My Happy Model (test)! : Model Card

Model Details

- Blackbox external model access: 1
- Capabilities demonstration: 1
- Capabilities description: 1
- Centralized model documentation: 1
- Evaluation of capabilities: 1
- External model access protocol: 1
- External reproducibility of capabilities evaluation: 1
- External reproducibility of intentional harm evaluation: 0
- External reproducibility of mitigations evaluation: 0
- External reproducibility of trustworthiness evaluation: 0
- External reproducibility of unintentional harm evaluation: 0
- Full external model access: 1
- Inference compute evaluation: 0
- Inference duration evaluation: 1
- Input modality: 1
- Intentional harm evaluation: 0
- Limitations demonstration: 0
- Limitations description: 1
- Mitigations demonstration: 0
- Mitigations description: 0
- Mitigations evaluation: 0
- Model architecture: 1
- Asset license: 1
- Model components: 1
- Model size: 1
- Output modality: 1
- Risks demonstration: 0
- Risks description: 0
- Third party capabilities evaluation: 0
- Third party evaluation of limitations: 1
- Third party mitigations evaluation: 0
- Third party risks evaluation: 0
- Trustworthiness evaluation: 0
- Unintentional harm evaluation: 0

Intended use

- Natural language processing tasks, including but not limited to translation, sentiment analysis, and question answering.

- Cross-lingual understanding and generation tasks.

- Instruction-based prompt generation for a wide range of languages.

- Zero-shot and few-shot learning applications.

- Exploratory data analysis and research in multilingual language model capabilities.

Factors

- Language support and proficiency across a broad spectrum of languages.

- The clarity and specificity of instruction prompts.

- Model scalability and performance across different sizes from 300M to 176B parameters.

- Generalization abilities to unseen tasks and languages.

- Accessibility and ease of use for researchers and developers with different levels of resources.

Metrics

TBD?

Evaluation data

- Description: A diverse set of evaluation tasks covering coreference resolution, natural language inference, sentence completion, and program synthesis across multiple languages.

- Description: Datasets from the Winogrande, ANLI, XNLI, and HumanEval evaluations, allowing for an extensive assessment of model performance in both seen and unseen languages.

- Description: Validation and test splits are utilized from the respective datasets to ensure unbiased evaluation.

- Description: Multilingual task evaluation employing prompts in both English and the respective native languages to gauge cross-lingual transfer capabilities.

- Description: Benchmarking against existing models like XGLM, T0, and GPT to understand the competitive landscape.

Training data

- Description: The model utilizes the BIG-bench xP3 dataset for training, promoting a wide coverage of tasks and languages.

- Description: Incorporation of code and programming languages alongside natural languages to enhance the model's versatility.

- Description: Utilized datasets such as BIG-bench, ROOTS, and a subset of the mC4 corpus to provide rich, diverse linguistic and task coverage.

- Description: Finetuning approach on xP3, xP3mt, and P3 datasets to enable cross-lingual generalization and effective prompt-based task performance.

- Description: Leverages both pretrained (BLOOM, mT5) and bespoke large language models across various sizes for targeted task learning.

Environmental

- Carbon emitted(tCO2eq):

- Energy consumption:

- Energy unit:

- Compute hours:

Ethical considerations

- Potential for biased or inaccurate outputs across less-supported languages, requiring careful validation.

- Use of the model in applications with impactful consequences should be approached with caution.

- Need for transparency regarding the training data sources and model limitations to users.

- Ethical considerations around data privacy and consent, especially in multilingual contexts.

- Awareness of cultural sensitivity and potential for reinforcing stereotypes must be considered in model application and development.

Recommendations

- Employment of early stopping, addition of long tasks, and minimum generation length forcing for improved generative task performance.

- Fine-tuning with both English and machine-translated multilingual prompts for enhanced cross-lingual abilities.

- Utilization of the model in research to explore and expand the boundaries of zero-shot learning across languages.

- Adoption of ethical and fair use practices, considering the model's broad linguistic capabilities.

- Engagement with the BigScience community for collaborative research and development efforts.

Additional information

- The project is conducted under the BigScience initiative, allowing for open collaboration and research.

- Models are released under RAIL and Apache 2.0 licenses for wide accessibility and use.

- Fine-tuned models incorporate biases towards short answers, affecting performance on generative tasks.

- Language contamination analysis in the pretraining corpus shows unintentional learning from 'unseen' languages.

- Recommendations include using a specific prompting format and considering model size according to task requirements.

Quantitative Analyses

TBD?