data	Tokenizer
GPT2-Untrained	-	-	pretrained
GPT2-Pretrained	from scratch	WebText	pretrained
GPT2-Scratch	from scratch	neuroscience	pretrained
GPT2-Finetuned	finetune	neuroscience	pretrained
GPT2-Neuro	from scratch	neuroscience	custom

2.4. Model configurations

We considered a number of variants of GPT-2 differ by their training strategies including training data and tokenization. Model variants are summarized in Table 1.

The pretrained GPT-2 and the tokenizer were loaded from Huggingface hub[4], which were trained on the WebText dataset collected by OpenAI (Radford et al., 2019). The neuroscience training data was collected by Luo et al. (2024) (see Sec. 2.5). The models trained from scratch and finetuned used the neuroscience data only. The neuro-tokenizer employs GPT-2's tokenization strategy (Radford et al., 2019), adapted from Byte Pair Encoding (BPE) (Gage, 1994) for word segmentation (Sennrich et al., 2016). It's trained anew on neuroscience data used for model training, maintaining a vocabulary size of 50,257 tokens.
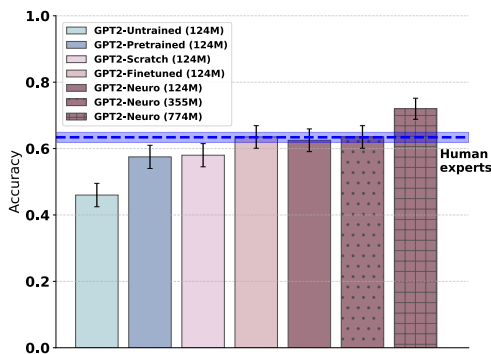
2.5. Neuroscience training data

The data we used to train GPT-2 from scratch, finetune the pretrained GPT-2 as well as train the neuro-tokenizer were collected by Luo et al. (2024). The training data spans Neuroscience publication (abstracts and full articles) dates 2002-2022, totaling 1.3 billion tokens. We randomly allocated 90% of the data for training, reserving the remaining 10% for validation. Training details see Appendix.

3. Results

We explored various training strategies and found that fine-tuning the pretrained GPT-2 on 20 years of neuroscience literature allowed it to achieve human-level performance on BrainBench, recording a 63.5% accuracy (Fig. 1; human experts: 63.4%). Training GPT-2 from scratch solely with neuroscience literature was less effective. However, developing a new tokenizer tailored to neuroscience literature and using it to retrain GPT-2 from scratch with the same data resulted in a performance on par with human experts, achieving 63% accuracy (Fig. 1). Notably, the amount of domain-specific data used to train GPT-2 from scratch is only about one-seventh of the text used to pretrain the original model. This indicates two effective approaches to reach human-level performance: pretraining on a broad general corpus followed by finetuning on domain-specific data, or using a specialized tokenizer and significantly less domain-specific data.

---

[3]https://github.com/braingpt-lovelab/BrainBench

[4]https://huggingface.co/openai-community/gpt2

Figure 1. **Performance of human experts and models on Brain-Bench.** Two configurations of GPT-2 models achieve human-level performance on BrainBench: one by fine-tuning the pretrained GPT-2 on neuroscience literature, and the other by using a new tokenizer (GPT2-Neuro) trained on neuroscience literature and retraining GPT-2 from scratch with only neuroscience data. Versions of GPT-2 that are untrained, pretrained, or trained solely on neuroscience data without these modifications underperform compared to experts on BrainBench. Larger GPT-2 (774M) trained using neuroscience literature surpassed human-level performance.

To assess the impact of a specialized tokenizer, we compared the tokens generated by the pretrained GPT-2 tokenizer with those from our neuro-tokenizer, trained on neuroscience data. The two tokenizers shared 47.9% of their vocabularies (Fig 2A). We utilized GPT-4 (zero-shot prompting) to analyze each vocabulary and identify tokens frequently associated with neuroscience. Our findings showed that the neuro-tokenizer contained twice the proportion of neuroscience-related tokens compared to the pretrained tokenizer (Fig. 2B-C). This significant improvement in specialized tokenization suggests that it is possible to pretrain GPT-2 from scratch with significantly less neuroscience data using the neuro-tokenizer, yet achieve performance comparable to both the finetuned model and human experts.

To better understand the differences in tokenization by the two tokenizers, we analyzed examples from BrainBench test cases where the pretrained GPT-2 answered incorrectly, whereas the GPT-2 trained with the neuro-tokenizer responded correctly. Figure 3 illustrates how the neuro-tokenizer more effectively preserves domain-specific terminologies, such as brain regions or neurotransmitters. We believe that this specialized tokenization allows the model to utilize limited domain knowledge more effectively and to consider a broader context within the fixed context window of the training data.

4. Discussion

In this contribution, we demonstrated that training a relatively small LLM (GPT-2) on limited domain-specific data
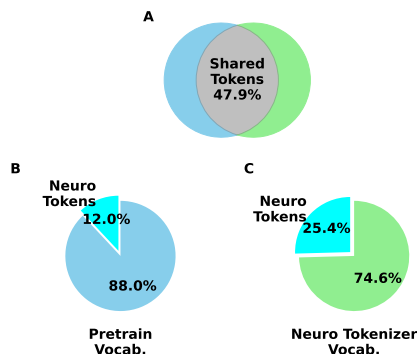
Figure 2. **Token Analysis.** (A) The pretrained GPT-2 tokenizer and the neuro-tokenizer share 47.9% of their vocabularies. (B-C) Of the vocabularies from the two tokenizers, 12.0% of the tokens from the pretrained tokenizer are commonly associated with neuroscience according to GPT-4, compared to 25.4% of the tokens from the neuro-tokenizer.

can match the predictive performance of human experts on BrainBench. By finetuning GPT-2 with just a fraction of its pretraining data, we elevated its performance to the level of trained neuroscientists. Additionally, we showed that training GPT-2 from scratch, with domain-specific knowledge incorporated into the tokenizer, yields comparable results. This highlights the importance of preserving domain-specific terminologies during tokenization to improve language models' performance on specialized tasks, as suggested by Yang et al. (2024) in the clinical science domain. Pretraining on small-scale knowledge with specialized tokenization offers a more efficient method for achieving human-like performance on domain-specific tasks. In addition, the resulting models show good calibration and ability to integrate information across context (Appendix B, C).

Achieving parity with and surpassing human experts using our simplified setup prompts questions about the essence of scientific progress. It suggests that a statistical machine, even one as basic as predicting the next word, can discern the intricate structure of a knowledge-rich field. Despite a significant performance gap (15%) between GPT-2 (124M) and more advanced LLMs on BrainBench tests, larger GPT-2 models—still much smaller than their counterparts—are narrowing this gap when pretrained with neuroscience data (Appendix D). We suspect that the performance is influenced by both model size and the quality and relevance of the training data (see Luo et al. 2024 for a discussion).

Working with smaller models does have benefits, such as enabling teams with modest resources to have full control over the training procedure. This control can minimize the risk of leakage and allow for additional hypotheses to be evaluated. For instance, in future work, we will evaluate whether training on adjacency fields like psychology

impacts performance on BrainBench, a neuroscience benchmark. That degree of control is not possible using pretrained LLMs and will allow us to evaluate the structure of scientific disciplines.

Software and Data

Human participant data, and intermediate data generated via simulations and analyses are publicly available at https://github.com/braingpt-lovelab/BrainBench. Model weights and training data are available at https://huggingface.co/BrainGPT.

All computer code associated with this work including model training, evaluation, data processing and analyses are publicly available at https://github.com/braingpt-lovelab/matching_experts.

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.

Acknowledgements

References

Gage, P. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994. URL https://api.semanticscholar.org/CorpusID:59804030.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks Are All You Need. 2023. doi: 10.48550/ARXIV.2306.11644. URL https://arxiv.org/abs/2306.11644. Publisher: arXiv Version Number: 1.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering, September 2019. URL http://arxiv.

org/abs/1909.06146. arXiv:1909.06146 [cs, q-bio].

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs, math].

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., Marinazzo, D., Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., Gu, N., Ferianc, M., Khona, M., Patil, K. R., Lee, P.-S., Mata, R., Myers, N. E., Bizley, J. K., Musslick, S., Bilgin, I. P., Niso, G., Ales, J. M., Gaebler, M., Murty, N. A. R., Loued-Khenissi, L., Behler, A., Hall, C. M., Dafflon, J., Bao, S. D., and Love, B. C. Large language models surpass human experts in predicting neuroscience results, March 2024. URL http://arxiv.org/abs/2403.03230. arXiv:2403.03230 [cs, q-bio].

Pal, A., Umapathi, L. K., and Sankarasubbu, M. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering, March 2022. URL http://arxiv.org/abs/2203.14371. arXiv:2203.14371 [cs].

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019.

Sennrich, R., Haddow, B., and Birch, A. Neural Machine Translation of Rare Words with Subword Units, June 2016. URL http://arxiv.org/abs/1508.07909. arXiv:1508.07909 [cs].

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., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., 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., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, 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. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman,

D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., 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., Pavlick, E., Rodola, 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., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, 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., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., 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., 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., Colón, L. O., Metz, L., Şenel, 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. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, 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., T, M. V., 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. S., 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., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., 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., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, 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., 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. 2022. doi: 10.48550/ARXIV.2206.04615. URL https://arxiv.org/abs/2206.04615. Publisher: arXiv Version Number: 3.

Strack, R. Visual proteomics. *Nature Methods*, 20 (12):1868–1868, December 2023. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-023-02104-6. URL https://www.nature.com/articles/s41592-023-02104-6.

Yang, T., Sucholutsky, I., Jen, K.-Y., and Schonlau, M. exKidneyBERT: a language model for kidney transplant pathology reports and the crucial role of extended vocabularies. *PeerJ Computer Science*, 10:e1888, February 2024. ISSN 2376-5992. doi: 10.7717/peerj-cs.1888. URL https://peerj.com/articles/cs-1888.

A. Comparison between pretrained and neuro-tokenizer


Figure 3. **Tokenization examples.** Compared to a pretrained tokenizer trained on general text, a neuro-tokenizer trained specifically on neuroscience literature better preserves domain-specific terminology, such as brain regions, in neuroscience.

B. Confidence calibration of domain-specific pretrained models

The absolute difference of perplexities of two versions of the abstract was used as a measure of model confidence. To assess the calibration of GPT-2 variants pretrained on neuroscience literature, we compared their accuracies with their confidence levels. First, we ranked and sorted model confidence across all test cases. Subsequently, we created 20 bins based on this sort. Within each bin, we calculated the mean accuracy. A well-calibrated model will exhibit a higher accuracy in bins associated with higher confidence rankings. We fit a linear regression model using the bin number as the independent variable and the mean accuracy of each bin as the dependent variable to evaluate calibration. We observed that human experts and models all show positive correlations between confidence and accuracy, indicating calibration (Fig. 4)

In addition, we fitted logistic regressions using perplexity differences to models' answers and from self-reported confidences of human experts to their responses. We confirmed a positive and significant correlation between model (355M and 774M) perplexities and their answers as well as human confidences and their responses (Table 2).
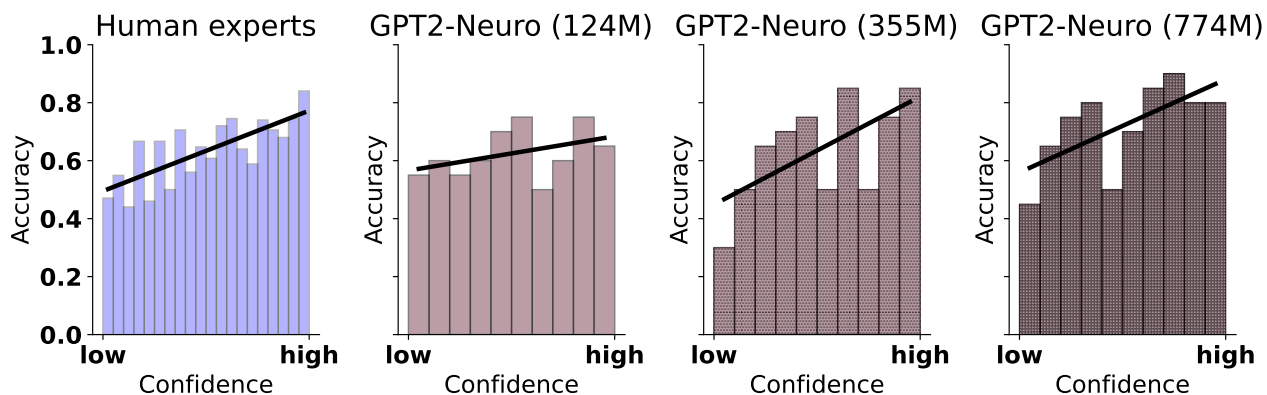
Figure 4. **Accuracy and confidence are calibrated for human experts and GPT-2 variants pretrained on neuroscience literature.** When human experts and models are confident in their BrainBench judgments, they are more likely to be correct. Confidence ratings were sorted and placed in equally-sized bins with the mean accuracy for items in that bin plotted. The positive slope of the black regression lines for human experts and all models indicates that confidence is well calibrated (i.e., higher confidence corresponds to higher accuracy). Calibration is beneficial for human-machine teams.

Model/Human	Coefficient	Intercept	Test Statistics
GPT2-Neuro (124M)	0.04 ± 0.02	0.41 ± 0.05	$t(4) = 2.14, p = 0.049$
GPT2-Neuro (355M)	0.60 ± 0.03	0.23 ± 0.05	$t(4) = 19.95, p < .001$
GPT2-Neuro (774M)	0.87 ± 0.09	0.57 ± 0.04	$t(4) = 9.21, p < .001$
Human experts	0.01 ± 0.00	0.04 ± 0.03	$t(4) = 19.11, p < .001$

Table 2. **Calibration analysis using logistic regression fits.** For models, logistic regressions were fitted between perplexity differences of a test case and its correctness given a LLM. For human experts, logistic regressions were fitted between their confidence of a test case and its correctness.

C. Contextual integration analysis

To investigate the extent to which GPT-2 models pretrained on neuroscience can integrate broad context from abstracts, we conducted an experiment involving the removal of contextual information from BrainBench test cases. Following the same evaluation procedure outlined for full abstract cases, we assessed the models using individual sentences extracted from abstracts containing at least one result alternation. In cases with multiple alternations, we computed the mean accuracy across these alternations as the final accuracy for the abstract. We then compared the level of performance degradation when these models were evaluated on full-length abstracts versus individual sentences where background and method information from the abstracts were removed.
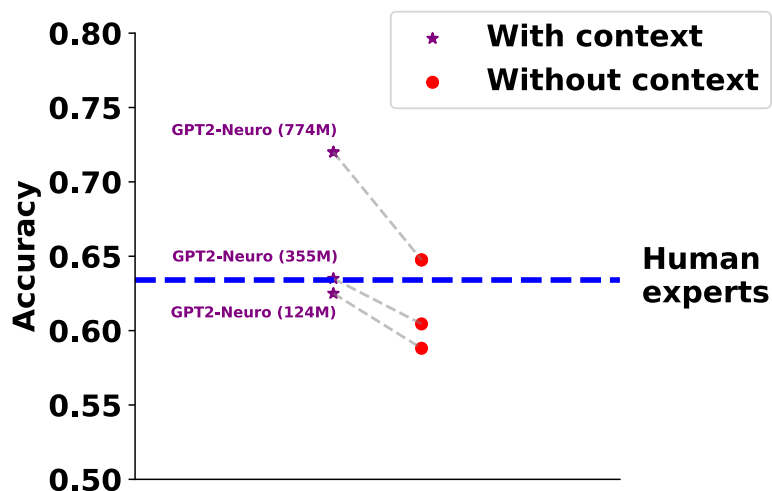

Figure 5. **GPT-2 variants pretrained on neuroscience literature integrate contextual information to succeed on BrainBench**. The removal of background and method sections from abstracts, with an evaluation based solely on individual sentences and result alternations, significantly impairs the performance of models on BrainBench. Much of the human-level performance appears to arise from integrating information across the abstract.

As a result, models performed worse when wider context was removed (Fig. 5), which provides strong evidence that these models are integrating information across the abstract, including information on background and methods.

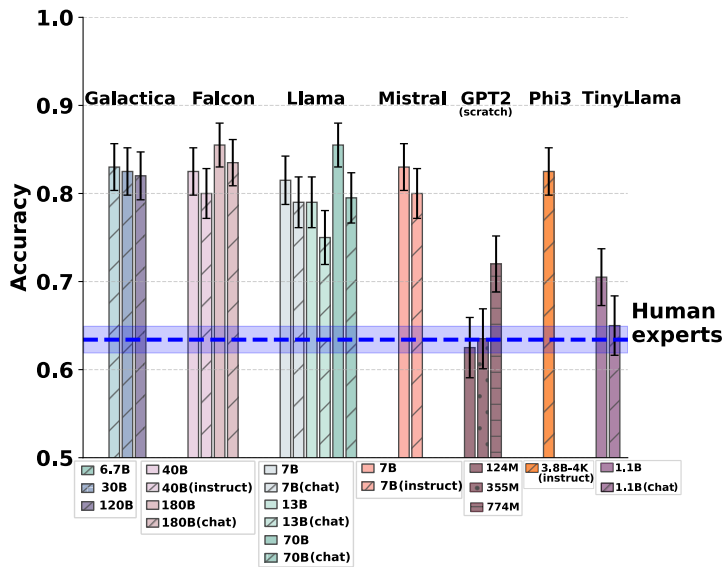## D. BrainBench performance across model sizes and training data

Figure 6. **BrainBench performance across models of varying sizes and training data.** The GPT-2 variants (124M, 355M, 774M) pretrained entirely on the neuroscience literature from scratch shows progressively better results, matching or surpassing human performance and closing the gap to larger models tested in Luo et al. (2024). Phi3, with half the size of the 7B models, achieves competitive results likely due to its high-quality training data. In contrast, TinyLlama, with 1.1 billion parameters, lags behind on BrainBench. Overall, performance on domain-specific tasks such as BrainBench is influenced by both model size and the quality and relevance of the training data.

E. Training details

Variants of GPT-2 models using Huggingface implementations. We used a batch size of 16 for GPT-2 124M (8 for GPT-2 355M and 4 for GPT-2 774M) and a chunk size of 1024. Training involved the use of the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 2e-5 and a cosine learning rate scheduler. We applied gradient accumulation steps set at 8. Five training epochs were performed, along with a warm-up step of 0.03 and a weight decay rate of 0.001. bf16 mixed precision training and data parallelism were employed. We used 4 Nvidia A100 (80GB) GPUs hosted on Microsoft Azure.