The State of Intent Detection in the Era of Large Autoregressive Language Models

Anonymous ACL submission

Abstract

In-context learning (ICL) using large pre-trained autoregressive language models (LLMs, e.g. GPT-3) has demonstrated effective classification performance at a variety of natural language tasks. Using LLMs for intent detection is challenging due to the large label space and limited context window, such that it is difficult to fit a sufficient number of examples in the prompt to allow the use of in-context learning. In this paper, dense retrieval is used to bypass this limitation, giving the model only a partial view of the full label space. We show that retriever-augmented large language models are an effective way to tackle intent detection, by-passing context window limitations effectively through the retrieval mechanism. Comparing the LLaMA and OPT model families at different scales, we set new state of the art performance in the few-shot setting with zero training for two of the three intent classification datasets that we consider, while achieving competitive results on the third one. This work demonstrates that the Retriever+ICL framework is a strong zero-training competitor to fine-tuned intent detection approaches. In addition, a small study on the number of examples provided at different model scales is done, showing that larger models are needed to make effective use of more examples in-prompt.

1 Introduction

In-context learning using large autoregressive natural language has recently exploded in popularity. Models pre-trained on massive amounts of textual data are able to reach reasonable performance on a wide variety of tasks with only a few examples of input and output for a given task provided in the model's input prompt in natural language (Brown et al., 2020) (Rae et al., 2022) (Chowdhery et al., 2022). In this work, we seek to push ICL to its limits as we tackle the problem of intent classification, to which ICL has not been directly applied before. Under ordinary ICL with a static set of examples, intent classification would not be possible, as the limited context window of these models would result in overflowing the context window if it were attempted to put at least one example from every class in the prompt. By augmenting the LLM with a dense retrieval model (Reimers and Gurevych, 2019) (Karpukhin et al., 2020) that retrieves the K nearest examples to the given input at a time, intent detection becomes possible to tackle directly with ICL. We evaluate LLMs in this setting with three intent classification datasets: BANKING77 (Casanueva et al., 2020), HWU64 (Liu et al., 2021), and CLINC150 (Larson et al., 2019) [1]. Experiments are done using the LLaMA models (Touvron et al., 2023) and the OPT models (Zhang et al., 2022). We compare the performance achieved against adapter-based fine-tuning of MLM models (DeBERTa-v2-XXLarge with the "Pfeiffer" adapter (Pfeiffer et al., 2020b) implemented with Adapter-Hub (Pfeiffer et al., 2020a)) and the previous SoTA for intent detection.

The contributions of this work are:

1. Showing that Retrieval-Augmented ICL is an effective way to tackle intent detection with zero training, either matching or outperforming fine-tuned adapter-based and contrastive-pre-training-based methods,

2. Analyzing ICL performance with different models and different numbers of examples, as well as performing a small ablation study based on resampling examples to determine what aspects of the inputs and outputs the model is using for ICL.

[1] HWU and CLINC made available under license CC-BY-SA 3.0, BANKING under CC-BY-4.0, use consistent with intended use

2 Related Work

2.1 Nearest neighbor selection of in-context examples

One of the earliest studies of the role of example selection in ICL is “KATE” (Liu et al., 2022). In this paper, the authors probe the performance of GPT-3 on NLP tasks using KNN retrieval (RoBERTa) for example selection. They compare this method against random selection and using the retrieval model directly (plain KNN). They also examine the effect of example ordering on performance and conclude that the most performant ordering (least-to-most and most-to-least similar orderings are tested) depends on the dataset. In our work, we also experiment with example ordering, and conclude that least-to-most ordering is the most effective on all intent detection datasets tested.

Another paper examining the effect of example ordering with regards to in-context learning is (Lu et al., 2022). The authors show that GPT-3 is extremely sensitive to example ordering, to the extent that certain permutations bring near SoTA performance on certain tasks while others perform at near random guessing.

2.2 Few-shot intent detection

The current state of the art in few-shot intent detection is the ConvFit method (Vulić et al., 2021). ConvFit uses a pre-trained LM in a dual-encoder configuration (e.g. BERT or RoBERTa) with two training stages. The first stage is a conversational fine-tuning stage using a generic conversational corpus with a retrieval task (using tuples of (context, response) retrieve the correct response for each context). The second stage is fine-tuning on the specific intent classification dataset with a contrastive loss, allowing the resulting LM to be used in a KNN fashion.

Another way LLMs have been used for intent detection is for data augmentation (Sahu et al., 2022). In this regime, LLMs are used to augment the few-shot datasets to train stronger traditional fine-tuned models (e.g. BERT).

3 Experimental Setup

For our sentence encoder/retriever, we use the SentenceTransformers library (Reimers and Gurevych, 2019), and use the pre-trained “all-mpnet-base-v2” model (pre-trained on over 1 billion training pairs) in frozen mode so that the entire pipeline is training-free. Experiments with contrastively fine-tuning the retriever model are also done.

All experiments were performed on a single A100 80GB GPU. For LLaMA 33B and 65B Huggingface 8-bit quantization was used. The main difference between the OPT and LLaMA models is the amount of pre-training data used. The LLaMA models were trained on 1T-1.4T tokens, while the OPT models were only trained on 180B tokens (see (Zhang et al., 2022) and (Touvron et al., 2023) for more details).

To reduce computational load and make inference easier, instead of using the logits of the LLM to rank our many classes (requiring multiple forward passes), we let the LLM generate freely, and encode the output to compare with our classes via our dense retriever model (SBERT). This allows us to restrict the model output to the set of classes we want without incurring additional inference cost.

4 Results

For direct comparison with previous works, especially ConvFit, we use the same 5-shot and 10-shot sets as DialoGLUE (Mehri et al., 2020). Experiments are run 3 times and the accuracies are averaged, except the zero-training LLM setups, which are deterministic. The contrastively fine-tuned retriever was trained for one epoch to avoid overfitting, using three times as many negative pairs as positive pairs (roughly 5-10 mins depending on the dataset). Otherwise default library parameters were used. The baseline “Pre-trained SBERT KNN” refers to using only the dense retriever to make predictions using 1-nearest-neighbor.

We provide a brief study regarding how to order examples in-prompt by similarity, since previous work has been inconclusive on this front. A brief analysis of ordering is provided in Appendix A.

Table 1 shows the performance comparison of all methods. Performance of the Retriever+ICL pipeline on BANKING and HWU is state of the art in the 10-shot setting. Not only this, but to match or surpass the previous state of the art for these datasets only LLaMA 7B is necessary, which with 8-bit quantization can be run on consumer hardware. In the case of CLINC, the DeBERTa baseline is slightly stronger than the Retriever+ICL. In the most challenging evaluation setting (the highly-specialized intent classes of the BANKING dataset in the most data-scarce 5-shot setting), the margin between DeBERTa and LLaMA 65B is 6.26%. In

Table 1: Performance comparison between all methods.

Model	BANKING 77		HWU 64		CLINC 150	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
Pre-trained SBERT KNN	78.41	85.39	69.89	75.46	82.51	84.84
ConvFit (reported)	-	87.38	-	85.32	-	92.89
DeBERTa (Pfeiffer)	81.47 ± 1.6	88.41 ± 0.19	79.80 ± 0.81	86.93 ± 0.052	91.86 ± 0.66	**95.05** ± 0.33
OPT 13B (20-ex)	81.23	85.65	78.90	83.64	85.27	89.24
OPT 175B (20-ex)	81.30	86.14	83.74	84.94	90.96	93.09
LLaMA 7B (20-ex)	84.42	87.63	85.87	87.55	88.58	91.73
LLaMA 13B (20-ex)	85.39	88.93	86.25	87.83	90.31	93.00
LLaMA 33B (20-ex)	87.37	90.52	86.71	88.75	**92.02**	94.13
LLaMA 65B (20-ex)	**87.73**	**90.71**	**89.03**	**90.06**	91.89	94.47

Table 2: Comparison of Models with Fine-tuned Retriever (20 examples in prompt), compared against non-fine-tuned performance

Model	BANKING	HWU	CLINC
	10-shot	10-shot	10-shot
SBERT KNN	87.40 ± 0.21	83.05 ± 0.47	91.48 ± 0.13
vs. frozen	+ 2.0%	+ 7.6%	+ 6.64%
OPT 13B	87.71 ± 0.18	83.83 ± 0.83	91.83 ± 0.22
vs. frozen	+ 2.06%	+ 0.19%	+ 2.59%
LLaMA 7B	87.39 ± 0.081	87.98 ± 0.75	94.17 ± 0.32
vs. frozen	- 0.24%	+ 0.43%	+ 2.44%
LLaMA 65B	88.93 ± 0.056	**90.12** ± 0.51	**95.62** ± 0.17
vs. frozen	- 1.79%	+ 0.062%	+ 1.16%

general the DeBERTa model struggled in the 5-shot scenarios, likely due to the extremely limited data.

We also provide a study of how performance changes given the number of examples provided in-context. Figures 1 and 2 show this performance. The x-axis value of 110 indicates a fully saturated context window, which is on average this number of examples.

A small ablation study is done to test if the nearest examples are actually necessary for effective performance, or if the LLMs are primarily using distributional information (e.g. the most frequent label in the prompt) or just the class label subset to perform well. By resampling from the classes initially retrieved by the retriever model, we preserve the distribution of labels but change the input examples themselves so that they are no longer the nearest in the embedding space. The result is shown in Figure 3. The ablation study was done on a different split (selected randomly) than the DialoGLUE split for each dataset.

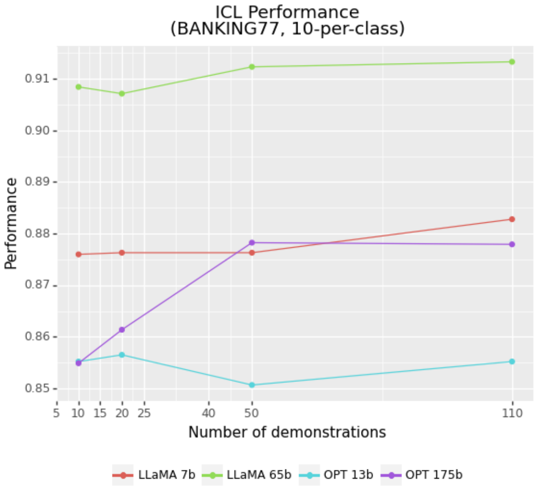

Figure 1: BANKING performance as a function of the number of examples in prompt

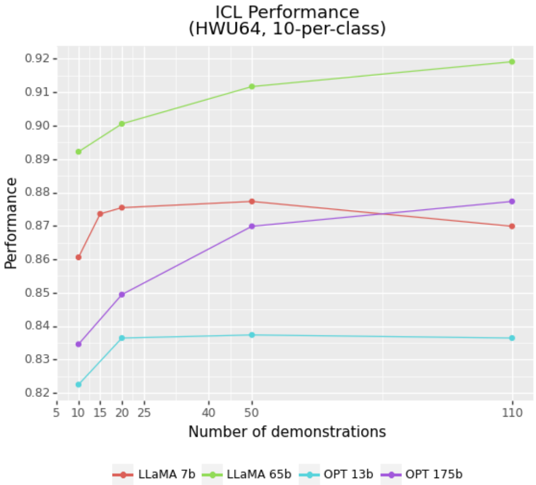

Figure 2: HWU performance as a function of the number of examples in prompt

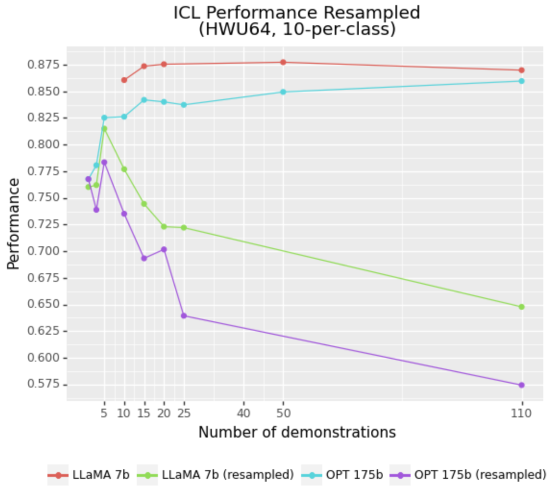

Figure 3: HWU performance when randomly resampling from the retrieved classes

5 Discussion

5.1 ICL at different scales

One trend noticeable from the performance graph as a function of the number of examples for BANKING and HWU (see Figures 1 and 2) is that there seems to be a difference in how effectively the models can use additional examples in relation to their size. The smaller OPT model is unable to effectively make use of the entire context window when it is filled (around 110 demonstrations) and remains at relatively low performance. In contrast, OPT 175B starts at a similar performance as OPT 13B at 10 demonstrations, but shows drastic improvement with more examples. With more examples, OPT 175B is able to match the performance of LLaMA 7B despite being trained on the same amount of data as OPT 13B (however is still unable to match LLaMA 65B). This seems to indicate a difference in ICL ability tied to scale. A similar trend is visible for the LLaMA models, where the performance of the 7B model does not change significantly (see 2), but the 65B model is able to continously improve, especially when going from 20 to 50 examples. All large models show non-negligible improvement from the 20 to 50 example regime, while none of the small models do. In general the LLaMA models have much stronger performance than the OPT models, which lines up with what we expect from LLaMA's stronger semantic priors/additional training data.

5.2 Fine-tuned Retriever

We note large improvements in the pure KNN mode accuracy, as expected, as we are optimizing a metric that is directly correlated with KNN performance. With fine-tuning, the pure KNN setup becomes near-competitive with ConvFit, the previous SoTA. In terms of Retriever+ICL performance, we see mixed results. In general the performance delta is quite small, suggesting that there is no significant retrieval quality bottleneck. LLaMA 65B with the fine-tuned retriever becomes the highest 10-shot CLINC score. In general, the fine-tuned CLINC retriever provides the most boost, which is also the least data-scarce scenario (retriever fine-tuning is expected to be more effective with more data).

5.3 Resampling Ablation

In the resampling ablation (see Figure 3) we see that resampling from the initial class distribution provided by the retriever model greatly hurts the performance across both OPT 175B and LLaMA 7B. This supports the strong performance numbers of the LLMs, showing that they are doing more than just selecting the most common class, using other distributional information, or just using the shortlist of class labels from the full set of classes to select in a more zero-shot fashion.

6 Conclusion

In this work, retriever-augmented in-context learning is shown to be a strong performer in the space of intent detection. State of the art accuracy is achieved in two of three datasets tested, and competitive accuracy is shown in the third. A frozen retriever is used to achieve this SoTA accuracy, making this pipeline as a whole training-free. A briefly fine-tuned retriever leads to somewhat stronger performance in certain cases. We also show how model performance changes as a function of the number of examples provided in-prompt, providing evidence that larger models are able to more effectively make use of many examples. Through a small ablation study, we demonstrate that the LLMs specifically make use of the most similar examples, rather than using the distributional information or just the class label set.

7 Limitations

One limitation of the research in this paper is that the experiments of this paper use the pre-existing

DialoGLUE few-shot splits for each dataset, following the example of prior works and to remain comparable to them (with the exception of the ablation study, which uses a separate split). However, since experiments were done only on this split, it is not necessarily the case that the results/model rankings are transferable to other splits as well (although it is worth noting from Figure 3 that performance on the separate split is very similar to the DialoGLUE split, and the model ranking remains the same).

8 Risks

This work is based on the use of large autoregressive language models (LLMs). LLMs have been trained on massive amounts of text data taken from the internet, which contain many representations of human biases. As such, these models frequently encode these same biases in their parameters and their predictions. They can sometimes perpetuate these harmful biases and stereotypes. As well, the environmental footprint of running these large models is not insignificant.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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. 2019. 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)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. *Benchmarking Natural Language Understanding Services for Building Conversational Agents*, pages 165–183. Springer Singapore, Singapore.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. 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)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pretrained transformer language models.

A Appendix

A.1 Ordering

As expected, least-to-most similar was the most effective ordering across all three datasets. Larger models are less sensitive to ordering, with a smaller delta between the two orderings. See Table 3 for more details.

Table 3: Comparison of LLaMA 7B and OPT 13B model prompt orderings (20 examples in prompt, 10-shot)

Model	**BANKING**		**HWU**		**CLINC**	
	MTL	**LTM**	**MTL**	**LTM**	**MTL**	**LTM**
OPT 13B	73.64	**85.65**	76.39	**83.64**	81.11	**89.24**
LLaMA 7B	83.64	**87.63**	86.99	**87.55**	90.20	**91.73**
LLaMA 65B	88.08	**90.71**	89.03	**90.06**	93.47	**94.47**