Are Llamas Sesquipedalian?
Analyzing Rare Words in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have changed the modern landscape of natural language processing (NLP). Due to their strong performance on multiple tasks, analyzing LLM performance in unusual or difficult scenarios is important. In this work, we investigate LLaMA's performance when using rare and unknown words, something previous transformer based models have been shown to struggle with. We apply various rare word experiments on Large Language Models, specifically LLaMA 7B and 13B. We demonstrate that LLMs still perform worse processing rare and unknown words compared to frequent words, but show that in contextualized scenarios, LLMs face far less deterioration using rare words than previous models.

1 Introduction

Large Language Models (LLMs) have had a large impact in Natural Language Processing and Artificial Intelligence in general. They have shown strong performance on many NLP tasks. Additionally, they have been shown to perform well in zero-shot and few-shot settings (Brown et al., 2020), making them powerful models for various tasks even without fine-tuning. As a result, LLMs have become a large focus of study.

While LLMs have been tested on various tasks, one area that has not been studied is LLMs' understanding of rare and unknown words. Rare and unknown words have always been a challenge in language representation. In static word embeddings like word2vec (Mikolov et al., 2013a,b) and GloVe (Pennington et al., 2014), these words either have weak or no representations. In contextualized embeddings produced by transformer models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), theoretically rare words should have good representations because they are influenced by the context; however, as shown in (Schick and Schütze, 2020), rare words still impede performance in these models as well.

In this work, we evaluate the ability of LLMs to understand and use rare words. We conduct experiments on the LLaMA model (Touvron et al., 2023), specifically the 7B and 13B versions. We make the following contributions: first, we adapt various rare word tasks to causal language models. Then, we apply these tasks to LLaMA 7B and 13B in order to evaluate their ability to understand rare words. We find that in both intrinsic and downstream tasks, the 7B and 13B LLaMA models have a weaker understanding of low frequency words compared to higher frequency ones. However, we find that in downstream tasks, LLaMA model face far less deterioration with rare words than previous models. We also show that some deterioration is due to the downstream rarification task itself, and not only the frequency of the words.

2 Related Work

2.1 Large Language Models

Language modeling has made large gains in recent years. Models like GPT (Brown et al., 2020), Megatron (Shoeybi et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023), have been shown to be proficient in many NLP tasks. In addition, these models are able to handle zero or few shot scenarios, performing well on tasks without finetuning (Brown et al., 2020). In this work, we focus on the LLaMA model. LLaMA is a transformer based model that has a smaller number of parameters than other LLMs, but is trained on much more data. (Touvron et al., 2023) shows that this approach can outperform models with more parameters on various tasks. LLaMA has four versions; 7 billion, 13 billion, 30 billion, and 65 billion parameters. We focus on the 7 billion and 13 billion models (7B and 13B respectively).

2.2 Rare Words

Rare and unknown words have always been a challenge with word embeddings. Static word embedding techniques like word2vec (Mikolov et al., 2013a,b) and GloVe (Pennington et al., 2014) only learn representations for words in the vocabulary of the training corpus. There have been attempts to estimate rare/unknown embeddings to make up for this issue; some approaches use an unknown word's context (Lazaridou et al., 2017; Horn, 2017; Herbelot and Baroni, 2017; Arora et al., 2017; Mu and Viswanath, 2018; Khodak et al., 2018), others use the word's roots (Bojanowski et al., 2017; Pinter et al., 2017; Sasaki et al., 2019), while others combine these approaches (Schick and Schütze, 2019a,c; Hu et al., 2019; Patel and Domeniconi, 2020, 2023). Contextualized models like Elmo (Peters et al., 2018) and BERT (Devlin et al., 2018) are able to produce representations influenced by its surrounding context, allowing for the ability to generate an embedding for an unknown/rare word on the spot. However, as demonstrated in (Schick and Schütze, 2020), contextualized models still struggle on rare words despite this, suggesting embedding estimation techniques are still necessary. This weakness is the main motivation for this work; if rare words are a challenge for smaller pretrained language models, are they still an issue in LLMs?

3 Experiments

3.1 WNLaMPro

First, we evaluate LLaMA's rare word representations using the Wordnet Language Model Probing (WNLaMPro) data set (Schick and Schütze, 2020). This data set was created to analyze a language model's ability to understand rare words. It contains a list of triples (which include keyword, relation, and target words) and pattern sentences for each relation. The goal of this task is to build a sentence out of the pattern and keyword, and then have the model predict the target words based on the inputted sentence. The language model is then evaluated based on where the target words rank in the probability of the output. For example, if we had the pattern "A <W> is a <MASK>" and our keyword is "lime", we would apply mask prediction on "A lime is a <MASK>" as input, and see the probability of the <MASK> token's output. We would then view the rankings of our target words, in this case words like "lemon" or "fruit". The task has defined multiple pattern sets, with relationships including Antonyms (opposites), Hypernyms (a category the word is in), Cohyponyms (words that share a Hypernym), and Corruptions (misspellings of frequent words). Schick and Schütze (Schick and Schütze, 2020) apply this task on BERT and RoBERTa, showing that rarer keywords perform worse at this task than common ones.

	Rare	Medium	Frequent
Overall	0.156	0.206	0.264
Antonym	0.333	0.321	0.550
Hypernym	0.360	0.438	0.475
Cohyponym	0.060	0.054	0.087
Corruption	0.135	-	-

Table 1: LLaMA 7B WNLaMPro (MRR)

	Rare	Medium	Frequent
Overall	0.146	0.197	0.256
Antonym	0.319	0.321	0.552
Hypernym	0.344	0.420	0.454
Cohyponym	0.066	0.051	0.088
Corruption	0.117	-	-

Table 2: LLaMA 13B WNLaMPro (MRR)

We adapt this task to causal language models, specifically a next token prediction task instead of mask prediction. For example, we adapt the pattern "A <W> is a <MASK>" to "A <W> is a", evaluating the next token predicted by the language model. We apply this adapted version of WNLaMPro to LLaMA 7B and 13B and compare the results of rare, medium, and frequent words. Word frequency is determined using the Westbury Wikipedia Corpus (WWC) (Shaoul, 2010) word counts, where occurrences of 0 to 10 instances are considered rare, 10 to 100 are considered medium, and everything higher is considered frequent. Performance is evaluated by looking at the ranks of the target words in the next token probability; the higher probability words have better ranks. This is measured using Mean Reciprocal Rank (MRR). We show the results in Tables 1 and 2[1].

As shown in the results, rare and medium words lag behind frequent words in all categories. In addition, the corruption MRR is low (if the corrupted word is matching its frequent counterpart, it should be close to 1), suggesting that when frequent words are misspelled, LLaMA may struggle with them as well. However, this task generally has weak contexts; the sentences do not contain other informative words to help LLaMA figure out what it could mean. To this end, we also investigate rare words in downstream tasks.

[1]We also report Precision@3 and Precision@10 in Appendix A.

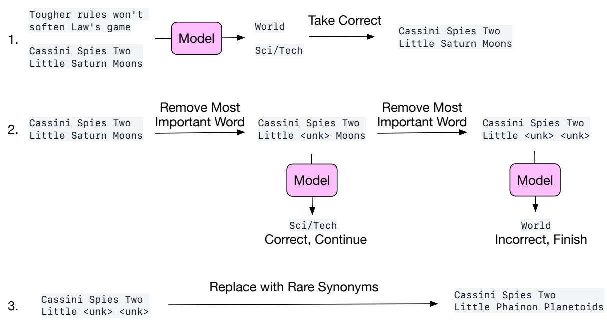

Figure 1: Example of Rarification. Test data is classified by the model, and the correct classifications are kept. Then, the most impactful word is removed from the data until the classification is incorrect. These words are then replaced with rarer versions using a substitution dictionary.

3.2 Rarification on Downstream Tasks

We now shift our focus to rare words occuring in more informative contexts, by investigating downstream tasks. Our main goal is to evaluate how rare words impact the performance on a downstream task. This introduces a challenge, however: due to their infrequency, it can be difficult to see their impact compared to more common words. This does not mean rare words are insignificant; as mentioned in (Schick and Schütze, 2019b), rare words comprehension is an important indicator of language understanding. Additionally, tasks on domains with specific terms or tasks with a large amount of named entities could depend on unusual terms that are extremely relevant to the domain, motivating rare word understanding for specific NLP tasks. Therefore, in order to evaluate rare words in downstream tasks, we use a process called *rarification* (Schick and Schütze, 2019b).

The goal of rarification is to replace important words in the data set with rarer synonyms, and to see how that impacts performance. First, using the WWC word counts used in Section 3.1 and synsets from WordNet (Fellbaum, 2010), we built a substitution dictionary. This dictionary maps frequent words to rare/medium words that are synonyms (from the same synset in WordNet). Similar to the approach in (Schick and Schütze, 2019b), we take the most common sense of each frequent word, and ensure that the corresponding rare/medium words share the same parts of speech. Then, using the data set of the downstream task, we extract a test set of examples that contain at least one word in the substitution dictionary. From this subset we take 10,000 examples. Our goal is to find important words to replace, so we take the following approach. First, we classify each example, and only take the ones that were correctly predicted. Then, for each example, we replace each word from our substitution dictionary with an "<unk>" token and compare how the classification probability changes for each replacement. We keep the replacement with the biggest change in probability and then repeat the process until the predicted class changes. The goal here is to find replaceable words that are needed for correct prediction. We then construct the rarified set by replacing all the chosen words with rarer synonyms. We show an example of this process in Figure 1. This data set has the following properties: with the original words, the classification accuracy should be 100%. With the chosen words replaced by "<unk>", it should be 0 %. Our goal is to see how well the model performs on the data set with the chosen words replaced by rarer versions. If LLaMA understands rare words, it should have an accuracy close to 100%.

We apply rarification to two tasks; AG News (Zhang et al., 2015) classification and Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018). AG News involves classifying news articles into four categories, "World", "Sports", "Business", and "Sci/Tech". For classification, we take the few shot approach. We formulate a prompt with some examples from the train set with their corresponding label, and then a test example without the label. Each example follows the format: "Article : [train article] Label : [train label]". It then ends with "Article : [test article] Label :". We then view the probability of the next token for each class name, selecting the highest as the chosen class. MNLI is an inference task that takes a premise and a hypothesis and assigns a relation between the two; either neutral, entailment, or contradiction. It follows the same approach as AG News, with a different prompt. It starts with "Given a Premise and a Hypothesis, state whether the relationship between the two is described as Entailment, Neutral, or Contradiction. Premise: [train premise] Hypothesis: [train hypothesis] Label : [train label]". It adds two more train examples, then ends with "Premise: [test premise] Hypothesis: [test hypothesis] Label : ".

In addition to the two LLaMA models, we include the results of rarification with BERT and RoBERTa from (Schick and Schütze, 2019b) (denoted with a "*"). We emphasize that the various models are not directly comparable with one an-

Model	AG News	MNLI
BERT(base)*	61.9%	53.4%
RoBERTa(large)*	65.7%	68.4%
BERT+BERTRAM*	66.6%	62.7%
RoBERTa+BERTRAM*	69.0%	73.2%
7B	86.3%	86.0%
13B	96.9%	74.0%

Table 3: Rarefication results. Results denoted by "*" are taken from (Schick and Schütze, 2019b). Each model should be compared to their unrarified data set, which has an accuracy of 100%.

other, as the rarification set is dependant on the type of model (it builds the set based on what the model gets correct). In addition, our experiments pull from a subsample of 10000 examples from each data set, and our word frequencies are based on WWC, as opposed to WWC combined with Book-Corpus in (Schick and Schütze, 2019b). Regardless, the results indicate how robust each model is to rare words, as each result is a measure on how much the model deteriorates when the original data subset is 100% accurate. We also include these models enhanced with BERTRAM (Schick and Schütze, 2019b), which improves rare word representation, to compare how enhanced models perform with rare words. We apply the rarification approach to each task using 7B and 13B. For each task/model combination, we get a rarified data set, on which we then apply the few-shot learning approach with the corresponding model. The results of rarification on AG News and MNLI are shown in Table 3.

As shown in the results, rarer words lead to some deterioration of results in both 7B and 13B. This demonstrates that even in downstream tasks with stronger contextualization, LLaMA has weaker performance, reducing from 100% to 86.3% and 86.0% in the 7B model for AG News and MNLI respectively, and to 96.9% and 74.0% in the 13B model. That being said, the high percentages suggests LLaMA does have smaller degradation from rare words compared to other models. This can especially be seen in the 13B model in AG News, which only degrades by 3.1% when rarification is applied. Compared to BERT and RoBERTa, LLaMA is far more robust to rarification, with much higher performance than the other models. This even holds true when BERT and RoBERTa use rare word estimation model BERTRAM to improve their rare word representations, suggesting LlaMA's representations are higher quality, despite not being as strong as their frequent word representations.

	AG News		MNLI	
	Rare	Freq	Rare	Freq
7B	88.0%	92.4%	85.4%	90.4%
13B	97.4%	98.0%	71.2%	76.0%

Table 4: Rarification using Rare vs Frequent Words

One potential risk of rarification is that the weaker performance can be attributed to the act of substituting the words, as opposed to the words themselves. To verify that weaker performance of LLaMA is due to rare words, we propose a variant on the rarification task. We build another substitution dictionary, this one with frequent word replacements (i.e. frequent synonyms of frequent words). We then take the overlap of replaceable words between this substitution dictionary and the rare word one, in order to create a comparable subset. We then repeat the rarification process, and compare the sets. Note that this creates a different data set, and therefore is not directly comparable to the results in Table 3.

We show the comparisons in Table 4. As shown in the results, replacing words with rare words does indeed make a difference, demonstrating that LLaMA has a weaker understanding of rare words compared to frequent ones. However, substitution in rarification does impact results, as shown by the fact that frequent replacements are not 100%. Overall, while the rarification process inherently leads to deterioration in the results, rare words still lead to more deterioration in LLaMA compared to frequent ones.

4 Conclusion

We investigate performance of LLaMA 7B and 13B on rare words. We find that in low context scenarios there is a sizable gap in language model understanding between frequent and rare words. We also find that LLaMA has weaker rare word performance in downstream tasks, but the deterioration is far less than previous models. This suggests that previous contextualized embedding estimation methods like BERTRAM (Schick and Schütze, 2019b) may still be applicable to modern LLMs, and worth considering. We plan to investigate this further in future work.

Limitations

There are some limitations to our work. First, we rely on building a specific prompt for the few-shot rarification task (Section 3.2), and extracted the predicted class by viewing the next token prediction probability. While this approach gave satisfactory results in the main classification task, it is possible that other prompt building methods and /or classifier methods could lead to stronger performance in general, and maybe even better understanding of the rare words. Secondly, our investigation does not cover the larger LLaMA models (the 30 billion and 65 billion parameter versions), due to computational capability. However, it would be very interesting to see how these larger models fit into these experiments, especially given the difference in performances between 7B and 13B in the rarification tasks.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309.

Franziska Horn. 2017. Context encoders as a simple but powerful extension of word2vec. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 10–14.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning for out-of-vocabulary words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4102–4112.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon M Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective post-processing for word representations. In *6th International Conference on Learning Representations, ICLR 2018*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Raj Patel and Carlotta Domeniconi. 2020. Estimator vectors: Oov word embeddings based on subword and context clue estimates. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Raj Patel and Carlotta Domeniconi. 2023. Enhancing out-of-vocabulary estimation with subword attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3592–3601.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112.

Shota Sasaki, Jun Suzuki, and Kentaro Inui. 2019. Subword-based compact reconstruction of word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3498–3508.

Timo Schick and Hinrich Schütze. 2019a. Attentive mimicking: Better word embeddings by attending to informative contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 489–494.

Timo Schick and Hinrich Schütze. 2019b. Bertram: Improved word embeddings have big impact on contextualized model performance. *arXiv preprint arXiv:1910.07181*.

Timo Schick and Hinrich Schütze. 2019c. Learning semantic representations for novel words: Leveraging both form and context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6965–6973.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8766–8774.

Cyrus Shaoul. 2010. The Westbury lab Wikipedia corpus. *Edmonton, AB: University of Alberta*, page 131.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A WNLaMPro More Results

We report all the results of WNLaMPro in Table 5 and Table 6.

B Implementation Details

All experiments were conducted using Pytorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2020) libraries.

	Rare			Medium			Frequent		
	MRR	P@3	P@10	MRR	P@3	P@10	MRR	P@3	P@10
Overall	0.156	0.064	0.027	0.206	0.085	0.043	0.264	0.112	0.057
Ant	0.333	0.111	0.033	0.321	0.107	0.032	0.550	0.189	0.059
Hyp	0.360	0.149	0.073	0.438	0.188	0.088	0.475	0.211	0.098
Coh	0.060	0.022	0.014	0.054	0.018	0.015	0.087	0.032	0.026
Cor	0.135	0.056	0.018	-	-	-	-	-	-

Table 5: WNLaMPro on LLaMA 7B

	Rare			Medium			Frequent		
	MRR	P@3	P@10	MRR	P@3	P@10	MRR	P@3	P@10
Overall	0.146	0.062	0.028	0.197	0.082	0.044	0.256	0.110	0.057
Ant	0.319	0.111	0.033	0.321	0.107	0.032	0.552	0.189	0.060
Hyp	0.344	0.146	0.072	0.420	0.182	0.086	0.454	0.202	0.093
Coh	0.066	0.023	0.015	0.051	0.017	0.016	0.088	0.034	0.028
Cor	0.117	0.053	0.018	-	-	-	-	-	-

Table 6: WNLaMPro on LLaMA 13B