Analyzing Cultural Common Sense Knowledge in ELECTRA

Samin Mahdizadeh
saminsani162@gmail.com

Mohsen Fayyaz
mohsen.fayyaz77@ut.ac.ir

1 Problem statement

Recent work has shown remarkable advancements in natural language understanding and textual inference, resulting in powerful models that can easily outperform humans in some cases. However, the challenging question is whether these models can go beyond pattern recognition and store commonsense knowledge over information not explicitly present in the text. A recent study (Yin et al., 2022) analyzed multilingual Pre-trained Language Models (mPLMs) for their cultural knowledge and introduced a framework for geo-diverse commonsense probing on multilingual PLMs. In this work, we concentrate on the performance of ELECTRA (Clark et al., 2020), and analyze it for cultural commonsense knowledge. Since ELECTRA predicts whether a token was replaced by a generator and considers all input tokens rather than masked ones, it can be assumed that the model can perform better on this task than models like BERT, and reach superior results.

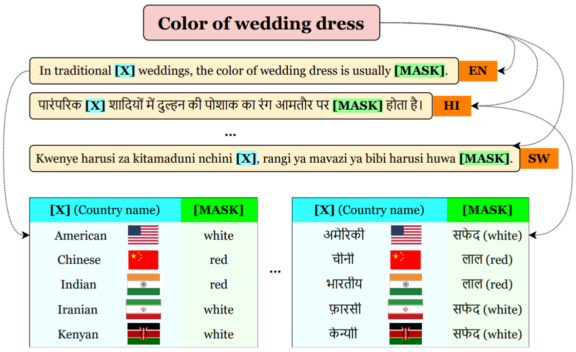
Our code is available at: Google Drive Folder

2 What you proposed vs. what you accomplished

- ~~Collect and preprocess dataset~~
- ~~EDA on our curated dataset~~
- ~~Adapt ELECTRA for our tests on collected dataset and examine its performance~~
- ~~Analyze the output of the model~~
- ~~Work on final reports~~

3 Related work

Language Model Probing. One way to detect what knowledge is encoded in PLMs is *probing*. Initially, this method was used to detect signs of morphology (Peters et al., 2018), semantic (Vulić et al., 2020), and syntax (Shi et al., 2016). Recently researchers have explored more complex knowledge like commonsense and tried to evaluate PLMs regarding these tasks. Razeghi et al. (2020) produced behavioral tests for commonsense knowledge using knowledge graphs. The LAMA probe (Petroni et al., 2019) extracts commonsense knowledge from masked language models with the help of a cloze-style QA task. COMET (Bosselut et al., 2019) is an autoregressive model fine-tuned on knowledge graphs (ConceptNet and ATOMIC) that predicts objects for each subject-relation pair. Recent works have focused on different aspects of common sense knowledge, including cultural (Yin et al., 2022) and metaphorical (Chen et al., 2022) knowledge. In addition, West et al. (2022) introduced a knowledge distillation method, extracting knowledge from large language models and training smaller models with the extracted knowledge, which can perform better than GPT in tasks that require commonsense knowledge.



In traditional [X] weddings, the color of wedding dress is usually [MASK].

पारंपरिक [X] शादियों में दुल्हन की पोशाक का रंग आमतौर पर [MASK] होता है।

Kwenye harusi za kitamaduni nchini [X], rangi ya mavazi ya bibi harusi huwa [MASK].

[X] (Country name)	[MASK]
American	white
Chinese	red
Indian	red
Iranian	white
Kenyan	white

[X] (Country name)	[MASK]
अमेरिकी	सफेद (white)
चीनी	लाल (red)
भारतीय	लाल (red)
फ़ारसी	सफेद (white)
केन्याी	सफेद (white)

Figure 1: Examples of GeoMLAMA dataset (Yin et al., 2022).

Geo-Diverse Commonsense. Existing data often include information from certain regions.

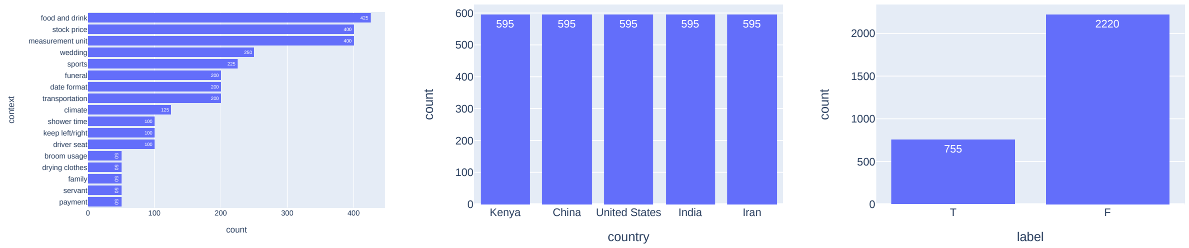


Figure 2: The statistics of the dataset we created for ELECTRA.

Hence the model may overlook the differences between different regions due to cultural diversities. Yin et al. (2021) constructed a geo-diverse visual commonsense reasoning dataset and evaluated model generality to understand geo-location commonsense. Yin et al. (2022) probed mPLMs to test their ability for cultural events (e.g., wedding dress color) in different languages. Liu et al. (2021) developed an ImageNet-Style dataset to represent more languages and cultures and found that cross-lingual transfer performance is not better than supervised performance in English.

4 Your dataset

The primary dataset we use is based on the GeoMLAMA dataset obtained from Yin et al. (2022), which contains 3125 prompts in English, Chinese, Hindi, Persian, and Swahili, with a wide coverage of concepts shared by people from American, Chinese, Indian, Iranian and Kenyan cultures. Since our purpose is to evaluate ELECTRA on cultural commonsense, we only use English data. In order to apply this dataset to ELECTRA, we need to fill mask inputs with candidate answers (both correct and wrong) and evaluate the model.

The GEOMLAMA dataset has 625 prompts for 5 languages(china, US, India, Iran, and Kenya). What is more,to obtain more robust probing results, 125 prompts are sentences that are generated directly while 500 of them are obtained by paraphrasing (most of them were generated using back-translation).

4.1 Data preprocessing

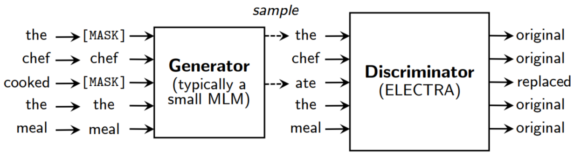


Figure 3: The pre-training objective of ELECTRA (Clark et al., 2020).

In order to reach the final dataset, We replaced the mask token in the prompts with one of the possible answers. The correct candidates were given the label T and the incorrect candidates were given the label F. Also, since the paraphrased sentences did not have the information related to the correct answer and the candidates. First, it was determined from which main sentence each of them was obtained, and then by grouping them by country, other information was obtained from their main sentence. Figure 2 shows some statistics related to the new dataset. This dataset contains 755 correct data and 2220 incorrect data. Since there can be several correct answers for a prompt, this dataset has more correct sentences than the original dataset. In the original dataset, instead of all the candidates, there was only the mask token. In addition, the number of data for all countries is the same and is equal to 595. It can be seen that 17 cultural concepts have been questioned in this dataset, the largest number of data is related to food and drink (425), and the lowest number of data (50) is related to concepts such as family, payment, servant, drying clothes, and broom usage.

5 Baselines

We will repeat the experiments with three other models (BERT, RoBERTa, and XLM-RoBERTa) and check their performance in zero- and few-shot settings. The better the models can perform the defined tasks, the more knowledge is encoded in their representations.

6 Your approach

6.1 Our approach

We hypothesize that ELECTRA can perform better in predicting cultural events than BERT. This

Instance	Correct Answers	Replaced Answer	Label
Most of Kenya is located in continental areas.	tropical	continental	False
Most of Kenya is located in dry areas.	tropical	dry	False
Most of Kenya is located in tropical areas.	tropical	tropical	True
The most popular sport in Iran is baseball.	soccer, volleyball	baseball	False
The most popular sport in Iran is soccer.	soccer, volleyball	soccer	True
The most popular sport in Iran is volleyball.	soccer, volleyball	volleyball	True

Table 1: Examples of our dataset.

is attributed to the fact that this model uses all input tokens to do the task. Furthermore, research on Prompting ELECTRA (Xia et al., 2022) shows that this model can be an excellent few-shot learner and outperforms masked language models in a wide range of tasks. Therefore, we feed curated sentences and observe the model's efficiency in predicting commonsense cultural knowledge. Our approach is to evaluate the model under a zero-shot setting in which we will replace country names and mask inputs. In order to extract the specific token probability (e.g., wedding dress color), the exact location for the token should be obtained. In the following steps, checking the model's performance with prompting and few-shot examples can provide helpful information for the analysis.

6.2 Tools

We leverage the deep learning library *PyTorch* (Paszke et al., 2019). For the language models, we use *HuggingFace Transformers* library (Wolf et al., 2020) implementation. And for resources to run our experiments, we use `https://colab.research.google.com`.

7 Experiments

In this section, we introduce different approaches to examine cultural knowledge in various language models and asses their performance.

After analyzing the logits of the Electra model, we found out that the sentences that ELECTRA detects as fake with high confidence are the sentences with answer words that are not the correct answer in any culture. For instance, in the sentence ”Normally when driving in the United States, people stick to the *anywhere* hand side.”, ELECTRA easily detects ”anywhere” as a replaced token, and it is not actually in any true answers of any culture. Therefore, we continue our next experiments only on the answers which are in the true option of at least one culture. We made this change because we want to explore the cultural knowledge of ELECTRA, not its general knowledge.

7.1 Use Model Representation

One of the ways to check whether modes have specific knowledge or not is to use sentence representation. In this section, the goal is to evaluate the representation of sentences by language models. The better the model can code the sentences, the easier for the classifier to extract knowledge. For this reason, in this method, all the layers of the language model are frozen and only the classifier weights are updated to predict whether the sentences are true or false. The results are shown in Table 2.

Since the dataset used is slightly unbalanced (60% with F labels and 40% with T labels), the F1 (weighted) criterion can be suitable for evaluating the models. It can be seen that the performance of the models was almost random. This shows that even if there is cultural information in language models, it is difficult to retrieve them and a stronger classifier is needed to extract information. However, the Electra model has achieved better results than other models and the F1 measure equals 0.64.

7.2 Question answering

In this section, an attempt is made to examine the knowledge stored in the models by the question-answering task. In this method, two types of yes-no, and multiple-choice questions are used.

Yes-No Question: for creating a yes-no question-answering dataset, we add “is this sentence true?” to each sentence in the original dataset. The model has two options (yes. This sentence is true – no, This sentence is false). We select the sentence with the highest probability

Model	Accuracy	F1	Recall	Precision
BERT	0.598404	0.553549	0.588567	0.59840
ELECTRA	0.664894	0.643751	0.648103	0.664894
ROBERTA	0.515957	0.351213	0.266212	0.515957
XLM-ROBERTA	0.590426	0.438376	0.348602	0.590426

Table 2: Use models representation to predict labels

Model	Accuracy	F1	Recall	Precision
ELECTRA	0.220267	0.233467	0.220267	0.248648
BERT	0.162667	0.196195	0.162667	0.247973
ROBERTA	0.107733	0.136617	0.107733	0.187099
XLM-ROBERTA	0.081067	0.129763	0.081067	0.325333
BERT-SWAG	0.414933	0.433452	0.453704	0.414933

Table 3: Multiple choice result in comparison to fine tuned model

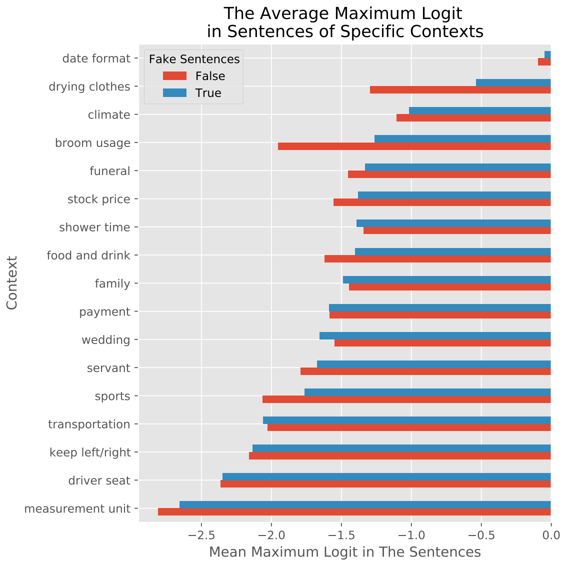


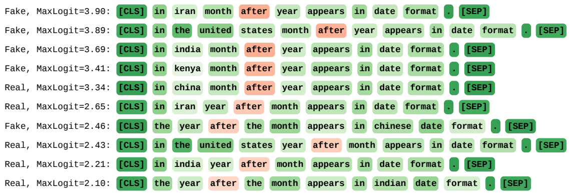
Figure 4: The average maximum logit in sentences of each context grouped by being fake or not. Higher logits are less real in ELECTRA's view. Generally, fake sentences have logits larger than real sentences demonstrating ELECTRA's relative knowledge of cultural concepts.

to be the model choice and compare it with the real answers. We do this experiment in zero and few-shot (k=16,128,500) settings. The results (reported in Appendix A) show that the language models have not been able to perform the task well in any of the settings. Even giving more samples to the model (up to 500 samples) has not helped much to improve the task performance. This may be attributed to the fact that rule or feature extraction is difficult to predict cultural contexts, and learning the task is not possible with small amounts of data.

Multiple Choice Question: Since language models are sensitive to the generated prompts and may reach different results by changing the combination of words in the sentence, in this section, each sentence in the dataset is first given to a question-generation model to generate a question about cultural context. Then, candidate answers are selected for each question and the models must predict the correct answer among these options(3 choices and 4 choices were tested in this experiment). When evaluating the models on three-choice questions in few-shot settings (up to k=128), the accuracy of the models was about 32%, which indicates a random performance. But, with the increase of data up to k=500, this number reached 38% for some models, which might show the improvement of the models if more data is observed(Appendix A). Also, we hypothesized that the random performance of the models might be due to not understanding the task and not the lack of cultural knowledge. Hence, we once again checked the performance of other models with the model fine-tuned on the SWAG data (four-choice questions). As the results in Table 3. show, this fine-tuned model is more effective than other models. Therefore, they may not have understood the question-answering task rather than lacking the knowledge.

7.3 Contexts

In this section, we analyze the different contexts in the dataset and how ELECTRA understands them. Figure 4 demonstrates the average maximum logit



```
Fake, MaxLogit=3.90: [CLS] in iran month after year appears in date format . [SEP]
Fake, MaxLogit=3.89: [CLS] in the united states month after year appears in date format . [SEP]
Fake, MaxLogit=3.69: [CLS] in india month after year appears in date format . [SEP]
Fake, MaxLogit=3.41: [CLS] in kenya month after year appears in date format . [SEP]
Real, MaxLogit=3.34: [CLS] in china month after year appears in date format . [SEP]
Real, MaxLogit=2.65: [CLS] in iran year after month appears in date format . [SEP]
Fake, MaxLogit=2.46: [CLS] the year after the month appears in chinese date format . [SEP]
Real, MaxLogit=2.43: [CLS] in the united states year after month appears in date format . [SEP]
Real, MaxLogit=2.21: [CLS] in india year after month appears in date format . [SEP]
Real, MaxLogit=2.10: [CLS] the year after the month appears in indian date format . [SEP]
```

Figure 5: The sentences attributed to the highest maximum logits of ELECTRA.

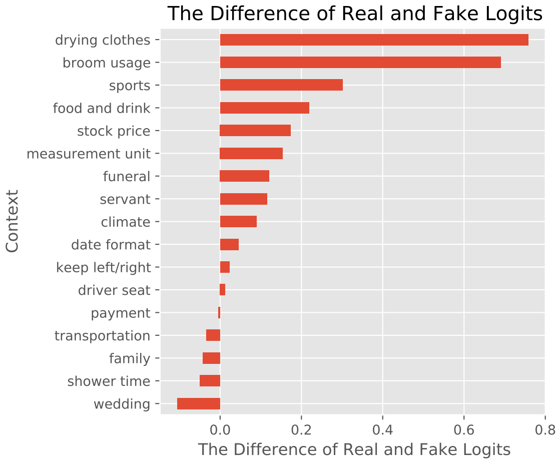


Figure 6: The difference between real and fake logits in each context. The higher difference shows that ELECTRA was better able to detect the cultural difference in that context.

of sentences in each context grouped by fake or real sentences. We observe that fake sentences have higher logits than real sentences. This reveals the relative knowledge of ELECTRA about different cultures. However, on average, ELECTRA did not assign positive logits which shows that it is not necessarily considering inaccurate cultural common sense sentences as completely fake. This can be attributed to its training procedure. As ”date format” had the closest scores to fake, we showed the sentences and ELECTRA's score for each token in Figure 5. This sentence is not very eloquent and if we give this sentence to BERT and mask the word ”after”, the first generated result would be ”and” not ”after”. Therefore, it is expected that ELECTRA finds this token fake. This observation also shows the importance of wording when leveraging the pre-training objective of ELECTRA as is.

To better grasp the knowledge of ELECTRA about different contexts, Figure 6 shows the difference of logits between true and fake sentences. The easiest contexts for ELECTRA have been

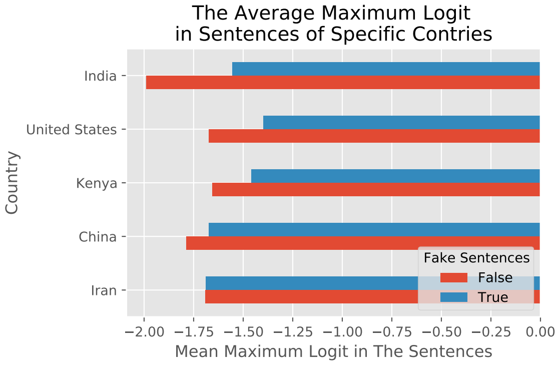
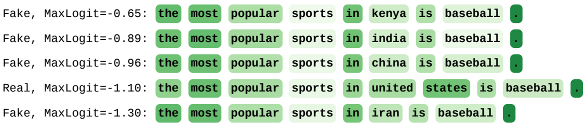


Figure 7: The average maximum logit in sentences of each country grouped by being fake or not. Higher logits are less real in ELECTRA's view. The countries are ordered by the difference between real and fake logits in their sentences.



```
Fake, MaxLogit=-0.65: the most popular sports in kenya is baseball .
Fake, MaxLogit=-0.89: the most popular sports in india is baseball .
Fake, MaxLogit=-0.96: the most popular sports in china is baseball .
Real, MaxLogit=-1.10: the most popular sports in united states is baseball .
Fake, MaxLogit=-1.30: the most popular sports in iran is baseball .
```

Figure 8: Examples of Sports context with the replaced token being baseball. Lower logits mean real for ELECTRA. The order is correct except for Iran.

about ”drying clothes” in sentences like ”American/Iranian/Kenyan/Indian/Chinese people dry their wet clothes in the machines/sun.”, and ”broom usage” in sentences like ”in Iran, it is rare/common that people use broom to clean the floor”. The hardest contexts were ”wedding”, ”shower time” and ”family” such as ”in Iran, it is rare/common that adults live with their parents.”. This example shows that some of these hard contexts are also controversial. For instance, based on the dataset, In India and China, it is common for adults to live with their parents; however, this also might be debated in Iran (Actually ELECTRA believes ”common” is more likely for Iran than ”rare”).

7.4 Countries

Another interesting topic to analyze is the knowledge of ELECTRA about different countries. To demonstrate this, Figure 7 shows the difference in logits for real and fake sentences in each country. It indicates that ELECTRA knew most about India even more than the United States. This is consistent with (Yin et al., 2022) results which hypothesized that knowledge about one country is more available in the texts of other countries and it is too common for one country to be mentioned in

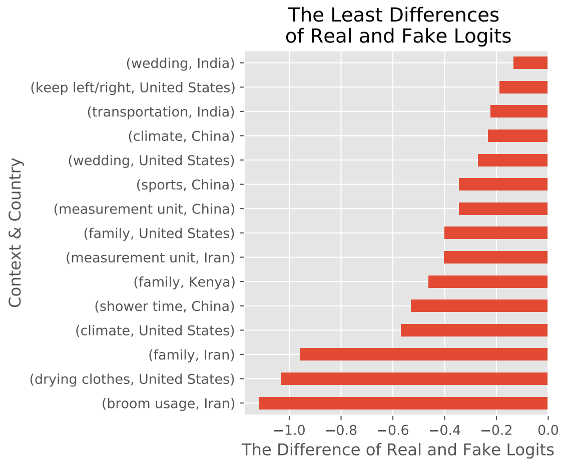


Figure 9: The least differences between real and fake logits in each context and country. These are the contexts and countries that ELECTRA made a mistake about.

its own language.

On the other hand, ELECTRA knows less about Kenya, China, and Iran. Figure 8 illustrates the context of sport where the replaced token is baseball in different countries. This sentence about baseball is only real for the united states. ELECTRA correctly assigns lower logit (more real) to the United States sentence than Kenya, India, and China. Nonetheless, Electra is assigning higher logit to Iran's sentence, which is incorrect and is consistent with the general trend seen earlier that ELECTRA's knowledge about Iran is more limited than about other countries.

8 Error analysis

In this section, we explore the sentences that ELECTRA got wrong. Figure 9 illustrates the context and countries which got the least difference between their true and fake sentences. As the difference is negative, it means that ELECTRA has assigned real and fake to these contexts and countries in reverse order. For instance, on broom usage in Iran, ELECTRA believes that ”in Iran, it is common that people use broom to clean the floor” instead of being rare. In this case, one can argue that ELECTRA's belief is correct, or argue that ELECTRA's knowledge about Iran is biased and outdated. in another case, ELECTRA believes that ”Chinese people usually take shower in the morning”, while the correct time is ”in the evening”. This is interesting trivia, but even ChatGPT cannot get this right[1]. In short, the errors of ELECTRA consist of controversial contexts, too specific information, and lack of knowledge such as not knowing the measurement units used in Iran.

9 Contributions of group members

- Samin Mehdizadeh: Did data collection/processing and video narration and wrote the reports. Tested BERT, ELECTRA, RoBERTa, and XLM-RoBERTa in various setups.

- Mohsen Fayyaz: Prepared the slides for video presentation, wrote reports, and conducted EDA on the prepared dataset. Explored the results and did the error analysis.

10 Conclusion

In this work, we explored the cultural knowledge of ELECTRA. Using the GeoMLAMA dataset and utilizing ELECTRA's inherent token evaluation system learned by its pre-training objective, we showed that ELECTRA generally knows the difference between real and fake cultural sentences. However, this knowledge varies in different contexts and countries. Our results showed that ELECTRA has more information about India and the United States than China and Iran. In our error analysis, we showed that some contexts of GeoMLAMA are controversial and in some cases, ELECTRA lacks cultural knowledge. We also conducted experiments in zero- and few-shot settings. Our results demonstrated that Cultural questions are difficult to answer for tested models, and these models perform similarly to a random model given a small number of training data. Testing a fine-tuned model on question-answering data shows that this poor performance could even be because the models did not understand the defined task well. Also, evaluations on the classifier show that when using the representation produced for a cultural sentence, the representation obtained by Electra was more useful for classification, which can indicate that this model has better encoded the cultural knowledge. Our work was restricted to using the maximum logit of all tokens, one dataset, and the base model of ELECTRA due

[1] When do Chinese people shower in the day? ChatGPT: Generally speaking, Chinese people tend to shower in the morning or evening, depending on their own individual routines.

to limited resources. Exploring different methods for aggregating token scores of ELECTRA, testing its knowledge on different cultural datasets, and leveraging the knowledge of larger ELECTRA can be interesting future work.

References

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Çelikyilmaz, A., and Choi, Y. (2019). Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chen, W., Chang, Y., Zhang, R., Pu, J., Chen, G., Zhang, L., Xi, Y., Chen, Y., and Su, C. (2022). Probing simile knowledge from pre-trained language models. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5875–5887. Association for Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., and Elliott, D. (2021). Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peters, M. E., Neumann, M., Zettlemoyer, L., and Yih, W.-t. (2018). Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Razeghi, Y., IV, R. L. L., and Singh, S. (2020). Deriving behavioral tests from common sense knowledge graphs.

Shi, X., Padhi, I., and Knight, K. (2016). Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

West, P., Bhagavatula, C., Hessel, J., Hwang, J., Jiang, L., Le Bras, R., Lu, X., Welleck, S., and Choi, Y. (2022). Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xia, M., Artetxe, M., Du, J., Chen, D., and Stoyanov, V. (2022). Prompting electra: Few-shot learning with discriminative pre-trained models. *arXiv preprint arXiv:2205.15223*.

Yin, D., Bansal, H., Monajatipoor, M., Li, L. H., and Chang, K.-W. (2022). Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *EMNLP*.

Yin, D., Li, L. H., Hu, Z., Peng, N., and Chang, K.-W. (2021). Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Zero-shot and Few-shot results

Model	**Accuracy**	**F1**	**Recall**	**Precision**
Zero-Shot				
ELECTRA	0.544000	0.544283	0.544000	0.544902
BERT	0.556267	0.556500	0.556267	0.558468
ROBERTA	0.465067	0.464882	0.465067	0.464729
XLM-ROBERTA	0.597333	0.597599	0.597333	0.599129
Few-Shot				
K = 16				
ELECTRA	0.501082	0.501377	0.501082	0.502052
BERT	0.500000	0.500310	0.500000	0.501239
ROBERTA	0.504329	0.504636	0.504329	0.505623
XLM-ROBERTA	0.500000	0.500289	0.500000	0.501563
K = 128				
ELECTRA	0.475995	0.476358	0.475995	0.477668
BERT	0.475995	0.476358	0.475995	0.477668
ROBERTA	0.475995	0.476358	0.475995	0.477668
XLM-ROBERTA	0.475995	0.476358	0.475995	0.477668
K = 500				
ELECTRA	0.516219	0.516496	0.516219	0.51706
BERT	0.516219	0.516496	0.516219	0.51706
ROBERTA	0.516219	0.516496	0.516219	0.51706
XLM-ROBERTA	0.516219	0.516496	0.516219	0.51706

Table 4: Yes-no question results

Model	**Accuracy**	**F1**	**Recall**	**Precision**
Zero-Shot				
ELECTRA	0.337067	0.402706	0.337067	0.318496
BERT	0.347200	0.416627	0.347200	0.520865
ROBERTA	0.342933	0.412367	0.342933	0.517268
XLM-ROBERTA	0.350400	0.417698	0.350400	0.518272
Few-Shot				
K = 16				
ELECTRA	0.317784	0.318038	0.317784	0.318496
BERT	0.312536	0.312521	0.312536	0.312515
ROBERTA	0.326531	0.326980	0.326531	0.327994
XLM-ROBERTA	0.325364	0.325540	0.325364	0.326065
K = 128				
ELECTRA	0.322232	0.322898	0.322232	0.324649
BERT	0.319532	0.320050	0.319532	0.320972
ROBERTA	0.338434	0.338763	0.338434	0.339636
XLM-ROBERTA	0.323132	0.323109	0.323132	0.324299
K = 500				
ELECTRA	0.362025	0.363005	0.362025	0.36498
BERT	0.372152	0.373223	0.372152	0.375652
ROBERTA	0.379747	0.378882	0.379747	0.378283
XLM-ROBERTA	0.346835	0.347152	0.346835	0.350459

Table 5: Multiple choice questions results