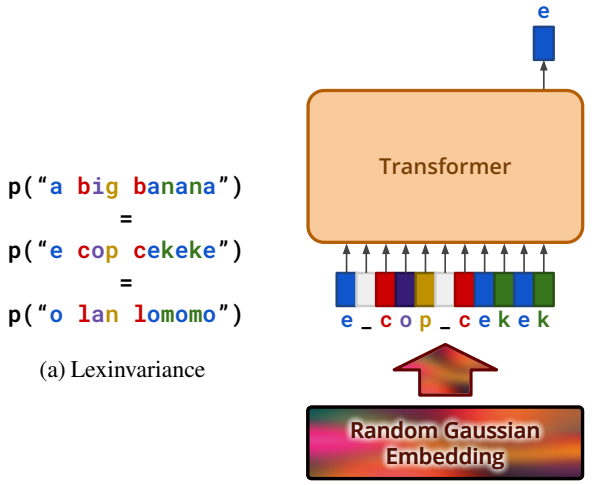
Lexinvariant Language Models

Qian Huang [1] **Eric Zelikman** [1] **Sarah Li Chen** [1] **Yuhuai Wu** [1 2] **Gregory Valiant** [1] **Percy Liang** [1]

Abstract

Token embeddings, a mapping from discrete lexical symbols to continuous vectors, are at the heart of any language model (LM). However, lexical symbol meanings can also be determined and even redefined by their structural role in a long context. In this paper, we ask: is it possible for a language model to be performant without *any* fixed token embeddings? Such a language model would have to rely entirely on the co-occurence and repetition of tokens in the context rather than the *a priori* identity of any token. To answer this, we study *lexinvariant* language models that are invariant to lexical symbols and therefore do not need fixed token embeddings in practice. First, we prove that we can construct a lexinvariant LM to converge to the true language model at a uniform rate that is polynomial in terms of the context length, with a constant factor that is sublinear in the vocabulary size. Second, to build a lexinvariant LM, we simply encode tokens using random Gaussian vectors, such that each token maps to the same representation within each sequence but different representations across sequences. Empirically, we demonstrate that it can indeed attain perplexity comparable to that of a standard language model, given a sufficiently long context. We further explore two properties of the lexinvariant language models: First, given text generated from a substitution cipher of English, it implicitly implements Bayesian in-context deciphering and infers the mapping to the underlying real tokens with high accuracy. Second, it has on average 4X better accuracy over synthetic in-context reasoning tasks. Finally, we discuss regularizing standard language models towards lexinvariance and potential practical applications.



(a) Lexinvariance

(b) Lexinvariant Language Model

Figure 1. Definition (a) and construction (b) of lexinvariant language model

1. Introduction

All language processing systems rely on a stable lexicon, which assumes that a token (a word or subword such as *tree*) has a consistent contribution to the meaning of a text (though of course this meaning is mediated by context). In neural language models (LMs), this contribution is the token embedding, which *stably* maps each token into a continuous vector (Schütze, 1993; Mikolov et al., 2013; 2010; Devlin et al., 2019; Brown et al., 2020). However, in real language, a token's contribution might be determined by its structural role; in math and code, novel variable names are arbitrarily defined to carry new meaning, and poems such as Jabberwocky exploit humans' lexical flexibility in interpreting novel words such as *vorpal*. Besides standard language understanding, this lexical flexibility also correlates with a stronger in-context reasoning performance. For example, GPT-3 (Brown et al., 2020) and other large language models that demonstrate high lexical flexibility show strong performance on tasks involving in-context reasoning over new concepts and rules.

Motivated by the above, we ask whether we can push this flexibility to the extreme: can we build a language model without *any* stable lexical mapping? To this end, we formulate and study such *lexinvariant* language models. We define a lexinvariant language model as a language model that assigns the same probability to all lexical permutations of a sequence.

*Equal contribution [1]Stanford University [2]Google Research. Correspondence to: Qian Huang <qhwang@cs.stanford.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Formally, we define a lexical permutation π to be a one-to-one mapping of a set of lexical symbols [1] onto itself. Then the lexinvariant language model is defined as a language model over the symbol sequence $x_1,...,x_n$ with the following property:

$$p(x_1,...,x_n) = p(\pi(x_1),...,\pi(x_n)) \; \forall \pi \tag{1}$$

For example, a lexinvariant language model (whose vocabulary is letters and space) should assign the same probability to the phrase "`a big banana`" as "`e cop cekeke`" because the two are the same up to the permutation $\pi = \{$`a:e,b:c,i:o,n:k,g:p,`$\cdots\}$ (Figure 1a).

The central question is: how well can lexinvariant language models predict the next token given an increasingly long context? We find the answer is almost as well as standard language models, both theoretically and empirically. This is rather surprising given that lexinvariance seems like a strong limitation (a model doesn't know what any individual symbol means!) However, the intuition is that given longer contexts, a lexinvariant model can both infer the latent permutation π (lazily) up to whatever ambiguity is present in the language model, and do the standard next word prediction task jointly.

Theoretically, we prove that a constructed lexinvariant language model can converge to the true language model as the context length increases—that is, the average L1 distance between the predictions of the two models decreases with a convergence rate of $O\left((\frac{d}{T})^{\frac{1}{4}}\right)$, where T is the length of the context and d is the vocabulary size, and where the big-O notation hides polylogarithmic factors of d and T and an absolute constant that is *independent* of the language model.

Empirically, we train a lexinvariant LM by replacing standard embeddings in a decoder-only Transformer (Vaswani et al., 2017) with per-sequence random Gaussian vectors, such that the same symbols get the same embedding within each sequence but get different embedding across sequences (Figure 1b). We indeed see that the perplexity gap between the lexinvariant LM and the standard LM shrinks as context length increases, as shown in Section 3.1. With a 150M parameters Transformer and a small character-level vocabulary (130 tokens), the average perplexity gap shrinks from 9X to less than 1X the average perplexity of a standard LM after observing 512 tokens over The Pile (Gao et al., 2020). With a larger 32K vocabulary, the gap also shrinks, especially on the more structured text like GitHub code, albeit at a much slower rate.

We then explore two additional properties of the lexinvariant LM: in-context deciphering and symbol manipulation. First, we show that given a ciphertext generated by applying a substitution cipher to English text, the lexinvariant LM can be seen as implicitly approximating Bayesian inference of the lexical permutation, i.e., cipher key, in-context. We show that the accuracy of this inferred cipher key quickly improves as context length grows, reaching 99.6% average accuracy. We also show examples in Section G that visualize the uncertainties over different possible lexical mappings maintained by the lexinvariant LM when the cipher key is ambiguous and that the semantic meaning of a symbol with very rare occurrence can be inferred efficiently relative to other common symbols in context. Second, we show that lexinvariant models perform better than traditional models over synthetic pure in-context reasoning tasks that involve symbol manipulation. We observe a significant 4X improvement over a standard language model. Finally, we discuss potential approaches to integrate the idea of lexinvariant LM into standard language modeling as a form of regularization, such that the LM assumes some form of partially stable symbol representations. The resulting LM can improve upon a standard language model over some BIG-bench tasks (Srivastava et al., 2022).

[1] We specifically consider lexical symbols as tokens, not necessarily words or other linguistic units.

2. Lexinvariant Language Model

We define a language model as a probability distribution $p(x_1,\ldots,x_n)$ over input token sequences $x_1,\ldots,x_n \in \mathcal{V}^n$, where $\mathcal{V}$ is some vocabulary over symbols. A language model is lexinvariant if for all permutations $\pi : \mathcal{V} \to \mathcal{V}$ and for all token sequences $x_1,\ldots,x_n \in \mathcal{V}^n$, $p(x_1,\ldots,x_n) = p(\pi(x_1),\ldots,\pi(x_n))$. For example, if $\mathcal{V} = \{\mathrm{a},\mathrm{b}\}$, then the model should assign the same probability to aab and bba. One example p that satisfies this could simply be

$$p(x) = \begin{cases} 1/2 & x \in \{aab,bba\} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Can such a lexinvariant language model predict language well, even though it can only make next token predictions based on the structure of co-occurence and repetition of input tokens in a single context?

2.1. Convergence on Language Modeling Performance

We show that we can construct a lexinvariant LM (as shown in Figure 2) to model the true language distribution faithfully, given a long enough context. The lexinvariant language model can essentially infer back the latent permutation π as it observes more symbols.

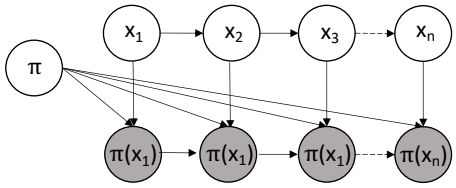


Figure 2. Probabilistic graphical model for the lexinvariant LM associated with the true language distribution $p(x_1,...,x_n)$.

As an intuitive example, suppose that $\mathcal{V} = \{a,b\}$ and the true language only contains two sequences $babbbb$ and $ababab$

(and their prefixes) with even probability. When given only the first three letters, a lexinvariant model can't tell the latent permutation and can only assign the same probability to a and b for the next letter: Due to the lexinvariant property, it assigns the same probability to $p(a|aba) = p(b|bab)$ as well as to $p(b|aba) = p(a|bab)$. Further, $p(a|aba) = p(b|aba)$ because the permutations to prefixes aba and bab are equally probable. In contrast, when considering the prefix $abab$, the fourth letter resolves the ambiguity in possible permutations π. (Since $baba$ is not in the true language distribution, π cannot map a to b.) Therefore, the model can correctly predict that $p(a|abab) = 1$ and $p(b|abab) = 0$.

Formally, for a given language model p, we define the associated lexinvariant language model $p'(x_1, \dots, x_n)$ as $\mathbb{E}_\pi[p(\pi^{-1}(x_1), \dots, \pi^{-1}(x_n))]$. Analyzing it, we have the following theorem:

Theorem 2.1. *Let $x_1,\dots,x_n$ be any token sequence generated by an arbitrary language distribution p with an alphabet of size d. Let $p'(x_1,\dots,x_n) = \mathbb{E}_\pi[p(\pi^{-1}(x_1),\dots,\pi^{-1}(x_n))]$. Then, for any $0 < \epsilon, \delta < 1/2$,*

$$\frac{1}{T}\sum_{t=1}^{T} \|p(x_t|x_1,\dots,x_{t-1}) - p'(x_t|x_1,\dots,x_{t-1})\|_1 \le \epsilon$$

with probability greater than $1 - \delta$, when $T \ge \frac{d}{\epsilon^4}\ \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$, where the polylogarithmic term hides an absolute constant that is independent of p.

This theorem says that this associated lexinvariant language model converges to modeling the true language distribution fairly efficiently—with polynomial rate and near-linear dependence on vocabulary size d. Strikingly, this holds irrespective of the properties of the language distribution p [2]. In other words, *a language model can indeed infer the operational meaning of the tokens in context based solely on the structure of the symbols*!

We give a complete proof of this theorem in Appendix A. At a high level, this convergence happens because at most timesteps t, the new observation x_t either provides new information about the permutation π, or x_t has similar likelihood under the permutations that are likely given $x_1,\dots,x_{t-1}$. In the simplest case, if the posterior $p(\pi \mid x_1,\dots,x_n)$ concentrates on the correct π, then we converge to the standard LM. But even if it doesn't, that means the uncertainty about π should not matter for predicting the next token. We make this precise by interpreting $p'(x_t|x_1,\dots,x_{t-1})$ as performing a multiplicative weights algorithm with the Hedge strategy of Freund and Schapire (Freund & Schapire, 1997), and then relate the regret bounds to the average KL divergence between the predictions of p and p', and ultimately the average $\mathcal{L}_1$ distance between these predictions.

[2] The convergence rate could be better depending on the language distribution, such as on math and code, where the symbols should have clear functional meaning in context. We explore this empirically in the experiment section.

2.2. Constructing a Lexinvariant Language Model

We now consider how to construct a lexinvariant LM in practice. A typical neural language model, such as a Transformer, converts input tokens to continuous vectors using token embedding and then passes these vectors as input to the rest of the neural network. Thus, the language model p it parameterizes depends on the token embedding $E : \mathcal{V} \to \mathbb{R}^d$:

$$p(x_1,\dots,x_n) = T(E(x_1),\dots,E(x_n)) \tag{3}$$

To make a neural LM lexinvariant, we can replace the standard stable token embedding E with a randomized E and take the expectation over E. Each token $x \in \mathcal{V}$ has an independent embedding $E(x) \sim \mathcal{N}(0, \mathcal{I}_d)$, and the language model becomes

$$p(x_1,\dots,x_n) = \mathbb{E}[T(E(x_1),\dots,E(x_n))] \tag{4}$$

Since $E \overset{d}{=} E \circ \pi$, the right-hand side is the same when x_i are applied with any permutation π, i.e., for any $x_1,\dots,x_n$:

$$\mathbb{E}[T(E(x_1),\dots,E(x_n))] = \mathbb{E}[T(E(\pi(x_1)),\dots,E(\pi(x_n)))], \tag{5}$$

showing that the Transformer with random E is lexinvariant as in Eq. 1. Now we can train this lexinvariant LM similarly to a standard LM. Concretely, we sample a new E for each training sequence and minimize the standard language modeling loss as in a standard neural LM. Here we are stochastically optimizing a variational lower bound of the standard language modeling loss with this randomized model by taking the expectation to the outside of the loss over log likelihood. Effectively, the same token gets the same random embedding within each training sequence, but different embedding across training sequences.

In practice, we focus on training decoder-only Transformers with a next token prediction objective in this work, where the model directly models $p(x_{n+1}|x_1,\dots,x_n)$ instead of the joint distribution. Our definitions and analysis above still hold in general. The only modification is that the final readout matrix also needs to be replaced with the same E, so that the Transformer can predict the embedding of the next token based on the embeddding of input tokens.

3. Experiments

See detailed experiment setup in Appendix C.

3.1. Convergence to Standard Language Models

We first show empirically that lexinvariant LMs can mostly recover the next token prediction performance of standard

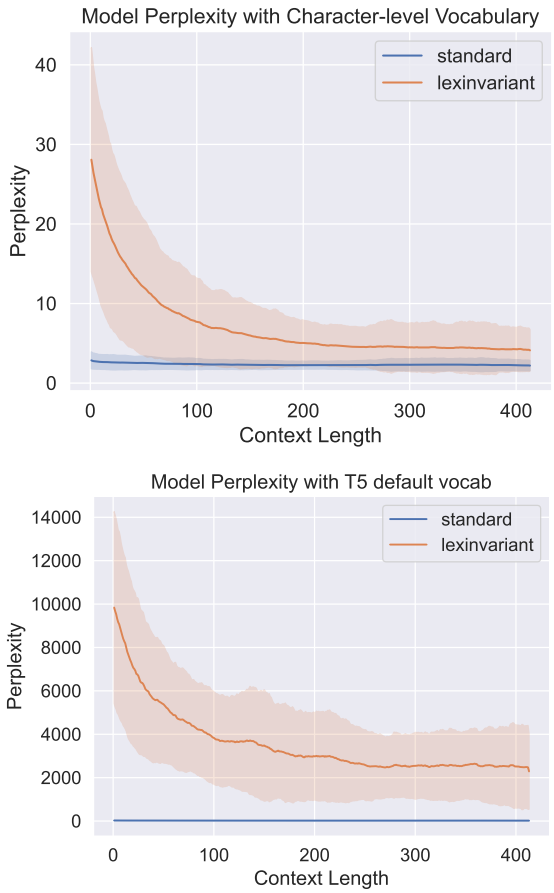


Figure 3. Perplexity over the Pile with character-level vocabulary (left) and T5 default vocab (right).

LMs after a long enough context. Here, we train lexinvariant and standard LMs with both character-level vocabulary (128 ascii characters) and T5 default vocab (32k tokens) over the three datasets. For each model, we measure the perplexity of each token in each sequence w.r.t. context length, smoothed by moving average within each sequence, i.e. $P(x_i,...,x_{i+k}|x_1,...,x_i)^{\frac{1}{k}}$ for context length i. We set the moving average window $k=100$. We plot results over 100 sequences. As shown in figure 3, the perplexity gap between lexinvariant LM and standard LM gradually shrinks as the prefix becomes longer and longer, albeit much more slowly with a larger vocabulary. This makes intuitive sense since a larger vocabulary has more possibilities of permutations and requires many more prefix tokens to disambiguate. Additionally, we observe that the gap shrinks significantly faster for models trained over Github than standard English text like Wiki-40B since code is more structured and it is easier to decipher the token permutation. We show the comparison across different datasets in Figure 4 in Appendix.

3.2. Recovering Substitution Ciphers

We show that lexinvariant LM is implicitly performing Bayesian in-context deciphering by testing its ability to recover cipher keys (e.g. Figure 5a) from character-level substitution ciphers. We discuss the intuition about implicit Bayesian in-context deciphering in Appendix B. To show this epirically, we train a small MLP probe on top of a frozen pretrained lexinvariant LM to predict the deciphered token corresponding to the last seen cipher token. We can then read out the inferred cipher key with each prefix of the sequence. As shown in Figure 5c, such a cipher key prediction generally has increasingly higher precision as the window is selected later in the context, and it becomes near-perfect by the end of the sequence. Specifically, the cipher key matrix produced by the last window has an average precision of 99.6% over 1000 input sequences. We show full analysis and examples in Appendix F and G.

3.3. Synthetic Reasoning Tasks

As discussed in the introduction, lexical flexibility is correlated with in-context reasoning performance, as demonstrated by existing large LMs. Thus, we study whether the lexinvariant model also learns in-context reasoning capabilities through the challenging lexinvariant training.

Specifically, we measure the performance of lexinvariant models over two pure in-context symbol manipulation tasks: LookUp, where the task is to predict the next token based on the given lookup table, e.g. `A->2⏎C->4⏎G->5⏎C->` (should predict `4` here); and Permutation, where the task is to permute an arbitrary subsequence of the given sequence the same way as in the given few demonstrations, e.g. `A 2 C->C A⏎4 1 D->` (should predict `D 4` here). In each of the tasks, the symbols are randomly sampled from the vocabulary so that we measure the pure reasoning ability independent from any knowledge of specific words. We measure the model performance in terms of generated token accuracy over 1000 examples. The results are shown in Table 1. As shown in the table, the lexinvariant models achieve drastically higher accuracy, with an average of 4X improvement. We discuss potential ways to extend this to more realistic tasks in Appendix I.

4. Conclusion

In this work, we define and study lexinvariant language models, which do not have stable embeddings and learn to infer the meaning of symbols in-context. We show several surprising properties of this model theoretically and empirically, including convergence to standard language modeling, in-context deciphering, and better reasoning capabilities. We also explore a less extreme lexinvariance regularized language model and demonstrate its potential for solving more practical tasks efficiently.

Acknowledgements

We thank Sang Michael Xie for the discussions and for providing feedback on our manuscript. Qian Huang is supported by Open Philanthropy AI fellowship.

References

The multiplicative weights algorithm. https://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15859-f11/www/notes/lecture16.pdf. Accessed: 2023-04-24.

Akyürek, E., Akyurek, A. F., and Andreas, J. Learning to recombine and resample data for compositional generalization. *ArXiv*, abs/2010.03706, 2020.

Aldarrab, N. and May, J. Can sequence-to-sequence models crack substitution ciphers? In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

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. J., 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., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, 1997.

Gao, L., Biderman, S. R., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020.

Guo, M., Dai, Z., Vrandecic, D., and Al-Rfou, R. Wiki-40b: Multilingual language model dataset. In *LREC 2020*, 2020.

Harnad, S. Symbol grounding problem. *Scholarpedia*, 2: 2373, 1990.

Hauer, B., Hayward, R. B., and Kondrak, G. Solving substitution ciphers with combined language models. In *International Conference on Computational Linguistics*, 2014.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.

Jia, R. and Liang, P. Data recombination for neural semantic parsing. *ArXiv*, abs/1606.03622, 2016.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., R'e, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L. J., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N. S., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T. F., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models. *ArXiv*, abs/2211.09110, 2022.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. Recurrent neural network based language model. In *Interspeech*, 2010.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

Nuhn, M. and Ney, H. Decipherment complexity in 1:1 substitution ciphers. In *Annual Meeting of the Association for Computational Linguistics*, 2013.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

Raiman, J. and Miller, J. Globally normalized reader. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

Schütze, H. Part-of-speech induction from scratch. In *31st Annual Meeting of the Association for Computational Linguistics*, pp. 251–258, Columbus, Ohio, USA, June 1993. Association for Computational Linguistics. doi: 10.3115/981574.981608. URL https://aclanthology.org/P93-1034.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4603–4611. PMLR, 2018. URL http://proceedings.mlr.press/v80/shazeer18a.html.

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., Rahane, A. A., Iyer, A. S., Andreassen, A., Santilli, A., Stuhlmuller, A., Dai, A. M., La, A. D., 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., 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. T., 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., 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., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E. P., Pavlick, E., Rodolà, E., Lam, E. F., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E. J., 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., Jaimovitch-L'opez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H. S., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H., Ng, I. A.-S., 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., 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., Berant, J., Frohberg, J., Rozen, J., Hernández-Orallo, J., Boudeman, 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. O., 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., Andrea, M., Farooqi, M. S., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M., Hagen, M., Schubert, M., Baitemirova, M., Arnaud, M., McElrath, M. A., Yee, M. A., 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., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., MukundVarma, T., Peng, N., Chi, N., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N. S., Doiron, 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. W., 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., Delgado, R. R., Millière, 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. J., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S., 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., hwan Lee, S., Torene, S. B., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S. R., Lin, S. C., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T. A., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T. N., Telleen-Lawton, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T. O., 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. F., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y. J., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Xinran, Z., Zhao, Z., Wang, Z. F., Wang, Z. J., Wang, Z., Wu, Z., Singh, S., and Shaham, U. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2021.

A. Convergence Proof

Theorem A.1. *Let $x_1,...,x_n$ be any token sequence generated by an arbitrary language distribution p with an alphabet of size d. Let $p'(x_1,...,x_n)=\mathbb{E}_\pi[p(\pi^{-1}(x_1),...,\pi^{-1}(x_n))]$. Then, for any $0<\epsilon,\delta<1/2$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|p(x_t|x_1,...,x_{t-1})-p'(x_t|x_1,...,x_{t-1})\|_1\leq\epsilon$$

with probability greater than $1-\delta$ when $T\geq\frac{d}{\epsilon^4}\operatorname{polylog}(d,\frac{1}{\epsilon},\frac{1}{\delta})$.

Proof. For any desired error $0<\epsilon<1/2$ and failure rate $0<\delta<1/2$, we will first prove the analogous statement for KL divergence instead of $\mathcal{L}_1$ distance, and then relate a bound on KL divergence back to $\mathcal{L}_1$ distance via Pinsker's inequality.

Throughout the rest of proof, we will work with a parameter $\epsilon'<O(\frac{\epsilon}{(\log(1/\delta))^{1/4}})<\frac{1}{2}$, and will bound our KL divergence by ϵ'.

To prove the bound in terms of KL divergence, it will be useful to ensure to work with a "smoothed" version of p, which we denote by $\tilde{p}$, for which every token has some nonzero probability, σ/d, of appearing at each timestep, for a parameter $\sigma=\delta\epsilon'/T$:

$$\tilde{p}(x_T|x_1,...,x_{T-1})=p(x_T|x_1,...,x_{T-1})(1-\sigma)+\frac{\sigma}{d}.$$

Similarly, let $\tilde{p}'(x_1,...,x_n)=\mathbb{E}_\pi[\tilde{p}^{-1}(\pi(x_1),...,\pi^{-1}(x_n))]$. We use $\tilde{\mathcal{P}}$ to denote the probabilities under this change. With probability at least $1-\sigma T\geq 1-\epsilon'\delta\geq 1-\frac{\delta}{2}$, the realized sequence $x_1,...,x_n$ drawn under p can be regarded as being drawn from $\tilde{p}$ (as these distributions can be coupled with this probability).

The key idea is then to show that $\tilde{p}'(y_{t+1}|y_{1:t})$, where $y_t=\pi^*(x_t)$ for some ground truth π^* unknown to p', is equivalent to using the multiplicative weights algorithm to predict y_{t+1} with the Hedge strategy, with the experts being each possible permutation of the tokens and the cost incurred by each expert being the negative log likelihood of the prediction. We denote $\tilde{\mathcal{P}}_{\pi'}(y_{1:n})=\tilde{\mathcal{P}}(y_{1:n}|\pi=\pi')=\tilde{p}(\pi^{-1}(y_1),...,\pi^{-1}(y_n))$ and show this in Lemma A.2.

With this equivalence, we can then bound the difference between the prediction of p and p' as the regret of the multiplicative weights algorithm. Concretely, we show in Lemma A.3 that the regret of p' to any expert π is bounded as

$$\frac{1}{T}\sum_{t}^{T}\log\frac{\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t})}{\tilde{p}'(y_{t+1}|y_{1:t})}\leq 2\epsilon'^2$$

for $T\geq\left(4\log^2(\frac{d}{\sigma})\log(d!)\right)/\epsilon'^4$.

We can see $\tilde{p}$ as the particular expert/permutation $\tilde{\mathcal{P}}_I$. And we can further only consider the special case that π^* is also the identity permutation, then the same result holds over x_t and with $\tilde{\mathcal{P}}_\pi$ replaced by $\tilde{p}$, i.e.

$$\frac{1}{T}\sum_{t}^{T}\log\frac{\tilde{p}(x_{t+1}|x_{1:t})}{\tilde{p}'(x_{t+1}|x_{1:t})}\leq 2\epsilon'^2$$

Now we want to convert this bound on regret in terms of log likelihood to KL divergence, and eventually to $\mathcal{L}_1$ distance. To convert it to KL divergence regret, we construct a martingale:

$$Z_i=\sum_{t=1}^{i}\left(D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})\|\tilde{p}'(x_{t+1}|x_{1:t}))-\log\frac{\tilde{p}(x_{t+1}|x_{1:t})}{\tilde{p}'(x_{t+1}|x_{1:t})}\right).$$

We verify that this is a martingale in Lemma A.4, with differences bounded by $2\log\frac{1}{\sigma}$, and bound the probability that Z_T exceeds $b=\log\frac{d}{\sigma}\sqrt{8T\log\frac{2}{\delta}}$ via Azuma's inequality Lemma A.6: with probability $1-\delta/2$, we have that $|Z_T|\leq b$.

Therefore, we have that with probability at least $1-\delta/2$

$$Z_T = \sum_{t=1}^{T}\left(D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})\|\tilde{p}'(x_{t+1}|x_{1:t})) - \log\frac{\tilde{p}(x_{t+1}|x_{1:t})}{\tilde{p}'(x_{t+1}|x_{1:t})}\right) \leq b$$

$$\sum_{t=1}^{T} D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})\|\tilde{p}'(x_{t+1}|x_{1:t})) \leq \sum_{t=1}^{T}\left(\log\frac{\tilde{p}(x_{t+1}|x_{1:t})}{\tilde{p}'(x_{t+1}|x_{1:t})}\right) + b$$

Putting this all together, since $\frac{1}{T}\sum_t^T \log\frac{\tilde{p}_i(x_{t+1}|x_{1:t})}{\tilde{p}'(x_{t+1}|x_{1:t})} \leq 2\epsilon'^2$ for $T \geq \left(4\log^2(\frac{d}{\sigma})\log(d!)\right)/\epsilon'^4$, we have the following:

$$\sum_{1}^{T} D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})\|\tilde{p}'(x_{t+1}|x_{1:t})) \leq 2\epsilon'^2 T + b.$$

We now convert our bound on KL divergence to a bound on $\mathcal{L}_1$ distance via Pinsker's inequality:

$$\|\tilde{p}(x_{t+1}|x_{1:t}) - \tilde{p}'(x_{t+1}|x_{1:t})\|_1 \leq \sqrt{\frac{1}{2} D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})||\tilde{p}'(x_{t+1}|x_{1:t}))}.$$

Further, at any given x_t, the difference between the redistributed probability distribution $\tilde{p}$ and a unmodified probability distribution p is at most σ, so

$$\|p(x_{t+1}|x_{1:t}) - p'(x_{t+1}|x_{1:t})\|_1 \leq \|\tilde{p}(x_{t+1}|x_{1:t}) - \tilde{p}'(x_{t+1}|x_{1:t})\|_1 + 2\sigma.$$

We are interested in the average $\mathcal{L}_1$ across time steps:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|p(x_{t+1}|x_{1:t}) - p'(x_{t+1}|x_{1:t})\|_1 &\leq \frac{1}{T}\sum_{t=1}^{T}(\|\tilde{p}(x_{t+1}|x_{1:t}) - \tilde{p}'(x_{t+1}|x_{1:t})\|_1 + 2\sigma) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{2} D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})\|\tilde{p}'(x_{t+1}|x_{1:t}))} + 2\sigma \\
&\leq \frac{1}{T}\sqrt{T\sum_{t=1}^{T}\frac{1}{2} D_{KL}(\tilde{p}(x_{t+1}|x_{1:t})\|\tilde{p}'(x_{t+1}|x_{1:t}))} + 2\sigma,
\end{aligned}$$

where in the last inequality we applied Cauchy–Schwarz. Hence for $T \geq \left(4\log^2\frac{d}{\sigma}\log(d!)\right)/\epsilon'^4$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|p(x_{t+1}|x_1,\ldots,x_t) - p'(x_{t+1}|x_1,\ldots,x_t)\|_1 &\leq \frac{1}{T}\sqrt{\frac{T}{2}(2\epsilon'^2 T + b)} + 2\sigma \\
&\leq \sqrt{\epsilon'^2 + \frac{b}{2T}} + 2\sigma.
\end{aligned}$$

Simplifying this for $b = \log\frac{d}{\sigma}\sqrt{8T\log\frac{2}{\delta}}$, $T \geq \left(4\log^2(\frac{d}{\sigma})\log(d!)\right)/\epsilon'^4$ and $\sigma = \epsilon'\delta/T$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|p(x_{t+1}|x_1,...,x_t) - p'(x_{t+1}|x_1,...,x_t)\|_1 &\leq \sqrt{\epsilon'^2 + \frac{\sqrt{2\log\frac{2}{\delta}}}{\sqrt{\log(d!)}}\epsilon'^2} + \frac{2\epsilon'\delta}{T} \\ &\leq \epsilon'(\frac{2\delta}{T} + \sqrt{1 + \sqrt{\frac{2\log\frac{2}{\delta}}{\log(d!)}}}) \\ &\leq \epsilon'(1 + \sqrt{1 + \sqrt{2\log\frac{2}{\delta}}}) \leq \epsilon' 2\sqrt{2}(2\log\frac{2}{\delta})^{1/4}.\end{aligned}$$

We can bound this average L_1 error by ϵ if we set $\epsilon' = \frac{\epsilon}{2\sqrt{2}(2\log\frac{2}{\delta})^{1/4}} < \frac{1}{2}$, in which case our condition that $T \geq \left(4\log^2(\frac{dT}{\delta\epsilon'})\log(d!)\right)/\epsilon'^4$ becomes $T \geq \left(512\log\frac{2}{\delta}\log^2\frac{dT}{\delta\epsilon'}\log(d!)\right)/\epsilon^4$. The theorem now follows by simplifying this expression. Since $\log\frac{2}{\delta} \leq 2\log\frac{1}{\delta}$, and $\log(d!) \leq d\log(d)$, we can relax the condition on T as

$$T \geq \left(1024\log\frac{1}{\delta}\log^2(\frac{d}{\delta\epsilon'})\log^2(T)d\log(d)\right)/\epsilon^4 = \log^2(T)\frac{d}{\epsilon^4}\,polylog(d,\frac{1}{\epsilon},\frac{1}{\delta})$$

To remove the $\log^2 T$ from the right side, note that for any $W > 10$, if T ¿ 10 $W\log^2 W$, then $T > W log^2 T$, yielding the further relaxed the condition on T as

$$T \geq \frac{d}{\epsilon^4}\,polylog(d,\frac{1}{\epsilon},\frac{1}{\delta}).$$

□

Lemma A.2. *Consider an arbitrary ground truth permutation π^*. For all time steps $t \in [1,n]$, let $y_t = \pi^*(x_t)$. Consider the online prediction game of predicting y_{t+1} at each time step given previous observation $y_{1:t}$ without knowing π^* but knowing $\tilde{p}$. Then, $\tilde{p}'(y_{t+1}|y_{1:t})$ is equivalent to the multiplicative weights algorithm's prediction of y_{t+1} with the Hedge strategy of Freund and Schapire (Freund & Schapire, 1997), where it*

- *Considers $d!$ experts corresponding to guessing each permutation π' is the ground truth permutation.*
- *Maintains a weight $w_{\pi'}^{(t)}$ for each expert at time step t, and the weights are initially as $\tilde{\mathcal{P}}(\pi)$.*
- *Picks a distribution across experts $p_{\pi'}^{(t)} = \frac{w_{\pi'}^{(t)}}{\Phi^{(t)}}$ where $\Phi^{(t)} = \sum_j w_j^{(t)}$.*
- *Produces prediction of y_{t+1} as $\sum_{\pi'} p_{\pi'}^{(t)} \tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t})$*
- *Receives a cost vector of $m_{\pi'}^{(t)} = -\frac{1}{\epsilon}\log\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t})$.*
- *Updates the weights $w_i^{(t+1)} = w_i^{(t)}\exp(-\epsilon m_i^{(t)})$ and repeat*

Proof. We can first see that $p_{\pi'}^{(t)} = \tilde{\mathcal{P}}(\pi'|y_{1:t})$ by induction:

Base case: $p_{\pi'}^{(0)} = \tilde{\mathcal{P}}(\pi)$ by assumption.

Inductive Case:

With the cost vector as $m_{\pi'}^{(t-1)} = -\frac{1}{\epsilon}\log\tilde{\mathcal{P}}_{\pi'}(y_t|y_{1:t-1})$, the update at step t is $w_{\pi'}^{(t)} = w_{\pi'}^{(t-1)}\tilde{\mathcal{P}}_{\pi'}(y_t|y_{1:t-1})$. Therefore, the

probability over any particular expert π' is

$$\begin{aligned}
p_{\pi'}^{(t)} &= \frac{w_{\pi'}^{(t)}}{\Phi^{(t)}} \\
&= \frac{w_{\pi'}^{(t-1)}\tilde{\mathcal{P}}_{\pi'}(y_t|y_{1:t-1})}{\sum_j w_j^{(t-1)}\tilde{\mathcal{P}}_j(y_t|y_{1:t-1})} \\
&= \frac{p_{\pi'}^{(t-1)}\Phi^{(t-1)}\tilde{\mathcal{P}}_{\pi'}(y_t|y_{1:t-1})}{\sum_j p_j^{(t-1)}\Phi^{(t-1)}\tilde{\mathcal{P}}_j(y_t|y_{1:t-1})} \\
&= \frac{p_{\pi'}^{(t-1)}\tilde{\mathcal{P}}_{\pi'}(y_t|y_{1:t-1})}{\sum_j p_j^{(t-1)}\tilde{\mathcal{P}}_j(y_t|y_{1:t-1})}
\end{aligned}$$

This is equivalent to the update given by Bayes' rule when plugging in $p_{\pi'}^{(t)} = \tilde{\mathcal{P}}(\pi'|y_{1:t})$:

$$\tilde{\mathcal{P}}(\pi'|y_{1:t}) = \frac{\tilde{\mathcal{P}}(\pi'|y_{1:t-1})\tilde{\mathcal{P}}_{\pi'}(y_t|y_{1:t-1})}{\tilde{\mathcal{P}}(y_t|y_{1:t-1})}$$

So we can conclude that $p_{\pi'}^{(t)} = \tilde{\mathcal{P}}(\pi'|y_{1:t})$, i.e. the process of updating the probability distribution across experts within the prediction game is equivalent to the process of the language model updating the probabilities $\tilde{\mathcal{P}}(\pi'|y_{1:t+1})$ across permutations π'. And this means that the algorithm's prediction $\sum_{\pi'} p_{\pi'}^{(t)}\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t}) = \sum_{\pi'}\tilde{\mathcal{P}}(\pi'|y_{1:t})\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t}) = \tilde{\mathcal{P}}(y_{t+1}|y_{1:t}) = \tilde{p}'(y_{t+1}|y_{1:t})$ □

Lemma A.3. *When using the Hedge strategy for the multiplicative weights algorithm, the average difference between the weighted distribution across experts and any particular expert π is bounded as*

$$\frac{1}{T}\sum_t^T \log\frac{\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t})}{\tilde{p}'(y_{t+1}|y_{1:t})} \leq 2\epsilon^2$$

for $\epsilon \leq 1$ and for $T \geq \left(4\log^2\left(\frac{d}{\sigma}\right)\log(d!)\right)/\epsilon^4$.

Proof. Consider an arbitrary expert π.

We first show that the cost vectors are bounded by $\rho = -\frac{1}{\epsilon}\log\frac{\sigma}{d}$: Recall we defined $m_\pi^{(t)} = -\frac{1}{\epsilon}\log\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t})$. By the definition of our redistributed probability distribution, at time step $t \in [1,...,T]$,

$$\begin{aligned}
\frac{\sigma}{d} &\leq \tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t}) \leq 1 \\
\log\frac{\sigma}{d} &\leq \log\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t}) \leq 0 \\
0 &\leq m_\pi^{(t)} \leq -\frac{1}{\epsilon}\log\frac{\sigma}{d} \\
0 &\leq m_\pi^{(t)} \leq -\frac{1}{\epsilon}\log\frac{\sigma}{d}.
\end{aligned}$$

By corollary 16.3 in (MW), if we have cost vectors $m^{(t)} \in [-\rho,\rho]^{d!}$, then for time $T \geq (4\rho^2\log(d!))/\epsilon^2$ where $\epsilon \leq 1$,

$$\frac{1}{T}\sum_t^T p^{(t)}\cdot m^{(t)} \leq \frac{1}{T}\sum_t^T m_\pi^{(t)} + 2\epsilon.$$

Note that we can simplify $T \geq \left(4\log^2\left(\frac{d}{\sigma}\right)\log(d!)\right)/\epsilon^4$.

We can now bound

$$\begin{aligned}
\frac{1}{T}\sum_t^T\left(p^{(t)}\cdot m^{(t)}-m_\pi^{(t)}\right)&\leq 2\epsilon\\
\frac{1}{T}\sum_t^T\left(\sum_{\pi'}p_{\pi'}^{(t)}m_{\pi'}^{(t)}-m_\pi^{(t)}\right)&\leq 2\epsilon\\
\frac{1}{T}\sum_t^T\left(\sum_{\pi'}\tilde{\mathcal{P}}(\pi'|y_{1:t})\left(-\frac{1}{\epsilon}\log\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t})\right)-\left(-\frac{1}{\epsilon}\log\tilde{\mathcal{P}}_{\pi}(y_{t+1}|y_{1:t})\right)\right)&\leq 2\epsilon\\
\frac{1}{\epsilon T}\sum_t^T\sum_{\pi'}\left(\tilde{\mathcal{P}}(\pi'|y_{1:t})\left(\log\tilde{\mathcal{P}}_{\pi}(y_{t+1}|y_{1:t})-\log\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t})\right)\right)&\leq 2\epsilon\\
\frac{1}{T}\sum_t^T\mathbb{E}_{\pi'}\log\frac{\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t})}{\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t})}&\leq 2\epsilon^2
\end{aligned}$$

By Jensen's inequality, we also have that

$$\begin{aligned}
\frac{1}{T}\sum_t^T\log\frac{\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t})}{\mathbb{E}_{\pi'}\tilde{\mathcal{P}}_{\pi'}(y_{t+1}|y_{1:t})}&\leq 2\epsilon^2\\
\frac{1}{T}\sum_t^T\log\frac{\tilde{\mathcal{P}}_\pi(y_{t+1}|y_{1:t})}{\tilde{p}'(y_{t+1}|y_{1:t})}&\leq 2\epsilon^2
\end{aligned}$$

□

Lemma A.4. *Let*

$$Z_i=\sum_{t=1}^i\left(D_{KL}(\tilde{\mathcal{P}}_I(x_{t+1}|x_{1:t})\|\tilde{\mathcal{P}}(x_{t+1}|x_{1:t}))-\log\frac{\tilde{\mathcal{P}}_I(x_{t+1}|x_{1:t})}{\tilde{\mathcal{P}}(x_{t+1}|x_{1:t})}\right)$$

Z_i is a martingale.

Proof. Consider

$$\begin{aligned}
\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}[Z_i]&=\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}\left[\sum_{t=1}^i\left(D_{KL}(\tilde{\mathcal{P}}_I(x_{t+1}|x_{1:t})\|\tilde{\mathcal{P}}(x_{t+1}|x_{1:t}))-\log\frac{\tilde{\mathcal{P}}_I(x_{t+1}|x_{1:t})}{\tilde{\mathcal{P}}(x_{t+1}|x_{1:t})}\right)\right]\\
&=\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}\left[D_{KL}(\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})\|\tilde{\mathcal{P}}(x_{i+1}|x_{1:i}))-\log\frac{\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})}{\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})}+Z_{i-1}\right]
\end{aligned}$$

Observe that Z_{i-1} has no dependence on x_{i+1}.

$$\begin{aligned}
\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}[Z_i]&=\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}\left[\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}\log\frac{\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})}{\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})}\right]-\mathbb{E}_{x_{i+1}\sim\tilde{\mathcal{P}}_I}\left[\log\frac{\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})}{\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})}\right]+Z_{i-1}\\
&=Z_{i-1}
\end{aligned}$$

Therefore, Z_i is a martingale. □

Lemma A.5. $|Z_i-Z_{i-1}|\leq c_i$ *where* $c_i=2|\log\frac{d}{\sigma}|$

Proof. We have

$$|Z_i - Z_{i-1}| = \left| D_{KL}(\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i}) \| \tilde{\mathcal{P}}(x_{i+1}|x_{1:i})) - \log \frac{\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})}{\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})} \right|$$

In our redistributed probability distribution $\tilde{\mathcal{P}}$, we have $\frac{\sigma}{d} \leq \tilde{\mathcal{P}}_\pi(x_i|x_{1:i-1}) \leq 1$ for any π at any time i. Therefore,

$$\log \frac{\sigma}{d} \leq \log \frac{\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})}{\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})} \leq \log \frac{d}{\sigma}.$$

Also, we can find an upper bound for the KL divergence by maximizing $\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})$ to 1 and minimizing $\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})$ to $\frac{\sigma}{d}$ so that

$$\begin{aligned} D_{KL}(\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i}) \| \tilde{\mathcal{P}}(x_{i+1}|x_{1:i})) &= \sum_{x_{i+1}} \tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i}) \log \frac{\tilde{\mathcal{P}}_I(x_{i+1}|x_{1:i})}{\tilde{\mathcal{P}}(x_{i+1}|x_{1:i})} \\ &\leq \log \frac{d}{\sigma} \end{aligned}$$

We can maximize $|Z_i - Z_{i-1}|$ by maximizing the first term and minimizing the second term, or vice versa. In the first case, $|Z_i - Z_{i-1}| \leq |\log \frac{d}{\sigma} - \log \frac{\sigma}{d}| = 2|\log \frac{d}{\sigma}|$. In the other case, $|Z_i - Z_{i-1}| \leq |0 - \log \frac{d}{\sigma}| = |\log \frac{d}{\sigma}|$.

Therefore, $|Z_i - Z_{i-1}| \leq c_i$ where $c_i = 2|\log \frac{d}{\sigma}|$. □

Lemma A.6. *By Azuma's inequality, with probability* $1-\delta$*, we have that* $\|Z_T\| \leq b$ *where* $b = 2\log \frac{d}{\sigma} \sqrt{-8T \log \frac{1}{\delta}}$

Proof. By Azuma's inequality, for all positive reals b,

$$\begin{aligned} P(Z_T - Z_1 \geq b) &\leq \exp\left(\frac{-b^2}{2\sum_{k=2}^T c_k^2} \right) \\ P(Z_T - Z_1 \leq b) &\geq 1 - \exp\left(\frac{-b^2}{2\sum_{k=2}^T c_k^2} \right) \\ &\geq 1 - \exp\left(\frac{-b^2}{8\sum_{k=2}^T \log^2 \frac{d}{\sigma}} \right) \end{aligned}$$

We can rewrite in terms of $\delta = \exp\left(\frac{-b^2}{8\sum_{k=2}^T \log^2 \frac{d}{\sigma}} \right)$ so

$$\begin{aligned} b &= \sqrt{-\left(8\sum_{k=2}^T \log^2 \frac{d}{\sigma} \right) \log \delta} \\ &\leq \log \frac{d}{\sigma} \sqrt{-8T \log \frac{1}{\delta}} \end{aligned}$$

Therefore,

$$P(Z_T - Z_1 \leq b) \geq 1 - \delta$$

□

B. In-context Bayesian Deciphering

We can see the associated lexinvariant language model as implicitly learning to approximate an in-context Bayesian deciphering process, i.e. inferring a probability distribution over possible lexical permutations based on seen tokens, with the language modeling prior:

$$\begin{aligned} & p'(x_{n+1}|x_1,...,x_n) \\ =& \sum_{\pi} \underbrace{p(\pi^{-1}(x_{n+1})|\pi^{-1}(x_1),...,\pi^{-1}(x_n))}_{\text{language modeling}} \underbrace{\mathcal{P}(\pi|x_1,...,x_n)}_{\text{inferring lexical permutation}} \end{aligned} \tag{6}$$

As shown above, p' can be reduced to two parts, where the first part is normal language modeling and the second part is the probability distribution of lexical permutations based on seen tokens. So the lexinvariant language model is implicitly learning to model $\mathcal{P}(\pi|x_1,...,x_n)$.

We can make this approximate in-context Bayesian deciphering explicit by training a small probe to predict $\mathcal{P}(\pi|x_1,...,x_n)$ given the internal representation of the lexinvariant language model. We will show that this indeed recovers π reasonably accurately in the experiment section.

C. Setup

Architecture. For all experiments, we use decoder-only Transformer architecture with T5 relative position bias (Raffel et al., 2022). We use models with 150M parameters, with 12 layers, 8 heads, head dimension 128, and MLP dimension 4096.

Training. We use the Adafactor optimizer (Shazeer & Stern, 2018), with a cosine decay learning rate schedule (Hoffmann et al., 2022) from 0.01 to 0.001 based on preliminary experiments. We train the models from scratch for 250K steps on all the settings, with 512 sequence length and 64 batch size. We ran all of our experiments on 8 TPU cores. Our models are implemented in JAX (Bradbury et al., 2018).

Datasets. For datasets, we mainly use the Pile (Gao et al., 2020), a large open-source corpus that contains text collected from 22 diverse high-quality sources. We also run experiments on two additional datasets to explore their effects on the behavior of lexinvariant models: Wiki-40B (Guo et al., 2020), which contains high quality processed Wikipedia text in 40+ languages, and GitHub (subset of the Pile), which contains code from GitHub repositories with more than 100 stars and less than 1GB files.

D. Model Architecture Details

In addition, we add a learnable scaling and bias parameter to the result of the embedding layer, so that the model can still learn to scale it as needed.

E. Convergence on other datasets

Figure 4 shows the perplexity of lexinvariant LMs across the three different datasets. Note that Github converges significantly faster than standard Engish text like Wiki-40B, since code is more structured and easier to decipher the token permutation.

F. Recovering Substitution Ciphers

Here we show that lexinvariant LM is implicitly performing Bayesian in-context deciphering by testing its ability to recover cipher keys (e.g. Figure 5a) from character-level substitution ciphers, e.g. `uC; kvR5W 4mfzd @f| Svcgn fw;m uCRmu;;d ]%~} :fBn`. For the lexinvariant LM, this cipher text is perceived as the same as `the quick brown fox jumps over thirteen lazy dogs`, due to the lexinvariant property. It will then proceed to complete the cipher text with `%d: uC; @f|` with the same probability as it will complete the normal text with `and the fox`.

Because of this, we cannot directly read out the distribution of possible cipher keys $\mathcal{P}(\pi|x_1,...,x_{n-1})$ implicitly inferred by the lexinvariant LM. To do this, we train a small two-layer MLP probe on top of a frozen trained lexinvariant LM. For each training sequence, we first embed the input sequence with a randomly sampled token embedding E as described in section 2.2 and obtain the hidden activation of the final layer generated by the frozen lexinvariant LM. Then, we pass this activation through the two-layer MLP probe. Finally, instead of decoding the output activations to classification logits with the same E as in the lexinvariant

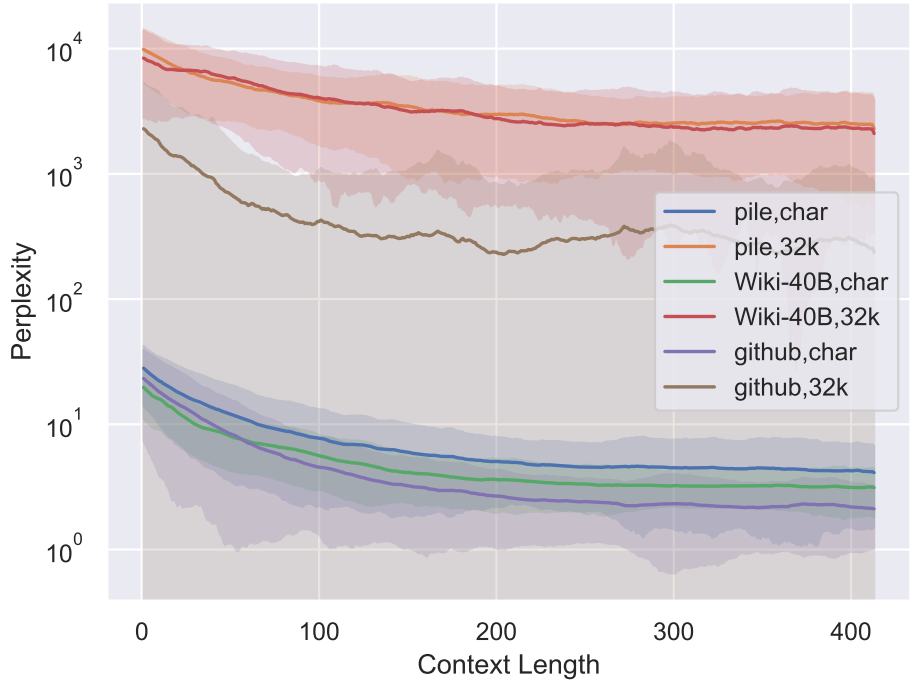
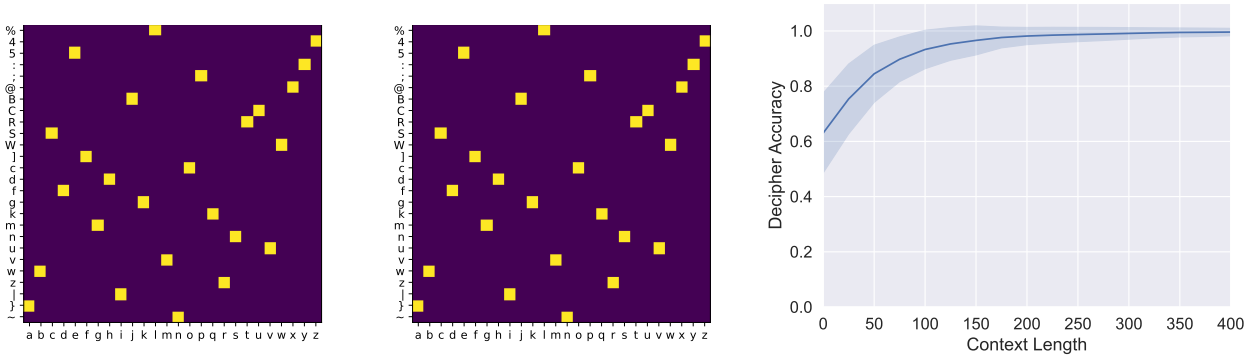


Figure 4. Smoothed Token Perplexity over the Pile, Wiki-40B and Github, with character-level and T5 default vocab

LM, we instead use another learnable non-randomized token embedding matrix E' so that the probe can recover the deciphered token with stable token embeddings. Overall, we train the probe jointly with this embedding matrix E' to predict the current token. Effectively, we are training the probe to decipher the current token using the representation provided by the lexinvariant LM. We train the probe over the same corpus as the original lexinvariant LM for 10k steps. With this probe, we can directly visualize $\mathcal{P}(\pi^{-1}(x_n)|x_1,...,x_n)$ inferred by the lexinvariant LM, which is effectively one row in the permutation matrix representing π.

Now we can use this probe to explicitly recover the cipher key. An example ground truth cipher key that we want to recover is shown in Figure 5a. Note that although the substitution cipher is only among lowercase letters, the character-level lexinvariant model we use assumes that all permutations among the 128 characters are possible , making the deciphering even more challenging.



(a) Ground truth (b) Majority vote prediction (c) Cipher key prediction accuracy

Figure 5. (a) (b): Cipher key matrix, where the vertical axis shows the cipher characters and the horizontal axis shows the deciphered letters. The highlighted entries show the correspondences between cipher characters and the actual letters, e.g. `%` deciphers to `l`. (c): Cipher key prediction accuracy, averaged across 1000 input sequences. Context length denotes the start index of the window.

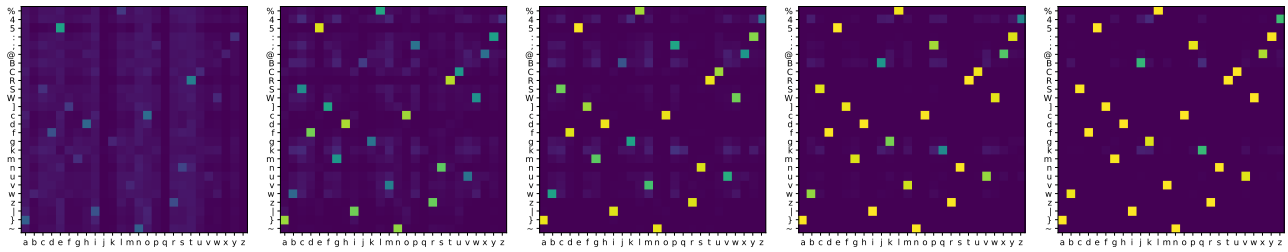


Figure 6. Predicted cipher key for windows of size 50, at indices 0, 50, 100, 200, and 400. Generated using temperature of $T = 1$.

Concretely, we first input ciphertext through the frozen lexinvariant LM with the probe to produce a deciphered sequence. We then select a window of size 100 in the middle of the sequence and perform a majority vote over the corresponding deciphered tokens of each cipher token seen in this window. This essentially produces a predicted cipher key matrix for each window, and we can measure its precision against the ground truth. As shown in Figure 5c, such a cipher key prediction generally has increasingly higher precision as the window is selected later in the context, and it becomes near-perfect by the end of the sequence. Specifically, the cipher key matrix produced by the last window has an average precision of 99.6% over 1000 input sequences.

Finally, we aggregate over the last window of the 1000 sequences to recover a full cipher key, in case certain letters never appear in the last window of certain sequences. We again recover a full cipher key via majority vote. In Figure 5b, we show the highly accurate predicted cipher key recovered from ciphertext produced using the example ground truth cipher key in Figure 5a.

To perform a more detailed analysis showing the Bayesian deciphering process of the lexinvariant model, we use the logits of the probe to recover the predicted distribution of the cipher key $\mathcal{P}(\pi|x_1,...,x_{n-1})$. Instead of taking the majority vote of the predicted decipher tokens in the window, we take the mean of logits predicted for each ciphered token. This essentially gives a locally averaged predicted distribution of cipher key matrices. Specifically, the cipher key matrices are generated across windows of 50 characters, and the probabilities are averaged over 1000 input sequences encoded using the same ground truth cipher. As shown in Figure 6, the predicted distribution of cipher key matrix becomes sharper as the prefix becomes longer.

G. In-context Bayesian Deciphering Examples

Here, we show several qualitative examples of in-context Bayesian deciphering. We first show how the lexinvariant LM maintains uncertainty over possible lexical permutations while iteratively updating them at each index, using examples from a character-level lexinvariant model. Then, we also show an example of semantic in-context deciphering with a 32K vocabulary lexinvariant model, where the meaning of a novel word is inferred relative to common words in-context.

G.1. Uncertainty over Lexical Permutations

In Figure 7a, we input the following ciphered sequence to the frozen character-level lexinvariant LM with the probe: "`I saw lots of people in town today, walking and talking around me. I greeted my friend Alice and my classmate Alex. I saw a guy, Joe, walking outside carrying a zat. Joe's zat was taken off zy wind. Today's wind was strong, so Joe's zat flew zackward. Joe lost Joe's zat for good. Joe will miss Joe's zat.`" For each instance of `z` in the sequence, we display the predicted deciphering of that instance as a row of probabilities across non-cipher letters `a-z`.

The lexinvariant model starts off assuming uniform probability for all possible lexical permutations π. After seeing more and more text, the lexinvariant model quickly realizes that `z` only has a few main plausible decipherings (`b`, `h`, `c`, `m`). Eventually, the lexinvariant model is able to narrow the possibilities down to `z` maps to `b` near the end of the sequence. The predicted probabilities shift with the seen context accordingly, demonstrating an example of how the predicted cipher key is iteratively updated at each index.

Figure 7b shows another example with a similar set up, but with text: "`I saw a man in the pazk with a zat. The man was walking with the zat zight beside him. I've nevez seen anything like that befoze.`" While context initially suggests that `z` may be deciphered as `c`, it becomes clear that `z` must correspond to `r` after the appearance of "`right`". The disambiguation is reflected in the depicted probabilities.

In Figure 7c and 7d, we show two deciphering examples over code. We consider two code examples in which it is initially ambiguous whether the character `z` deciphers to `:` or {. The ambiguity is eventually resolved by the use of Python-like or Java-like syntax.

G.2. Semantic Deciphering

In addition to character-level deciphering, we show examples of semantic deciphering with the larger vocabulary of 32k. Although the lexinvariant LM could not possibly figure out the true lexical permutation among 32k tokens using a small 512 context, it is possible to construct a simple context that repetitively uses simple words so that these words are easier to decipher. Then the lexinvariant LM can decipher the approximate semantics of the rare symbols relative to other easier-to-decipher words.

One example is the following: given the prompt `'crash!' 'aaah!' i looked up from my cup of coffee. 'crash!' - that was the cafe window. and 'aaah!'` [... more text...] `what one here is a drink ⏎ - restaurants ⏎- music ⏎- coffee ⏎ - father ⏎ the one here that drink is`, where the word `coffee`, `music`, and `father` all only appear once before the question and `restaurants` appeared 4 times, the model is able to correctly answer that `coffee` is drinkable. See the full example in Appendix K.

H. Synthetic Reasoning Tasks

As discussed in the introduction, lexical flexibility is correlated with in-context reasoning performance, as demonstrated by existing large LMs. Thus, we study whether the lexinvariant model also learns in-context reasoning capabilities through the challenging lexinvariant training.

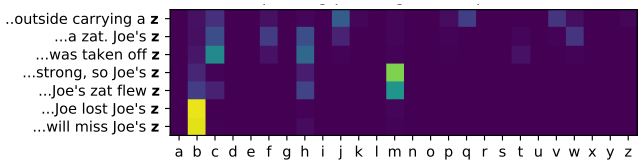
Specifically, we measure the performance of lexinvariant models over two pure in-context symbol manipulation tasks: LookUp, where the task is to predict the next token based on the given lookup table, e.g. `A->2⏎C->4⏎G->5⏎C->` (should predict `4` here); and Permutation, where the task is to permute an arbitrary subsequence of the given sequence the same way as in the given few demonstrations, e.g. `A 2 C->C A⏎4 1 D->` (should predict `D 4` here). In each of the tasks, the symbols are randomly sampled from the vocabulary so that we measure the pure reasoning ability independent from any knowledge of specific words. We measure the model performance in terms of generated token accuracy over 1000 examples. The results are shown in Table 1. As shown in the table, the lexinvariant models achieve drastically higher accuracy, with an average of 4X improvement.

Table 1. Accuracy over synthetic reasoning tasks.

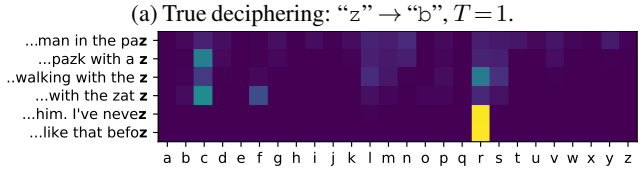
Dataset	Vocab	LookUp Acc		Permutation Acc	
		Standard	LI	Standard	LI
Pile	char	48.50	91.80	27.66	59.35
	32k	21.45	92.10	22.84	55.63
Wiki-40B	char	38.25	59.70	20.77	60.51
	32k	8.75	59.35	9.94	50.91
Github	char	42.40	86.65	21.03	71.59
	32k	4.25	80.20	8.59	67.39

I. Regularizing Language Models with Lexinvariance

Although lexinvariant LM has various interesting properties , it is not suitable for practical tasks since it would require the context to be extremely long so that all required words and knowledge are defined in the context. Here, we explore how to construct more practical semi-lexinvariant LMs that maintain some properties of lexinvariant LMs via regularization. We



(a) True deciphering: "z" → "b", $T = 1$.



(b) True deciphering: "z" → "r", $T = 1$.

(c) True deciphering: "z" → "{", $T = 2$.

(d) True deciphering: "z" → ":", $T = 3$.

Figure 7. Probe predictions for deciphering "z" at each occurrence of "z" in context.

emphasize that this exploration is intended to be illustrative rather than directly improving state-of-the-art.

Instead of using random Gaussian embedding matrices in place of a learned embedding matrix entirely, we can use random embeddings for only some of the tokens in each sequence, while others use the learned embedding. This means that the learned LM assumes that certain tokens have stable meanings but not others, which can be seen as a form of regularization towards lexinvariance. Specifically, we randomly select tokens to randomize based on a Bernoulli distribution, which can essentially be seen as a form of dropout on token embeddings. On the BIG-bench tasks, we found that a model with dropout rate $p = 0.2$ for randomization was 25% more likely to improve performance than to harm performance when evaluated with three shots, relative to a comparably-sized LM, with improvements especially over retrieval type of tasks. See full details in the Appendix M.

More broadly, this regularization view could potentially bring the benefit of lexinvariant LMs to practical applications. For example, the regularization could improve 1) the robustness of LMs by making them less sensitive to adversarial attacks or noise in the input data, 2) generalization across different languages or domains by being less tied to specific lexical items and more prone to learn the shared language structure, and 3) reasoning over more realistic tasks as we have started to explore with BIG-Bench. These areas are promising directions for future work to explore.

J. Code Deciphering Full Examples

Java:

```
binary_search()z
  if (high >= low)z
    mid = (high + low) / 2;
    if (arr[mid] == x)
      return mid;
    if (arr[mid] > x)z
      high = mid - 1;
      return binary_search();
    } elsez
      low = mid + 1;
      return binary_search();
    }
  } elsez
    return -1;
  }
}
void func2()z
```

Python:

```
binary_search()z
  if (high >= low)z
    mid = (high + low) // 2
    if (arr[mid] == x)z
      return mid
    if (arr[mid] > x)z
      high = mid - 1
      return binary_search()
    elsez
      low = mid + 1
      return binary_search()
  elsez
      return -1
def func2()z
```

K. Semantic Deciphering Full Example

```
'crash!' 'aaah!' i looked up from my cup of coffee. 'crash!' - that was the cafe
window. and 'aaah!' - that was kate. people in the cafe shouted. kate and i ran
to the window. there was no one there. then i turned to kate and put my arm around
her. 'are you all right?' i asked. 'yes,' she said. 'i think so.' 'what is it?'
some one shouted and a short red-faced man ran into the room. the man took my arm.
'matt! what are you doing to kate?' he asked. 'nothing, papa,' kate replied. 'it
wasn't him. it was from out in the street.' the red-faced man looked at the window
and then at me. he turned to his daughter. 'are you ok, kate?' he asked. kate
gave him a little smile. 'yes, i think i am, papa,' she said. then her father
spoke to me. 'sorry, matt. i heard kate and i thought...' 'that's ok, paolo,'
i answered. it was ok. you see, this is soho, in the centre of london. in the
day it's famous for music and films. at night people come and eat and drink in the
restaurants. expensive restaurants and cheap restaurants; italian restaurants and
chinese restaurants. and day and night there are internet cafes like the web cafe.
in soho you can buy any thing and any one. there are lots of nice people in soho.
but there are also lots of people who are not very nice. i know because i live and
work here. i often take a drink to a shop or cafe. i'm not rich and famous. and
i don't know a lot. but i do know soho. what one here is a drink - restaurants -
music - coffee - father the one here that drink is
```

Example prediction of the lexinvariant with 32k vocabulary train on the Pile:

```
 - coffee. and i
```

L. Synthetic Reasoning Task

Table 2 shows a variant of the synthetic reasoning task results in Subsection 1, where the symbols are instead sampled proportion to the token frequencies. Although the improvement still generally holds, the standard LM with character-based vocabulary becomes significantly better. We believe that this is because the model can get a significant advantage by guessing among the most common letter.

Dataset	Vocab	LookUp Acc		Permutation Acc	
		Standard	LI	Standard	LI
Pile	char	72.80	90.95	40.63	60.47
	32k	61.20	90.95	40.55	54.55
Wiki-40B	char	75.55	63.45	42.71	59.86
	32k	41.05	57.95	26.81	51.86
Github	char	66.00	86.75	36.62	70.77
	32k	59.25	78.45	37.46	65.04

Table 2. Synthetic Reasoning Tasks (adjusted for token frequencies)

M. Language Models Regularized with Lexinvariance and BIG-bench Results

As described in the main paper, we implement a lexinvariance regularized Model in a way similar to embedding dropout. Note that one problem in implementing it naively by using random Gaussian embeddings and learned embedding in a mixture is that the two would become quickly distinguishable from each other during training since learned embeddings often have larger norms, allowing the model simply ignore the randomized tokens. So instead of using random Gaussian embedding matrices in place of a learned embedding matrix, we explored another approach for training a lexinvariant regularized LM: training a standard LM with learnable embedding matrix over sequences partially applied with a random token permutation $B_p(x_1,\pi),...,B_p(x_1,\pi)$, where $B_p(x_i,\pi) = \pi(x_i)$ with probability p and $B_p(x_i,\pi) = x_i$ with probability $1-p$. Since each

token can be remapped to any other token with equal chance, the produced model should ideally also be lexinvariant when $p = 1$, though with no strict guarantees. In practice, we found the models trained this way behave very similarly to models with random Gaussian embedding.

We evaluate our model over BIG-bench tasks where the language model performance scales well, and we prioritize evaluating generative tasks over multiple-choice tasks. Tasks we evaluated on:

gre reading comprehension.mul, linguistics puzzles.gen, linguistics puzzles.gen, rhyming.gen, tellmewhy.gen, simple arithmetic multiple targets json.gen, simple arithmetic json subtasks.gen, disfl qa.gen, arithmetic.gen, bridging anaphora resolution barqa.gen, matrixshapes.gen, sufficient information.gen, logical args.mul, novel concepts.mul, code line description.mul, unnatural in context learning.gen, unit interpretation.mul, english proverbs.mul, general knowledge.mul, geometric shapes.gen, human organs senses.mul, contextual parametric knowledge conflicts.gen, crass ai.mul, auto categorization.gen, penguins in a table.gen, hindu knowledge.mul, english russian proverbs.mul, modified arithmetic.gen, cryobiology spanish.mul, evaluating information essentiality.mul, intent recognition.mul, understanding fables.mul, figure of speech detection.mul, empirical judgments.mul, simple ethical questions.mul, swahili english proverbs.mul, language identification.mul, phrase relatedness.mul, nonsense words grammar.mul, undo permutation.mul, object counting.gen, identify odd metaphor.mul, elementary math qa.mul, social iqa.mul, parsinlu qa.mul, metaphor understanding.mul, timedial.mul, causal judgment.mul, list functions.gen, implicatures.mul, date understanding.mul, codenames.gen, fact checker.mul, physics.mul, abstract narrative understanding.mul, emojis emotion prediction.mul, metaphor boolean.mul, strategyqa.gen, ascii word recognition.gen, auto debugging.gen, cause and effect.mul, conlang translation.gen, cryptonite.gen, cs algorithms.mul, dyck languages.mul, gender inclusive sentences german.gen, hindi question answering.gen, international phonetic alphabet transliterate.gen, irony identification.mul, logical fallacy detection.mul, movie dialog same or different.mul, operators.gen, paragraph segmentation.gen, parsinlu reading comprehension.gen, repeat copy logic.gen, rephrase.gen, simple arithmetic json.gen, simple arithmetic multiple targets json.gen, sports understanding.mul, word unscrambling.gen, hyperbaton.mul, linguistic mappings.gen, anachronisms.mul, indic cause and effect.mul, question selection.mul, hinglish toxicity.mul, snarks.mul, vitaminc fact verification.mul, international phonetic alphabet nli.mul, logic grid puzzle.mul, natural instructions.gen, entailed polarity.mul, list functions.gen, conceptual combinations.mul, goal step wikihow.mul, logical deduction.mul, conlang translation.gen, strange stories.mul, odd one out.mul, mult data wrangling.gen, temporal sequences.mul, analytic entailment.mul, disambiguation qa.mul, sentence ambiguity.mul, swedish to german proverbs.mul, logical sequence.mul, chess state tracking.gen, reasoning about colored objects.mul, implicit relations.mul, riddle sense.mul, physical intuition.mul, simple arithmetic json multiple choice.mul, geometric shapes.gen, gem.gen, simp turing concept.gen, common morpheme.mul, qa wikidata.gen, international phonetic alphabet transliterate.gen, similarities abstraction.gen, rephrase.gen, emoji movie.gen, qa wikidata.gen, word sorting.gen, emoji movie.gen, qa wikidata.gen, periodic elements.gen, hindi question answering.gen

Bellow, we plot the net percentage of tasks improved/deproved in each of the BIG-bench categories, out of the tasks that are changed by at least a threshold amount.

N. Compute

We use one TPU v3-8 for all our pretraining runs. It takes approximately 23 hours for each pretraining run.

O. Related Work

O.1. Symbol Grounding

Beyond a modeling choice, the main question of our paper (that being whether an LM can learn language without a stable token representation) is also analogous to the symbol grounding problem: Can meaning be acquired when symbols are not even grounded stably, i.e. they can be mapped to completely random meanings in different sequences? It has long been argued by the symbol grounding literature that symbolic representations must be grounded bottom-up in nonsymbolic representations (Harnad, 1990), with famous arguments like Searle's Chinese room. And this leads to an ongoing debate on whether LMs can learn meaning purely from large amounts of text, without grounding to any real-world objects (Bender et al., 2021). Conceptually, lexinvariant LM is one step further away from physical grounding. We show that they can still infer the meaning of symbols based on lexical structures within the context.

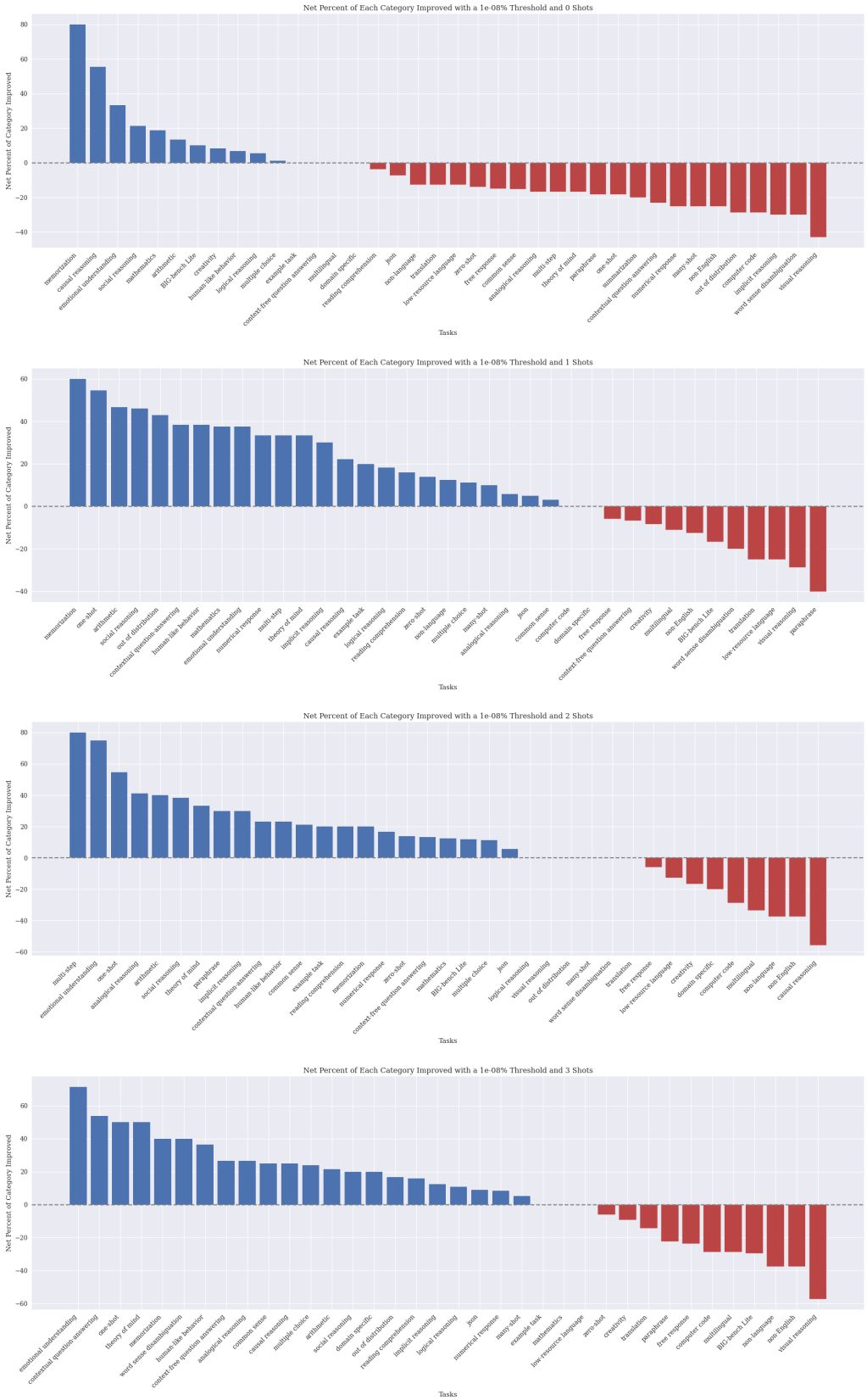


Figure 8. BIG-bench results with 0,1,2 and 3 shots.

O.2. Byte-level T5

There is existing work on absorbing tokenization completely into part of language modeling by using extremely small tokens, such as Byte-level T5 (Xue et al., 2021). In the extreme, such a model would become closer and closer to lexinvariant LM, since bytes or bits have almost no stable meaning, so their embeddings are likely not used for prediction. In this paper, we study general lexinvariant LMs with the lexinvariant property baked in and without requiring specific tokenizers.

O.3. Group invariances and Data augmentation

Our implementation of lexinvariant LMs can be seen as performing a form of very aggressive data augmentation, where we randomize the identity of each token in each sequence. From this perspective, it is somewhat similar to the data recombination in (Jia & Liang, 2016; Akyürek et al., 2020) and augmentation of named entities in (Raiman & Miller, 2017), where certain parts of the sentence are swapped with other words while still maintaining the original grammatical structure. In contrast to these augmentations, the training for our lexinvariant LMs completely swaps out all parts of the input text.

O.4. Deciphering Substitution Cipher using LMs

In general, solving substitution ciphers, where the cipher key is a permutation of the original alphabet, is a NP-hard problem when only having access to LMs that can assign probabilities to sequences (Nuhn & Ney, 2013). There have been several works focusing on solving substitution ciphers using LMs, including approaches from searching over the permutation space guided by LMs' scores (Hauer et al., 2014) to training a seq-to-seq model directly to perform deciphering as translation (Aldarrab & May, 2020). Although our work does not focus specifically on the task of deciphering substitution ciphers, we show that our lexinvariant model can efficiently perform in-context deciphering as a byproduct of language modeling.

O.5. Reasoning

It has been shown that large language models acquire surprising in-context reasoning capabilities (Brown et al., 2020; Liang et al., 2022; Srivastava et al., 2022). Many of them are related to lexical flexibility through training for purely next-token prediction, such as modified arithmetic, data reformatting, and redefining single word etc. However, LLMs also memorize an enormous amount of knowledge along the way, which is often unnecessary. This work can also be seen as an exploration of whether a (semi-)lexinvariant LM can discount knowledge and prioritize learning the diverse structural reasoning patterns in language, therefore achieving the strong reasoning capability of LLMs with a smaller model.

P. Broader Impacts

Our work primarily provides a scientific exploration and understanding of the properties of lexinvariant language models. More broadly, these properties could potentially help improve the robustness, generalizability, and reasoning ability of LMs in the future works. In general we don't foresee more specific negative societal impacts from this work other than general misuse of language models.