In-Context Learning with Long-Context Models: An In-Depth Exploration

Amanda Bertsch[γ]
abertsch@cs.cmu.edu

Maor Ivgi [τ]
maor.ivgi@cs.tau.ac.il

Uri Alon[γ*]
urialon@cs.cmu.edu

Jonathan Berant[τ]
joberant@cs.tau.ac.il

Matthew R. Gormley[γ]
mgormley@cs.cmu.edu

Graham Neubig[γ]
gneubig@cs.cmu.edu

[γ] Carnegie Mellon University [τ] Tel Aviv University

Abstract

As model context lengths continue to increase, the number of demonstrations that can be provided in-context approaches the size of entire training datasets. We study the behavior of in-context learning (ICL) at this extreme scale on multiple datasets and models. We show that, for many datasets with large label spaces, performance continues to increase with hundreds or thousands of demonstrations. We contrast this with example retrieval and finetuning: example retrieval shows excellent performance at low context lengths but has diminished gains with more demonstrations; finetuning is more data hungry than ICL but can sometimes exceed long-context ICL performance with additional data. We use this ICL setting as a testbed to study several properties of both in-context learning and long-context models. We show that long-context ICL is less sensitive to random input shuffling than short-context ICL, that grouping of same-label examples can negatively impact performance, and that the performance boosts we see do not arise from cumulative gain from encoding many examples together. We conclude that although long-context ICL can be surprisingly effective, most of this gain comes from attending back to similar examples rather than task learning.[2]

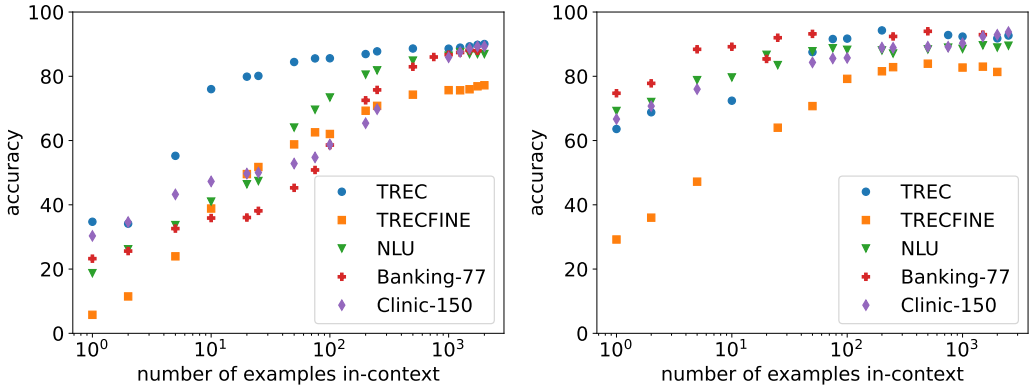
1 Introduction

When a few examples are provided in-context, large language models can perform many tasks with reasonable accuracy. While questions remain about the exact mechanism behind this phenomena (Min et al., 2022b; von Oswald et al., 2023), this paradigm of *in-context learning* (ICL) has seen widespread adoption in both academic and industry applications, thanks to its ease of implementation, relatively small computational cost, and ability to reuse a single model across tasks.

However, most work in this area has focused on models where the maximum number of demonstrations is severely limited by context length. As more and more methods are developed to adapt language models to extreme context lengths (Deepmind (2024); Fu et al. (2024), *inter alia*), in-context learning over large quantities of data becomes a potential alternative to finetuning. The properties of ICL in this regime are not well-understood; additionally, as the cost of inference over many thousands of tokens can be steep, the efficiency and performance tradeoff between many-shot ICL and finetuning on the same data is complex.

[*] Now at Google DeepMind

[2] Data and code are available at `https://anonymous.4open.science/r/long-context-icl/`



(a) Using randomly selected examples.

(b) Using retrieved examples.

Figure 1: The performance increases with more demonstrations far beyond the context window of the base Llama-2. Results are on Fu et al. (2024)'s long-context finetuned Llama-2-7b model, using a context of up to 80K tokens.

We conduct a systemic study of long-context in-context learning. Namely, we consider: a) the performance of prompting the base model naively, b) retrieving examples to use in-context for each test example, c) comparison to finetuning the base model, and d) using models trained to adapt to longer contexts. Performance continues to increase past 2000 demonstrations (see Figure 1), approaching and sometimes *exceeding* the performance of models finetuned (with LoRA (Hu et al. 2022)) on thousands of examples from the same dataset (§ 2).

We find that, as the number of demonstrations in-context increases to extreme values, the behavior of ICL shifts (§ B). In-context learning becomes less sensitive to example order, and the benefits of retrieval over using a random set of demonstrations diminishes — allowing the use of a single set of demonstrations, encoded once through the model and cached, rather than re-encoding a custom set of demonstrations for each inference example.We demonstrate that long-context ICL is strongly impacted by grouping examples of the same label. We also find that the effectiveness of long-context ICL is not because of the continual refinement of a decision boundary during encoding but because of the retrieval from more relevant examples (§ D). Our work furthers the understanding of in-context learning and suggests that long-context ICL is a strong alternative to retrieval and finetuning.

2 Long-context ICL

We consider 5 classification datasets: TREC (Hovy et al., 2001), TREC-fine (Hovy et al., 2001), NLU (Xingkun Liu & Rieser, 2019), Banking-77 (Casanueva et al., 2020), and Clinic-150 (Larson et al., 2019). We evaluate across 4 models– 3 versions of Llama 2 finetuned for various amounts of context and Mistral Instruct. We then take three common settings– naively sampling a fixed subset to use in-context, retrieving relevant data for each example at inference time, or finetuning on the dataset. Appendix A contains more details about the experimental settings.

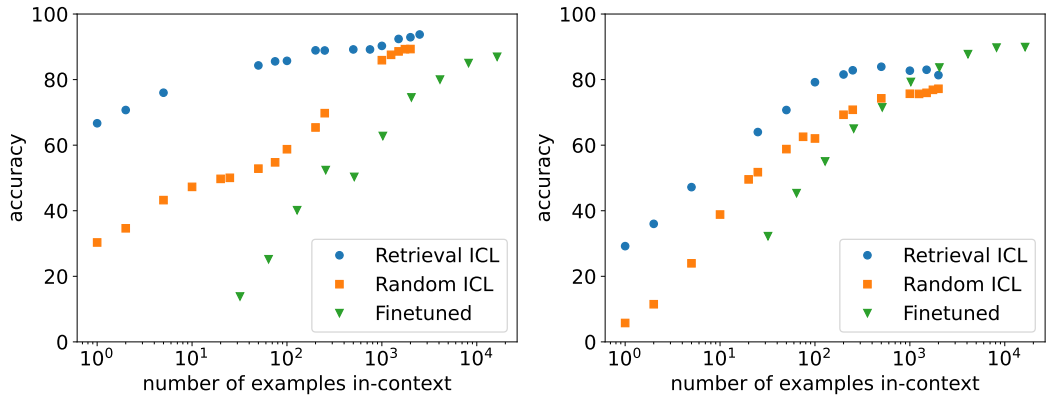
2.1 In-context results

Scaling up ICL to many examples leads to surprisingly strong results Figure 1 and Table 1 show the performance of models in both in-context learning settings. Scaling up in-context learning from 10 to 1000 demonstrations results in accuracy gains of up to *50.8* points (an average of 36.8 points across the 5 datasets).

Longer context lessens the importance of carefully selecting in-context examples Retrieving relevant examples for each test set example far outperforms using a randomly selected subset in the short-context regime. This is true even if the order of retrieved examples is shuffled (rather than

Dataset	Llama2	Llama2-32k	Llama2-80k	Mistral
Randomly selected				
TREC	**82.32** / 80.52	**93.12 / 93.12**	**90.04 / 90.04**	**87.28** / 85.00
TREC-fine	**61.40 / 61.40**	**75.56** / 75.08	**77.20 / 77.20**	**72.68** / 70.48
NLU	**76.88 / 76.88**	**85.04** / 85.00	**87.52** / 86.92	**86.44 / 86.44**
Banking-77	**56.36 / 56.36**	**82.44 / 82.44**	**88.08** / 87.96	**86.76** / 86.68
Clinic-150	**60.92 / 60.92**	**84.40 / 84.40**	**89.32 / 89.32**	**90.56 / 90.56**
Retrieval				
TREC	**90.80** / 85.64	**94.84** / 94.64	**94.28** / 92.68	**90.80 / 90.80**
TREC-fine	**78.80 / 78.80**	**83.88** / 81.12	**83.92** / 81.36	**80.80** / 79.60
NLU	**90.00** / 88.40	**89.80 / 89.80**	**89.64** / 89.52	**90.40** / 89.20
Banking-77	**93.20** / 92.40	**94.32 / 94.32**	**94.00** / 92.96	**93.20 / 93.20**
Clinic-150	**87.60 / 87.60**	**89.84 / 89.84**	**93.76 / 93.76**	**93.20** / 92.40

Table 1: For all datasets, performance of ICL continues to increase with additional demonstrations. The results are the best accuracy (left) and accuracy at maximum data (right) for each model. Bold indicates the best performance for that model/dataset pair.



(a) Clinic-150

(b) Trecfine

Figure 2: Comparing retrieval ICL, random selection ICL, and finetuning on two representative datasets. Finetuning sometimes, but not always, exceeds ICL at high numbers of demonstrations. Note that, while retrieval ICL uses the listed number of examples in context, it assumes access to the larger test set to draw examples from (Perez et al., 2021). See Appendix G for results on other datasets.

.

ordered by relevance).[3] However, adding additional examples does continue to slightly improve performance; this is especially surprising because, after all examples with non-trivial lexical overlap are retrieved the remaining examples are randomly selected.

While retrieval continues to outperform random selection on most datasets, the importance of the selection strategy diminishes with additional examples. On Banking-77, the dataset where retrieval is most beneficial, the performance gain from retrieval drops from 51.5 points at 1-shot ICL to 4.9 points at 1500-shot ICL. In the long context regime, using the more computationally efficient but less effective strategy of a single randomly selected set of demonstrations is more feasible; the performance penalty for doing so is never more than 5 points of improvement, and as low as 1.8 points (for 2000-shot ICL on TREC).

[3]We perform three random shuffles of the retrieved inputs and test for difference in distribution from the original results; this does not significantly change performance for any dataset (2-sided t-test, $p < 0.05$).

Why does long-context ICL help? We analyze this phenomenon in detail in Appendix D, and find that the performance of long-context ICL is largely unaffected by encoding the input in blocks, suggesting that the model is not aggregating a decision boundary over long context but instead attending to relevant examples in-context to make the prediction.

2.2 Comparison with finetuning

While we have demonstrated that in-context learning with hundreds or thousands of examples is effective, this amount of data is also appropriate for finetuning a model. Finetuning has higher upfront cost but allows for reduced inference-time cost. We compare in-context learning with the popular parameter-efficient finetuning (PEFT) strategy LoRA (Hu et al., 2022).

PEFT is more data-hungry than ICL– especially ICL with retrieval When a relatively small set of examples is available, ICL generally outperforms LoRA finetuning on the same model.[4] For most datasets, finetuning performance never exceeds long-context ICL performance even with additional examples (e.g. Figure 2a). The exceptions are on TREC and TREC-fine, where finetuning outperforms ICL at the highest numbers of examples (but continues to underperform at low numbers of examples) (e.g. Figure 2b). Generally, the datasets with larger label spaces show the least strong finetuning performance, likely because these are more open-ended classification problems and require more data to train the classifier.

For some datasets, PEFT does win out overall– finetuning with more examples than even the 80k model can fit in-context does result in higher performance. In datasets where PEFT performance never exceeds ICL performance, it nevertheless has dramatically reduced inference costs for similar performance; thus, finetuning on 4096 examples may still be preferable to prompting with 1000 if efficiency of inference is a major priority. This is because, even if demonstration encodings can be cached across inference examples, cross-attention to a long context of demonstrations is expensive.

3 Conclusion

We have demonstrated that ICL with large demonstration sets can be surprisingly effective, and shed light on a few surprising properties in its behavior. Namely, long-context ICL exhibits a reduced dependence on example selection, relatively stable performance with respect to example order, and performance often approaching or exceeding parameter-efficient finetuning on the same data, all properties that make this an appealing option for a variety of tasks. We have also shown that long-context ICL's effectiveness is largely due to retrieval from the long context during prediction, rather than cross-attention within the large demonstration set during encoding.

Our work also highlights that our understanding of ICL remains incomplete. Though much work has studied the potential mechanisms behind ICL, these works have largely focused on simple tasks with small (< 10 examples) demonstration sets; as our work demonstrates that ICL's properties shift in the long context regime, more work is necessary to validate hypotheses about ICL at larger scales.

While prior work has focused on two strategies for performing inference on a new task– either finetuning on task-specific data or selecting a subset of that data to use in-context– our results points to a potential third paradigm: adapting the *model* to fit as much of that data in-context as possible, caching and reusing the encoding of the long demonstration set. While finetuning with full datasets is still a powerful option if the data vastly exceeds the context length, our results suggest that long-context ICL is an effective alternative– trading finetuning-time cost for increased inference-time compute. As the effectiveness and effiency of using very long model context lengths continues to increase, we believe long-context ICL will be a powerful tool for many tasks.

Acknowledgments

We would like to thank Vijay Viswanathan, Sewon Min, Akari Asai, Xiang Yue, and Simran Khanuja for useful discussions about this work.

[4]Note that some prior results have showed strong PEFT performance in the few-example setting on different tasks; see Appendix F for more discussion.

This work was partially supported by The Yandex Initiative for Machine Learning, the Len Blavatnik and the Blavatnik Family foundation, and the European Research Council (ERC) under the European Union Horizons 2020 research and innovation programme (grant ERC DELPHI 802800). AB was supported by a grant from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE2140739. MI also acknowledges the support of the Israeli Council of Higher Education. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Stephanie Chan, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning, 2024.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J. Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, Jamie Sully, and Alex Hernandez. Many-shot jailbreaking, 2024. URL https://www-cdn.anthropic.com/af5633c94ed2.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. Buffet: Benchmarking large language models for few-shot cross-lingual transfer, 2023.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. Unlimiformer: Long-range transformers with unlimited length input. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6f9806a5adc72b5b834b27e4c7c0df9b-Paper-Conference.pdf.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Gregory Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Iq0DvhB4Kf.

Necva Bölücü, Maciej Rybinski, and Stephen Wan. impact of sample selection on in-context learning for entity extraction from scientific writing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.338. URL https://aclanthology.org/2023.findings-emnlp.338.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023.

Hyunsoo Cho, Hyuhng Joon Kim, Junyeob Kim, Sang-Woo Lee, Sang goo Lee, Kang Min Yoo, and Taeuk Kim. Prompt-augmented linear probing: Scaling beyond the limit of few-shot in-context learners, 2023.

Google Deepmind. Our next-generation model: Gemini 1.5, 2024. URL https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#build-experiment.

Gilad Deutch, Nadav Magar, Tomer Bar Natan, and Guy Dar. In-context learning and gradient descent revisited, 2024.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context, 2024.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models, 2024.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. Prototypical calibration for few-shot learning of language models, 2022.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1, 000 examples. *ArXiv preprint*, 2022. URL https://arxiv.org/abs/2212.06713.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001. URL https://aclanthology.org/H01-1069.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 2023. doi: 10.1162/tacl_a_00547. URL https://aclanthology.org/2023.tacl-1.17.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *ArXiv preprint*, 2023. URL https://arxiv.org/abs/2312.03732.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131.

Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.78. URL https://aclanthology.org/2023.acl-long.78.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How long can context length of open-source LLMs truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a. URL https://openreview.net/forum?id=LywifFNXV5.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts?, 2023b.

Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jun Zhang, Zhiyong Wu, and Lingpeng Kong. In-context learning with many demonstration examples, 2023c.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning, 2024.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning, 2024.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context, 2023.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=rBCvMG-JsPd.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00638. URL https://doi.org/10.1162/tacl_a_00638.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. In-context learning for text classification with many labels. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni (eds.), *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.genbench-1.14. URL https://aclanthology.org/2023.genbench-1.14.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.201. URL https://aclanthology.org/2022.naacl-main.201.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.759.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.779. URL https://aclanthology.org/2023.findings-acl.779.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto,

Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5c04925674920eb58467fb52ce4ef728-Abstract.html.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud D. Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows for large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:258686160.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 2009. doi: 10.1561/1500000019.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2023.

Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, and Jinwoo Shin. Hierarchical context merging: Better long context understanding for pre-trained LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ulaUJFd96G.

TogetherAI. Llama-2-7b-32k-instruct - and fine-tuning for llama-2 models with together api, 2023. URL https://www.together.ai/blog/llama-2-7b-32k-instruct.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Szymon Tworkowski, Konrad Staniszewski, Mikoł aj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Mił oś. Focused transformer: Contrastive training for context scaling. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8511d06d5590f4bda24d42087802cc81-Paper-Conference.pdf.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024.

Pawel Swietojanski Xingkun Liu, Arash Eshghi and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, Ortigia, Siracusa (SR), Italy, 2019. Springer. URL http://www.xx.xx/xx/.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models, 2023.

Paiheng Xu, Fuxiao Liu, Zongxia Li, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps, 2023.

Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding, 2024.

Cecilia Ying and Stephen Thomas. Label errors in BANKING77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.insights-1.19. URL https://aclanthology.org/2022.insights-1.19.

LILI YU, Daniel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. MEGABYTE: Predicting million-byte sequences with multiscale transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JTmO2V9Xpz.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://aclanthology.org/D17-1004.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c.html.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training, 2024.

A Additional experimental setup information

A.1 Models used

We compare ICL performance across several long- and short-context models:

1. **Llama2** (Touvron et al., 2023) is a decoder-only model trained with a 4096 context length. We use the non-instruct (non-chat) variant because it is more commonly used as a base model for long-context finetuning and because we observed very similar performance between the chat and non-chat variants in our initial experiments.
2. **Llama2-32k** (TogetherAI, 2023) is a version of Llama-2-7b finetuned by TogetherAI for a 32k context window. We use the non-instruct version.
3. **Llama2-80k** (Fu et al., 2024) is a version of Llama-2-7b finetuned with 80k context and a carefully designed long-document data mixture.
4. **Mistral-7b-v0.2** (Jiang et al., 2023). To verify that the trends we observe are not specific to the Llama series, we use the instruct version of Mistral-v0.2 (the non-instruct model is not publicly available). The trained context length of Mistral-7B-Instruct-v0.2 is 32k tokens.

While all of these models can extrapolate to inputs longer than their trained context length, we restrict the lengths of inputs to fit within the trained context length; this represents the best case performance without the additional confound of the extrapolation strategy.

A.2 Compared settings

We consider three common methods for using a large dataset.

Random sampling ICL We use 10 random shuffles of the training dataset, averaging the results across these shuffles. Across models and across varying numbers of demonstrations in-context, we draw the first n examples from each shuffle. In this setting, the encoding of demonstrations can be performed once and cached across all inference examples.

Retrieval ICL A strong alternative for in-context learning is to retrieve a relevant subset of examples as demonstrations for each test set example. In this setting, we use a BM25 (Robertson & Zaragoza, 2009) retriever with stopwords removed and retrieve the most relevant demonstrations by comparing the test input text to the full demonstration texts. When doing k-shot prompting, if less than k examples are retrieved by the retriever,[5] we randomly sample additional examples until we reach k. We compare putting examples in the order they were retrieved and in three random shufflings. Note this is more computationally expensive than using a random sample as demonstrations, as a new set of retrieved demonstrations must be encoded for each test set example. However, prior work has found that, in some scenarios, retrieval of good examples can make the difference from near-zero to high test accuracy (Levy et al., 2023).

Finetuning We finetune Llama2-7b with a classification head on varying amounts of data from each dataset with several random seeds, and plot performance at convergence on the same held-out test data. We initialize the classification head from the parameters of the pretrained language modeling head by subsampling the values of the first token of each label; this creates a better-than-random initialization for finetuning. For more details on the finetuning procedure, see Appendix H.

A.3 Constrained decoding

For each dataset, we use *constrained decoding* to only produce valid labels as output;[6] all ICL results in the paper, across all methods, use this constrained decoding. Note that, without constrained decoding, these models may produce invalid labels in the few-shot regimes. For finetuning, we use a classification head so that no invalid outputs may be produced.

[5]This occurs when there are less than k examples with any (non-stopword) overlap with the test example.

[6]Following Ratner et al. (2022): at each generation step, we simply multiply the logits of tokens that could lead to a valid label by a large constant.

A.4 Evaluation

Following prior work (Zhao et al., 2021; Lu et al., 2022; Han et al., 2022; Ratner et al., 2022), we subsample 250 examples from the test set of each dataset. We release the subsampled test set and full prediction outputs for each experiment in the project repository. We evaluate on each dataset with accuracy and macro-F1, to capture both overall performance and the performance on minority classes; as the trends for both metrics are very similar, we report accuracy in the paper for readability.

B Properties of long-context ICL

In this section, we compare the properties of long-context ICL with the known properties of short-context ICL. We consider using ICL as a testbed for properties of long-context models in Appendix E.

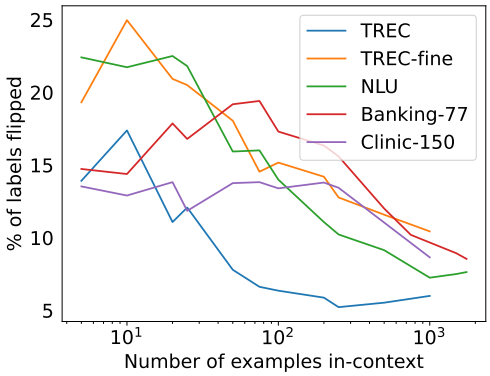


Figure 3: The impact of (randomly) reordering examples in-context decreases with additional demonstrations.

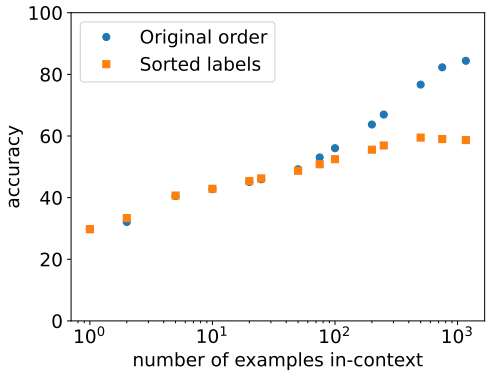


Figure 4: By contrast, sorting examples by label has an increasingly negative impact on performance in longer context regimes. Results on Llama2-32k with Clinic-150.

Is it best to use the entire context? Prior work suggested that, for some simple tasks, providing additional input can *reduce* performance (Levy et al., 2024). However, we observe monotonically increasing performance on nearly every dataset; after the performance curve begins to flatten, small variation occurs, but no significantly lower performance occurs at higher example counts. While using the full context window is computationally costly, and may not be necessary to achieve high performance on these datasets, it does not appear to be harmful to performance; and the additional cost of more input is minimal, as the key-value pairs can be cached and reused across test samples.

Sensitivity to example order Prior work has shown that many models exhibit strong sensitivity to example order in-context (Lu et al., 2022). We examine this by measuring the percentage of predictions that change when the input is reordered, averaged over 3 re-shufflings. Figure 3 shows that, while there is some sensitivity to example order at all context lengths, this effect weakens substantially with additional context. Across all datasets, the percent of labels flipped by reshuffling in 1000-shot ICL is less than *half* the percentage flipped by reshuffling in 10-shot ICL.

Label sorting We also consider an adversarial case for example ordering: we sort the examples so that examples with the same label appear together. At small numbers of examples, this has very little impact; if the average number of examples per class is low, label sorting is similar to a random sort. However, as the number of examples grows, label sorting begins to have a dramatic impact on performance. Figure 4 shows the performance of Llama2-32k on Clinic-150 with and without label sorting. As the number of examples in-context increases, the penalty for input sorting increases as well; at 1169-shot ICL, label sorting decreases accuracy by 25.7 percentage points. This suggests that contextualization of examples with *different* labels is important to performance, and that this contextualization only occurs effectively over relatively short distances in the context window.

Dataset	Llama2	Llama2-32k	Llama2-80k	Mistral
TREC	20 (140)	100 (1129)	75 (2000)	50 (1129)
TREC-fine	75 (131)	250 (1056)	500 (2000)	500 (1091)
NLU	100 (162)	500 (1309)	500 (2000)	250 (1309)
Banking-77	- (100)	500 (838)	750 (1750)	500 (860)
Clinic-150	- (145)	750 (1169)	1000 (2000)	750 (1212)

Table 2: We measure the saturation point as the point at which the model reaches 95% of its maximum accuracy on the dataset; "-" in a column indicates that the maximum performance is achieved by using the full context window. The number in parenthesis represents the maximum number of examples that fit in the context window. As the label space of the dataset increases (from top to bottom row), so does the number of examples that can be used before saturation.

C Saturation

One metric we are interested in is the point where the model performance *saturates*, which we define informally as the point where adding more examples is unlikely to meaningfully improve performance. More formally, we define the saturation point as the smallest number of examples tested such that performance reaches 95% of the model's maximum performance.

Saturation points vary by dataset. We define saturation as the first point at which performance reaches 95% of the model's maximum performance on that dataset. Table 2 shows the number of examples at saturation and the maximum number of examples that fit in the context window for each model. For datasets with larger label spaces, saturation generally occurs later; Banking-77 and Clinic-150 do not saturate within the context window of Llama2 (4096 tokens, which represents between 100-162 in-context examples for these datasets). In the longer-context regime, saturation points generally occur slightly later on Llama2-80k, but in both models occur before the model's maximum context length.

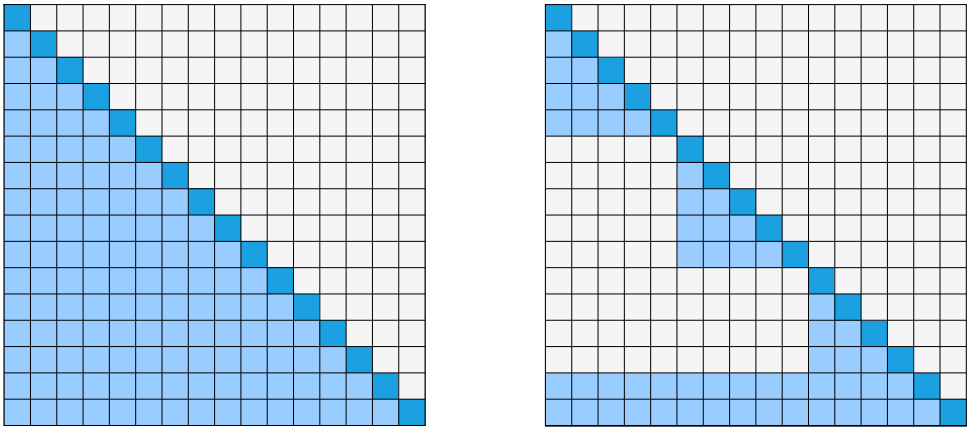
This suggests two things. First, given a fixed model, it is often not necessary to use the full context length to extract high performance from that model. Second, current models do not make use of the full potential of ICL; models often saturate in performance before the maximum number of examples, despite longer-context versions revealing that further performance improvements are possible.

The number of classes has some impact on saturation point– but is not fully explanatory. Our results show datasets with more classes benefit from more demonstrations in-context, on average, before saturation. This is to be expected, as the expected number of demonstrations necessary before seeing the correct label increases with the number of total label classes. To test if this is an intrinsic property of these datasets, or truly linked only to the number of label classes, we construct subsets of two high-label-space datasets, Banking-77 and Clinic-150, by randomly selecting half of the labels to exclude from the dataset. These subsets remain in the same domain, but with a smaller label space; if saturation is tied to the number of examples, then this should move the saturation point. Note that this is distinct from *combining* labels (e.g. TREC vs TREC-FINE), as combining finegrained labels into general labels makes the classification task simpler. It's still possible that the subset chosen is a simpler task (e.g. by removing one of a pair of frequently confused labels); to moderate the effect of this change, we average results over 3 randomly chosen subsets.

Table 3 compares the saturation point of the full- and half-label-space runs. For the datasets with the most number of labels, halving the number of labels also reduces the amount of examples that are useful before saturation, albeit not by half. However, the trend is less clear for the tasks with fewer labels; in some cases, reducing the label space actually *increases* the number of demonstrations before saturation. While the size of the label space clearly has some impact on the saturation point, more investigation is necessary to identify other factors impacting this behavior.

Dataset	Llama2	Llama2-32k	Llama2-80k	Mistral
TREC	14.29 / 24.69	8.86 / 1.77	3.75 / 1.71	4.43 / 6.0
TREC-fine	57.25 / 80.71	23.67 / 27.9	25.0 / 32.38	45.83 / 33.33
NLU	61.73 / 80.71	38.2 / 17.21	25.0 / 26.67	19.1 / 36.67
Banking-77	100.0 / 88.0	59.67 / 37.91	42.86 / 28.57	58.14 / 35.56
Clinic-150	100.0 / 64.04	64.16 / 35.14	50.0 / 22.86	61.88 / 56.67

Table 3: We compare the saturation point between the full-label-space (left) and half-label-space (right) for each model+dataset pair. Here we represent the label space as a percentage of the full context window.



(a) Causal attention (b) Blockwise attention

Figure 5: Normal versus block-causal attention. When performing blockwise attention, we allow full attention for the example we are predicting a label for (here: the last 2 tokens).

.

D Why does long-context ICL help?

To study the underlying mechanism behind the improved performance of the model at longer context lengths, we consider a modified attention pattern where demonstrations can only attend to a small block of nearby demonstrations. The test example we are predicting a label for can always attend to all demonstrations. Figure 5 compares this block-attention to the usual causal attention.

If improved performance is due predominately to the development of a more fine-grained task understanding from embedding many examples together (e.g. continually refining a decision boundary, in the way finetuning the model would do), then encoding many examples in small blocks would be far worse than encoding them all together. If the improvements come largely from seeing more relevant examples to attend to, then the performance should not be strongly reliant on how many other examples each demonstration is contextualized with. Note that this is distinct from methods that overload the same positions with multiple embeddings in order to process longer contexts (e.g. Ratner et al. (2022)); here, we are not modifying any positional information, only restricting attention between demonstrations to a local context block.

Rather than breaking the context window into fixed-length chunks, we fix a number of examples per-block. If the block size is equal to the number of examples in-context, this is equivalent to normal attention; if the block size is 1, each example can attend only to itself. Figure 6 shows results on Banking-77 and Clinic-150, with a comparison to both full attention over all examples and full attention over a single block of examples.

When attending in very small blocks (e.g. for banking, < 5 examples), performance is *worse* than attending to a small set of fixed examples. We hypothesize that this is due to inadequate contextualization of each example leading to less informative embeddings. However, performance quickly climbs; 95% of the performance of full attention is recovered by a block of 50 examples in

the case of Banking-77 or 75 examples for Clinic-150 (more generally: this occurs between 20- and 75-example block sizes in all datasets). In some dataset/model pairs, performance of block attention even slightly exceeds performance of full attention.

To determine how much of the benefit of encoding multiple examples together is due to task learning, we consider the case of ICL with examples sorted by label. In the block attention case, this ensures that most blocks have examples with only one label represented. While sorting examples by label is harmful to performance in the block attention case as well, it is not *more* harmful to the blocked attention model than it is to the full attention model. This suggests that much of the performance of the model is not due to learning decision boundaries in each block and aggregating them, as most blocks in the label-sorted case do not see more than one or two labels. This supports the theory that the primary performance improvement from long-context modeling is due to retrieving from more relevant examples in-context, rather than learning a better task boundary; this is supported as well by the retrieval results in Table 1, where retrieval performance at short contexts is close to (though never exceeding) very-long-context ICL performance.

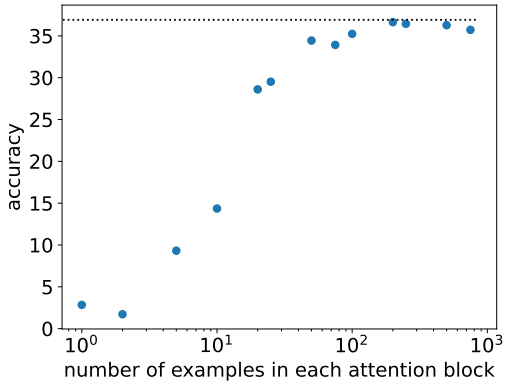
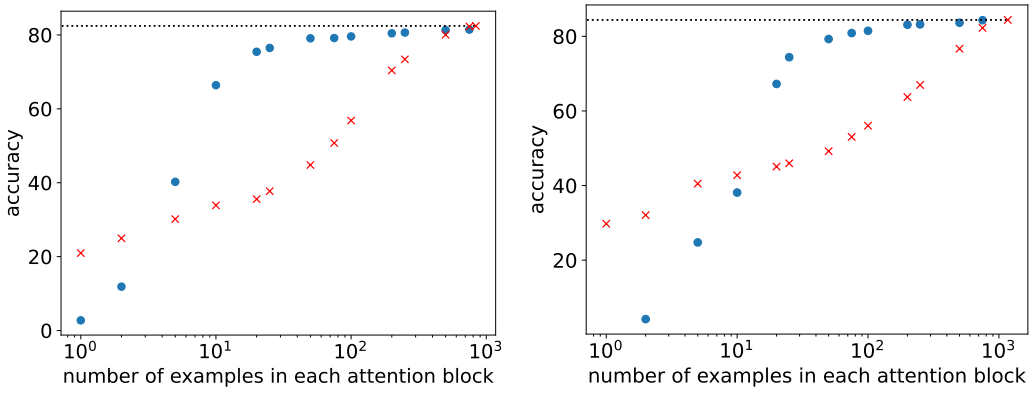


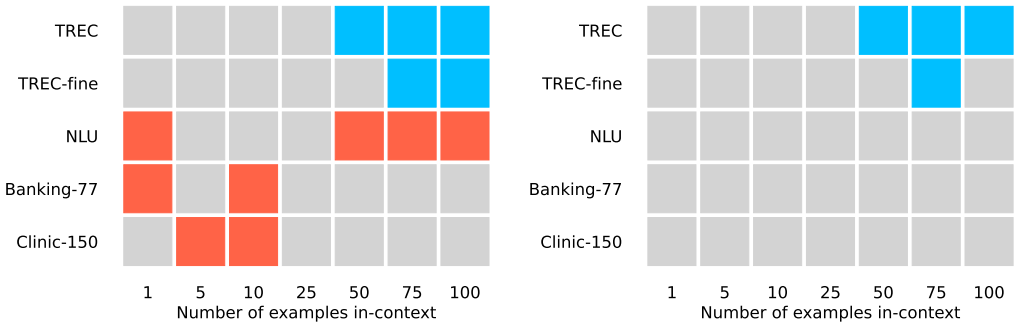
Figure 7: Even when the examples in-context are sorted, block attention can recover similar performance to full attention on the same ordering.



(a) Banking-77

(b) Clinic-150

Figure 6: Comparing block attention to full attention over the same set of examples with Llama2-32k. Block attention approaches full attention with relatively small block size. The blue dots represent block attention with blocks of that size; the red 'x's represent full attention over that number of examples. The black line represents the performance of full attention over all examples.



(a) Llama2-32k (b) Llama2-80k

Figure 8: Short-context behavior of long-context models. Each model's performance is compared to the performance of the base model it was finetuned from, on the same amount of data. **Red** represents significantly worse performance; **blue** represents significantly better performance ($p < 0.05$).

.

Number of examples	1	5	10	25	50	100	200	250
Llama2	**100.0**	93.0	96.5	97.0	98.6	97.95	-	-
Llama2-32k	80.0	95.0	97.0	96.6	98.4	98.5	98.5	**98.9**
Llama2-80k	90.0	94.0	95.0	98.2	**98.5**	98.3	98.0	97.9

Table 4: Copying behavior given the test examples in the context window. Results are averaged over Banking-77 and Clinic-150; bold indicates the best performance for that model.

E Using ICL as a testbed for long-context model properties

In this section, we use in-context learning as a testbed to examine several properties of long-context models.

How do long-context models perform in the short-context regime? Llama2-32k and Llama2-80k are finetuned variants of Llama2-7b, adapted for longer contexts. We evaluate how these models perform relative to the base model in short-context tasks (e.g. ICL using less than 4096 tokens of demonstrations) by testing whether the difference in performance is statistically significant (2-sided t-test, $p < 0.05$). Performance is generally similar, with some areas of slight improvement from the base model; Figure 8 shows full results. We observe degradation in performance in some settings for Llama2-32k, highlighting the importance of testing for behavior regression when finetuning for additional capabilities.

Input utilization We analyze the performance of all methods on a naturalistic needle-in-the-haystack Ivgi et al. (2023); Liu et al. (2024) style test. If the model is effectively using the context, then it should be able to exactly recover the label for any example it has seen in-context. Note that, while a model trained on some set of data is not uniformly capable of exact copying from that training data (Biderman et al., 2023), in nearly all of our finetuning runs, the model fits the training data with 100% accuracy.

We examine this behavior by selecting the same set of examples to use in-context and in evaluation; all models should then be able to achieve 100% accuracy. Table 4 shows the results; while all models achieve very high accuracy on the copied data, no model is able to uniformly copy correctly from the input. Surprisingly, performance improves slightly with additional demonstrations for most models, possibly due to additional specification of the task.

F Related work

Augmenting decoder-only models with long context Many methods for extending the context of language models have been introduced in the last few years. One series of work has focused on positional embedding extrapolation strategies (Peng et al., 2023; Rozière et al., 2024; Chen et al., 2023; Liu et al., 2023; Zhu et al., 2024; Xiao et al., 2024; Han et al., 2024). When extrapolating past pretraining length, models also generally benefit from additional finetuning on long-context data (Xiong et al., 2023). Other methods include adding retrieval-based attention (Bertsch et al., 2023; Tworkowski et al., 2023; Yen et al., 2024) or hierarchical merging of information (Song et al., 2024; YU et al., 2023). The two long-context Llama variants we consider in this work are both examples of finetuning for length extrapolation.

Separately, methods for longer context for ICL have also been proposed. Parallel context windows (Ratner et al., 2022) and structured prompting (Hao et al., 2022) propose methods of re-using the same positional embeddings multiple times to encode more demonstrations; this is quite effective for small numbers of overlaps, albeit with diminishing returns as the number of overlapping windows increases. Li et al. (2023c) propose a new efficient attention mechanism for longer context and motivate its use using long-context ICL on a model tuned with many-shot instruction finetuning. Cho et al. (2023) propose a hybrid of ICL and linear prompting which improves beyond few-shot ICL performance.

Several works have also critiqued the efficacy of long context models. Liu et al. (2024) demonstrate that some long-context models fail to effectively use the middle of the context window; the models we use were released after this work and have generally high scores for middle-of-context retrieval. Li et al. (2023a) suggest that some long-context models are only effective at utilizing inputs that are shorter than their context window's intended supported length; we do not observe this effect strongly, though we do see saturating performance slightly before the maximum number of examples in-context for some models. Li et al. (2023b) show that many models fail at tasks that require reasoning over long dependency lengths; this is unlikely to be an issue in our setting.

Properties of in-context learning Milios et al. (2023) study ICL for many-class classification with models up to 4k context length and find that, when retrieving demonstrations, smaller (7b) models show early performance saturation on many tasks. This is consistent with our findings for Llama2-7b with retrieval; however, the same model continues to learn from demonstrations in the random selection case, and the same size model finetuned for longer context does not show the same performance dropoff and continues to see improvements from additional context. Our results suggest that this failure to use longer context effectively is not an inherent property of 7b models, but instead a type of shallow heuristic used by this model when the demonstrations are of sufficiently high quality.

Xu et al. (2023) study the impacts of ground-truth label, input distribution, and explanations on ICL performance; Bölücü et al. (2023) study the impact of example selection in a specific domain. Lin & Lee (2024) argue that ICL occurs in two modes: learning tasks and retrieving tasks, and that retrieval of similar-but-not-quite-correct tasks can explain "early ascent" behaviors where ICL performance peaks once in a fewshot regime and then performance improves again with a much higher number of examples. Similarly, Pan et al. (2023) argue for a distinction between task recognition and task learning, and suggest that task learning continues to benefit from additional examples at scale. von Oswald et al. (2023) suggest in-context learning can be viewed as gradient descent, although Deutch et al. (2024) argue against this interpretation. Hendel et al. (2023) view ICL as compressing the demonstrations into a "task vector" that maps from inputs to outputs; the surprising effectiveness of block encoding initially appears contrary to this theory, though it is possible that multiple similar task vectors are learned from the separate blocks and then ensembled via attention for the final prediction.

Concurrently to our work, Agarwal et al. (2024) study many-shot prompting of Gemini 1.5 and show improvements from the fewshot setting across both classification and generation tasks. Our work differs in its evaluation of multiple open-source models, our comparison to finetuning the same base model, and our use of ICL as a testbed for analysis of long context behaviors.

Tasks where long context does not help Concurrently to our work, Li et al. (2024) identify a set of tasks where long context is not uniformly helpful. However, we observe that the tasks that show this trend either have near-0 performance at short demonstration lengths or also display an inverse performance trend on the short context scale (e.g. for TacRED (Zhang et al., 2017), we observe that

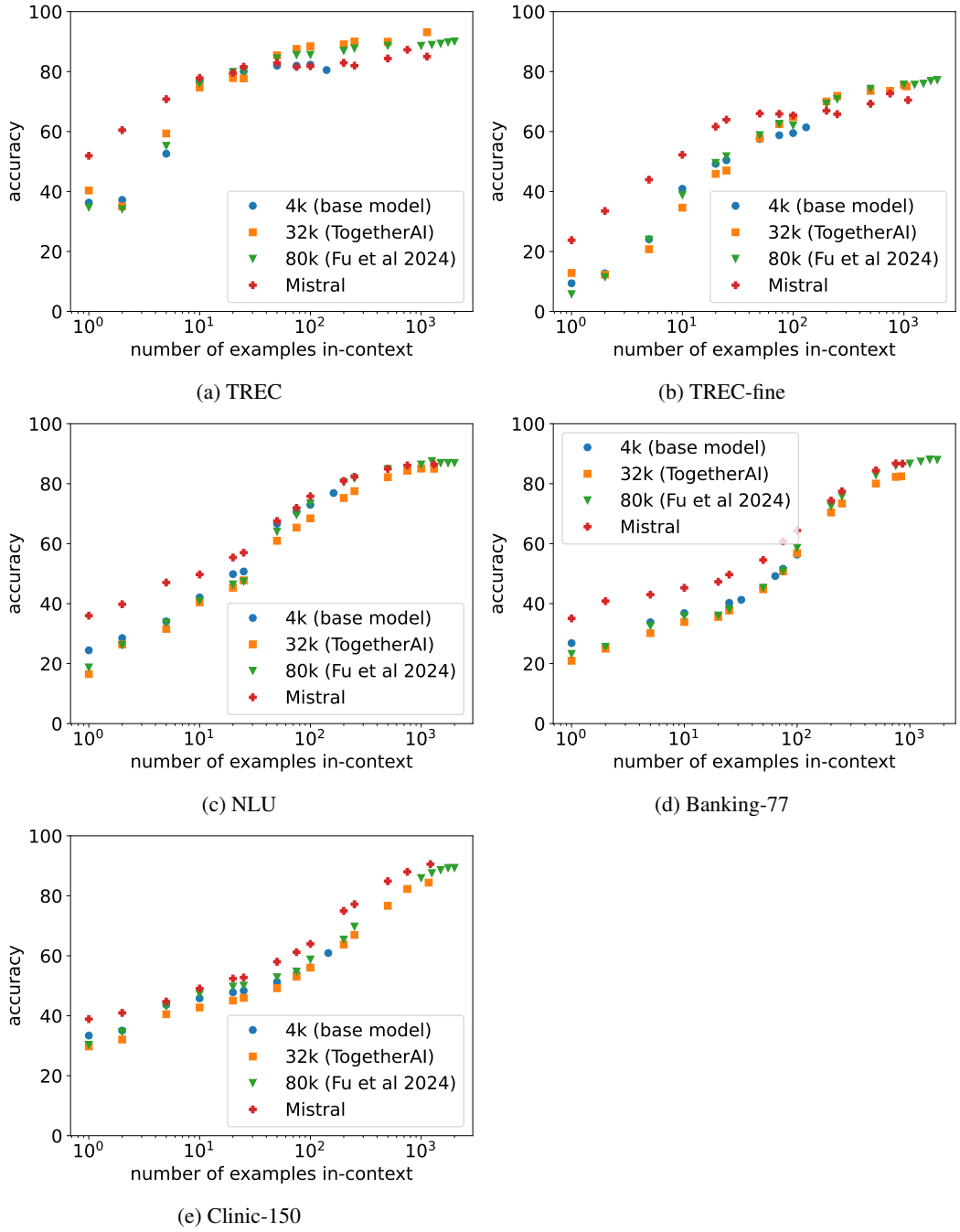
performance decreases from 1 to 10 total demonstrations). While these are important failure modes of language models, we restrict our analysis to tasks without these confounding issues. In Banking-77, the one dataset that our works share, both papers observe similar trends of improved performance with additional context.

Comparing in-context learning and finetuning Min et al. (2022a) show that models trained on fewshot learning can generalize to perform fewshot learning on new tasks; in some cases, this can outperform finetuning directly on the new task. Mosbach et al. (2023) compare finetuning to ICL more directly; they find that finetuning generally outperforms ICL with the same number of examples both in-domain and out-of-domain, when comparing 16-example ICL to finetuning on the same 16 examples. Their setting differs from ours in their choice of model (OPT), the amount of data considered (16 for ICL, 16 or 128 for finetuning), and the use of full finetuning rather than PEFT. Liu et al. (2022) find that PEFT generally outperforms ICL in their setting, where they finetune an encoder-decoder model with a language modeling objective using their T-few method and 20-70 samples. Asai et al. (2023) compare finetuning and ICL for mT5 on cross-lingual transfer and find that ICL outperforms finetuning in some, but not all, of the tasks studied. To the best of our knowledge, no prior work has considered the relative performance of finetuning and ICL in the many-shot regime, where there are hundreds or thousands of examples in-context.

G Full ICL results across datasets

For space, we show 1-2 representative datasets for each point of analysis in the paper. In this appendix, we present results across all datasets for completeness.

G.1 Random selection ICL across all models



(a) TREC

(b) TREC-fine

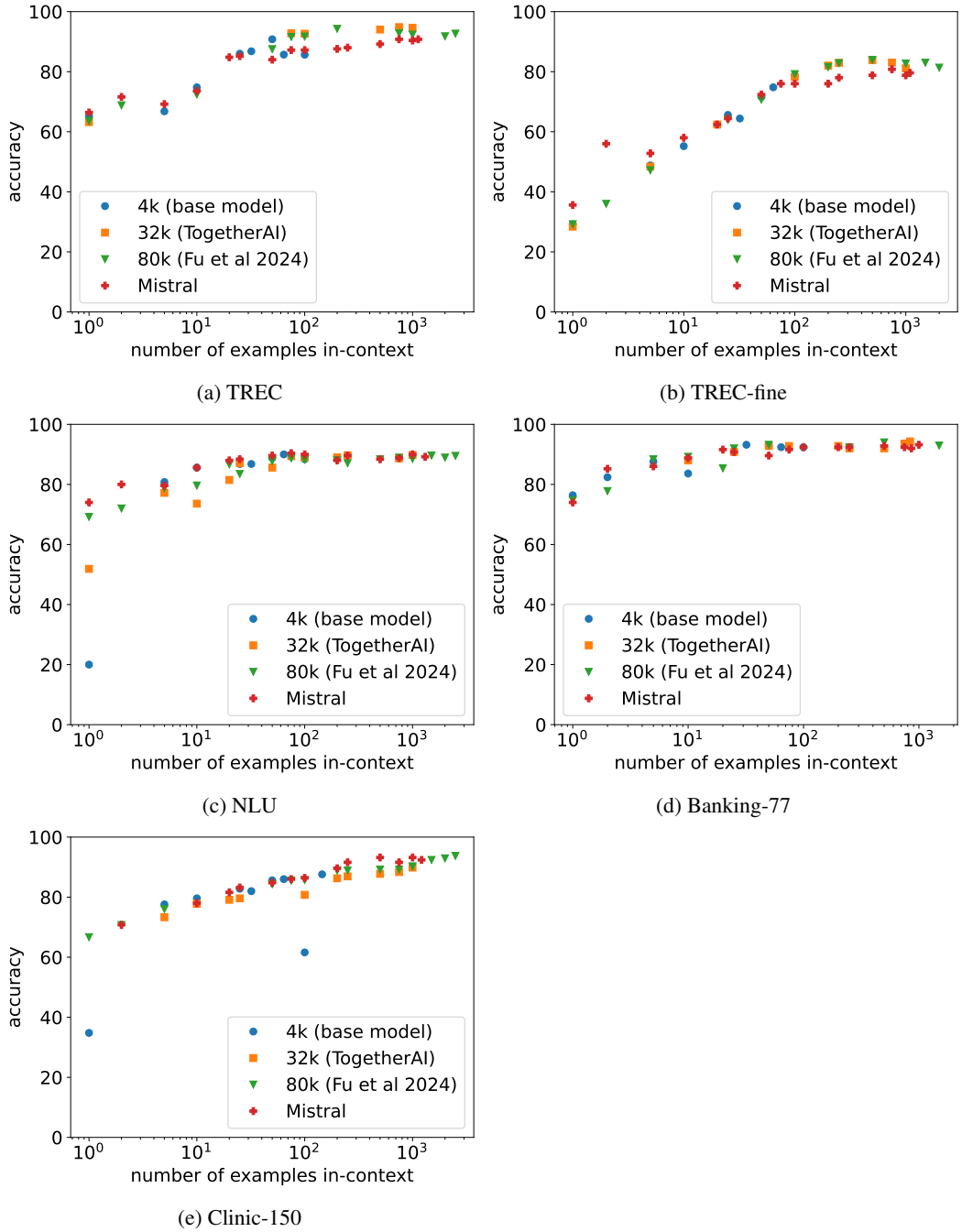
(c) NLU

(d) Banking-77

(e) Clinic-150

Figure 9: Performance of random-selection ICL across all models for each dataset. Performance continues to increase with additional examples in-context.

G.2 Retrieval ICL across all models



(a) TREC

(b) TREC-fine

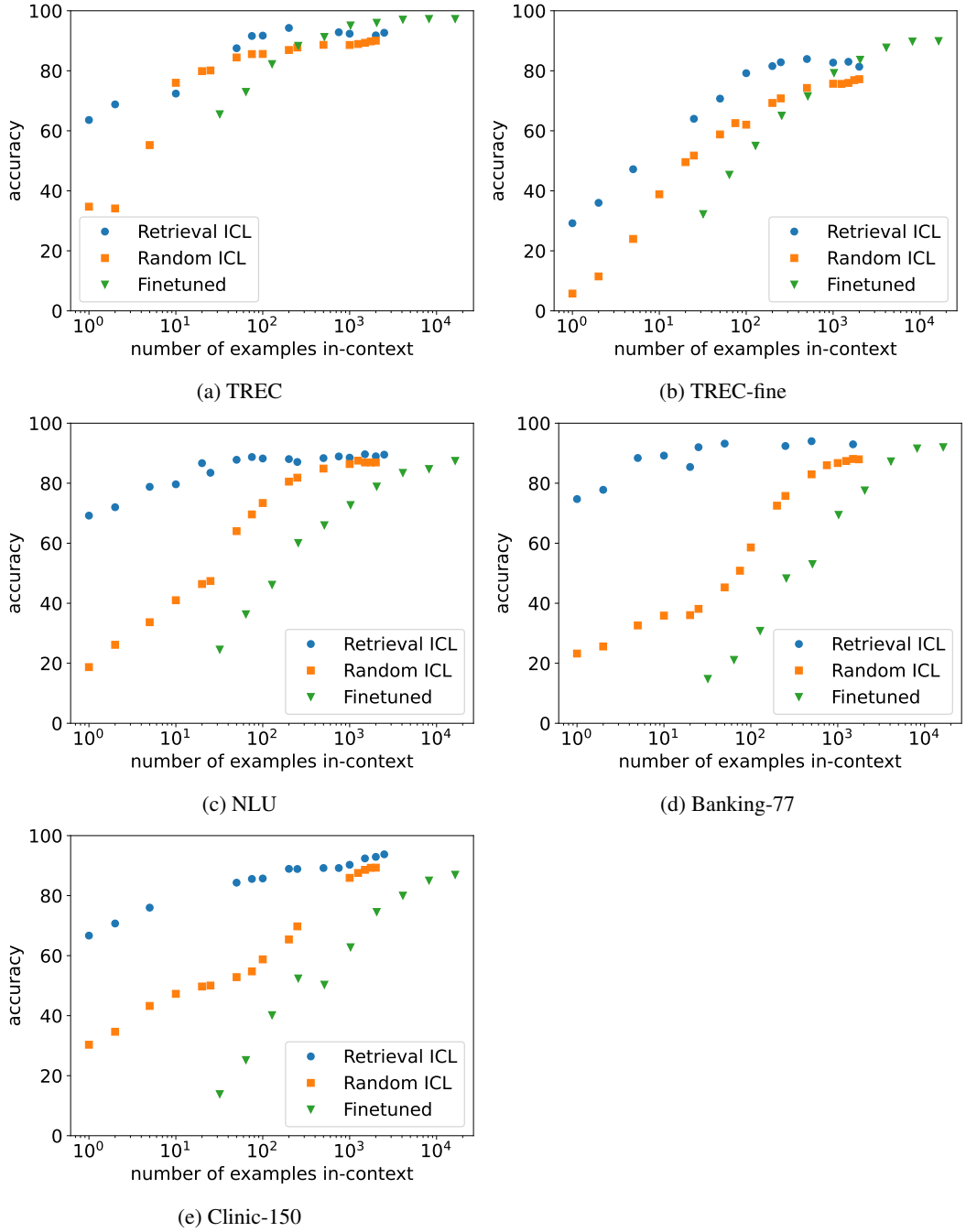
(c) NLU

(d) Banking-77

(e) Clinic-150

Figure 10: Performance of retrieval-based ICL across all models for each dataset. Short-context performance here is higher than for random-selection, but performance continues to improve with more examples until a saturation point, where performance flattens out.

G.3 Comparing retrieval, random selection, and finetuning



(a) TREC

(b) TREC-fine

(c) NLU

(d) Banking-77

(e) Clinic-150

Figure 11: Performance of retrieval-based ICL, random-selection ICL, and finetuning across 5 datasets. At small example counts, ICL outperforms finetuning; when several thousand examples are used, finetuning outperforms ICL in some datasets.

H Finetuning

To perform parameter efficient finetuning (PEFT), we used the peft (Mangrulkar et al., 2022) package (version 0.9.0). We finetune the model for 30 epochs, evaluating it every epoch on the test set, and ultimately choosing the checkpoint with the highest test accuracy. We note that using the test set to perform model selection presents an unfair advantage to PEFT (compared to ICL) and may not be truly indicative of the generalization error. However, doing so provides the advantage of being both comparable to ICL in terms of the data used, as well as giving an upper bound on the true generalization accuracy of the finetuned model, further emphasizing any observed efficacy gap between it and ICL.

Initialization of the classification head While in our default setting we initialize the classification head from the pretrained LM head, subsampled at the representation of the first token in each label, we investigate the efficacy of this approach by contrasting with a randomly initialized classification head. Figure 12 shows that while in the few-shot regime, this approach has significant advantage, as the training set grows in size the difference shrinks to become negligible. In no case was random initialization better than this approach.

Hyperparameter tuning To remain comparable in terms of compute efficient finetuning, we did not perform extensive hyper-parameter tuning per task, and instead experimented with a good global setting on a single dataset (Banking-77). Specifically, we experimented with different learning rates, different LoRA ranks (r) and α (Hu et al., 2022) and also tried applying RSlora (Kalajdzievski, 2023) which sets the scaling factor to $\frac{\alpha}{\sqrt{r}}$ as some evidence suggest it can outperform the original method. Figure 13 summarizes the results, depicting average test accuracy against training examples with different settings.

Ultimately, we found that using HuggingFace's (Wolf et al., 2020) default parameters of $r = 8$, $\alpha = 32$, LoRA dropout of 0.1 and a learning rate of $1e-3$ to work best. In all cases, we used batch sizes of 32 and weight decay of 0.01.

It is possible that methods specialized for finetuning in small-data regimes, such as T-few Liu et al. (2022), might close the gap between ICL and PEFT in the small-data regimes. We did not consider T-few in our analysis because of its additional pretraining stage, which imposes substantial additional cost, and because T-few was developed with a focus on encoder-decoder models and we consider only decoder-only models in our setting.

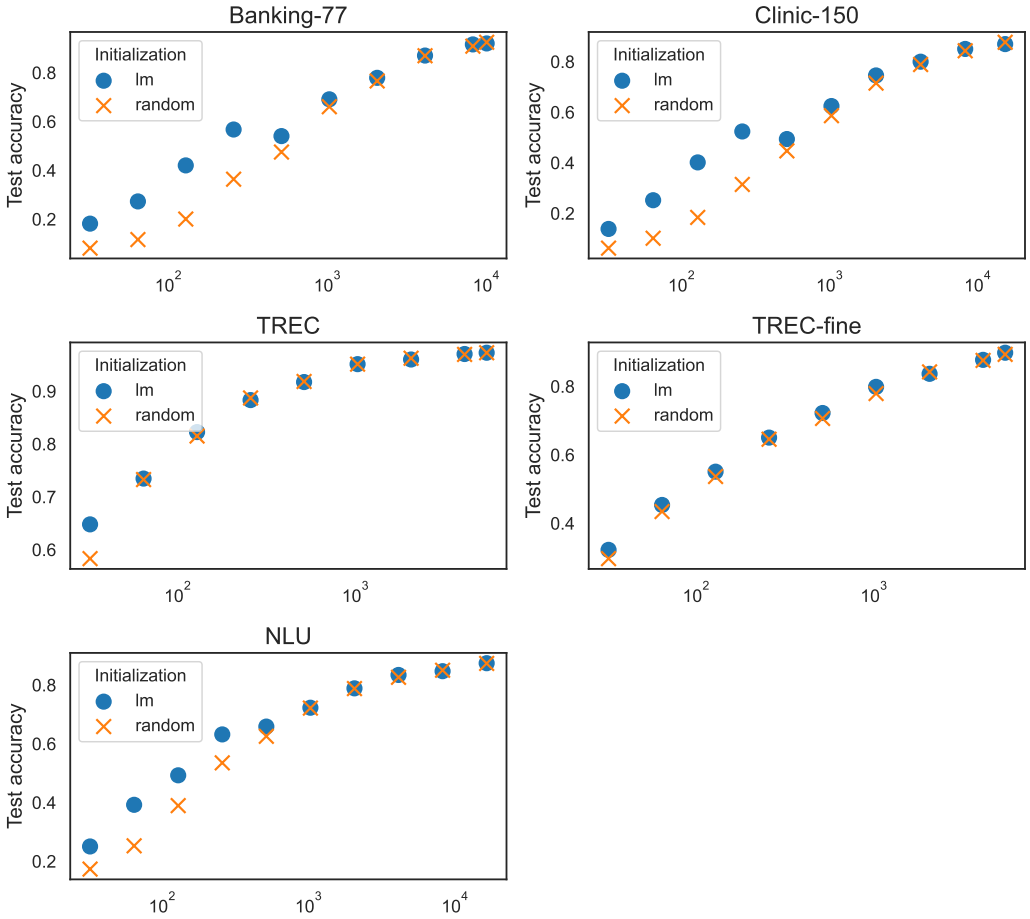


Figure 12: Comparing initialization methods of the classification head when finetuning a PEFT llama-2-7b model. Averaged (best) test accuracy over 5 random seeds. Initialization with *lm* subsamples the pretrained language-modeling head at the first token of the target label, while random samples random weights.

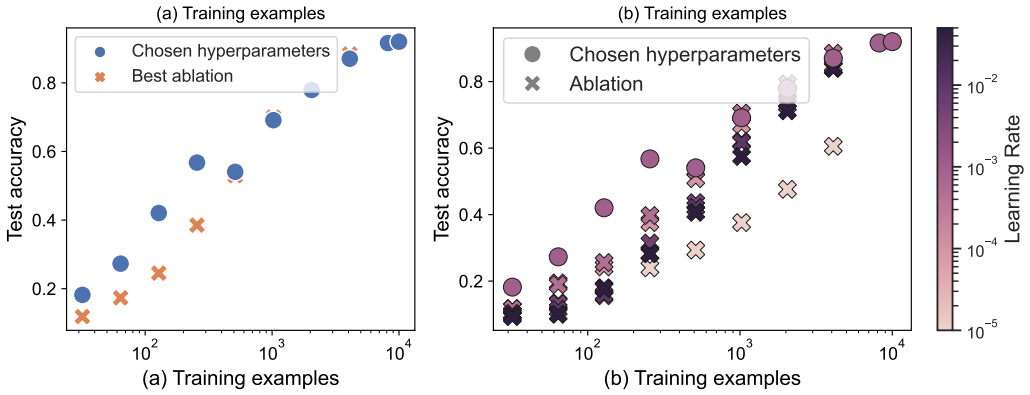


Figure 13: Comparing hyperparameters when finetuning a PEFT llama-2-7b model on Banking-77. Averaged (best) test accuracy over 3 random seeds. (a) Comparing our fixed LoRA configurations to the best alternative configuration (at each scale) we tried. (b) Comparing different learning rates.

Dataset	Domain	# Labels	Avg demo length	Training set size	Example labels
TREC	questions	6	22.7	5,452	location, entity
TREC-fine	questions	50	23.7	5,452	abbreviation expansion, location city
NLU	conversational	68	20.7	19,286	takeaway query, iot hue light up
Banking-77	financial	77	27.4	10,003	top up failed, lost or stolen card
Clinic-150	multiple	151	22.3	15,250	rollover 401k, meal suggestion

Table 5: The datasets we consider in this work span diverse label spaces and domains. The average demonstration length is the average combined length of input, output, and formatting tokens per demonstration provided in the context window.

I Prompt formatting and examples from datasets

Table 5 contains summary statistics for each dataset. We also provide additional description and examples for each dataset.

As a demonstration of the datasets, we provide an example of 3-shot prompting for each dataset with the prompt formatting we used (and with examples drawn from the training set of each dataset).

Prompt formatting and instruction phrasing can have significant impact on performance (Sclar et al., 2023); we keep the formatting consistent with prior work (Ratner et al., 2022), with prefixes for the input and output for each exemplar. Because we use predominately non-instruction-tuned models, we do not add an additional instruction or system prompt.

I.1 TREC

TREC (Hovy et al., 2001; Li & Roth, 2002) is a question classification dataset with two granularities of labels. We refer to the 6-label coarse classification as TREC.

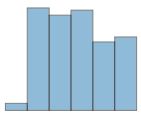


Figure 14: The label distribution of TREC. One label is much less frequent than the rest.

Question: How does light travel through the void of space if there is no medium for it to ‘ wave ’ or ‘ pulse ’ .
Type: description
==
Question: What cathedral was Thomas Becket murdered in ?
Type: location
==
Question: The lawyer who represented Randy Craft , what was his name ?
Type: human
==
Question: What are the rites accompanying the circumcision of a newly born-child in Judaism called ?
Type:

Figure 15: Example 3-shot prompt for TREC.

I.2 TREC-fine

We refer to TREC's 50-label finegrained classification as TREC-fine.

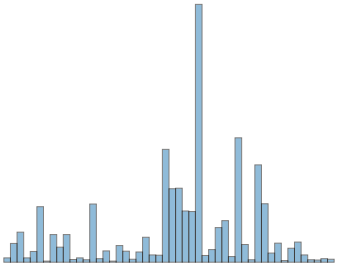


Figure 16: The label distribution of TREC-fine.

Question: How does light travel through the void of space if there is no medium for it to ‘ wave ’ or ‘ pulse ’ .
Type: description manner
==
Question: What cathedral was Thomas Becket murdered in ?
Type: location other
==
Question: The lawyer who represented Randy Craft , what was his name ?
Type: human individual
==
Question: What are the rites accompanying the circumcision of a newly born-child in Judaism called ?
Type:

Figure 17: Example 3-shot prompt for TREC-fine.

I.3 NLU

NLU (Xingkun Liu & Rieser, 2019) is a 68-way intent classification dataset in the conversational domain. The original paper evaluates on 64 of the intents; we use all 68. The data is licensed under CC BY 4.0.

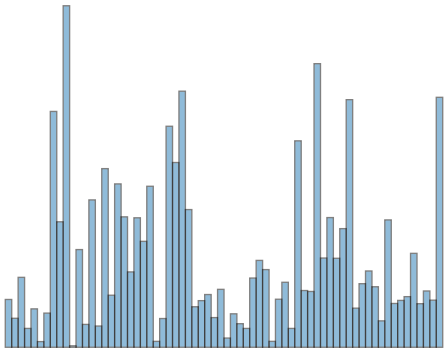


Figure 18: The label distribution of NLU.

utterance: oh it is nice one, olly.
intent: general praise
==
utterance: nope wrong.
intent: general negate
==
utterance: what events near hear are happening this week
intent: recommendation events
==
utterance: play fishing podcasts that are favorited
intent:

Figure 19: Example 3-shot prompt for NLU.

I.4 Banking-77

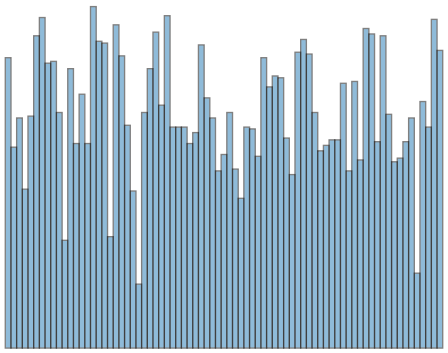


Figure 20: The label distribution of Banking-77.

Banking-77 (Casanueva et al., 2020) is a 77-way intent classification task in the financial domain. Although the accuracy of some labels in BANKING77 has been criticized (Ying & Thomas, 2022), we report results here on the original dataset for consistency with prior work. The data is licensed under CC BY 4.0.

query: How long will my payment be pending?
intent: pending card payment
==
query: My physical card is not working
intent: card not working
==
query: i cant seem to activate card
intent: activate my card
==
query: I didn't set up a direct debit payment on my account.
intent:

Figure 21: Example 3-shot prompt for Banking-77.

I.5 Clinic-150

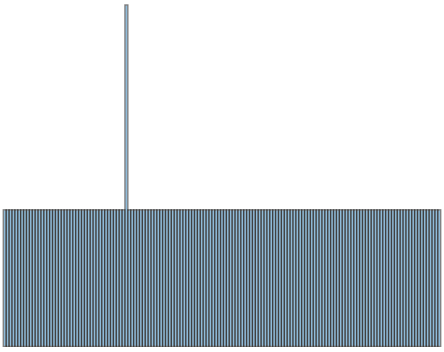


Figure 22: The label distribution of Clinic-150. It is balanced except for the "out of scope" label, which has additional data points in the split we use.

Clinic-150 (Larson et al., 2019) is a 151-way, multi-domain intent classification task; examples are either labeled with an intent from one of 10 domains or with the catch-all "out-of-scope" label. We use the "plus" train split from the original paper, which adds additional "out-of-scope" examples to the dataset. The data is licensed under CC BY 3.0.

utterance: how much is my comcast bill
intent: bill balance
==
utterance: tell me about yourself
intent: what is your name
==
utterance: how to build up my credit score
intent: improve credit score
==
utterance: are you employed by me
intent:

Figure 23: Example 3-shot prompt for Clinic-150.

J Additional details

J.1 Computational cost

For our finetuning experiments, we used approximately 25 GPU-days of compute on 80GB A100 GPUs. The computational cost of the in-context learning experiments was approximately 50 GPU-days, primarily using 48GB A6000 GPUs. All experiments were run on local clusters (i.e. not using cloud providers).

J.2 Broader impacts

Any work that studies general capabilities of language models can be used for both positive and negative applications downstream. In-context learning work is perhaps particularly vulnerable to this dual use, in part because it is a relatively accessible method, requiring far less compute than finetuning or pretraining models.

Independent of our work, Anil et al. (2024) observed that many-shot prompting can be used to jailbreak some models. We do not study or suggest the use of long-context ICL for jailbreaking in our work, but this is a potentially harmful use case of this paradigm. Outside of this risk, we do not foresee any additional harms introduced by long-context ICL methods.

We hope that our work opens up additional possibilities for people who are compute-constrained to customize models to their use cases or tasks. We also aim to carefully explore the boundary between where finetuning and long-context ICL is appropriate, to enable practitioners to make informed choices on when to apply each method.

J.3 Limitations

Our work focuses on open-source models (and predominately on the Llama-2 family); more work is necessary to establish whether this trend holds across model families, although we are encouraged by the strong results Agarwal et al. (2024) observed for many-shot prompting on Gemini 1.5. Additionally, we do not consider full finetuning or other non-LoRA PEFT methodologies; it is possible that some of these methodologies may outperform in-context learning. Finally, we focus on classification tasks, and care should be taken before attempting to generalize to generation tasks.