petals-team / StableBeluga2

Downloads last month: 132,919

Model Description

Stable Beluga 2 is a Llama2 70B model finetuned on an Orca style Dataset. This repository contains the model from the stabilityai/StableBeluga2 repository with the following changes:

- **Storing weights in bfloat16 instead of float32.** This leads to 2x smaller files and a small quality loss, which is not significant compared to the loss caused by NF4 quantization used in Petals by default.
- **Storing weights in small shards.** Each transformer block is stored in its own shard (1.71 GB each). The input and output embeddings and adjacent layernorms are in a separate shard (1.05 GB) too. This way, Petals clients and servers don't have to download any excess data besides the layers they actually use.
- **Using Safetensors instead of Pickle.** This allows faster loading with smaller RAM requirements.

Model Details

Developed by: Stability AI
Model type: Stable Beluga 2 is an auto-regressive language model fine-tuned on Llama2 70B.
Language(s): English
Library: HuggingFace Transformers
License: Fine-tuned checkpoints (Stable Beluga 2) is licensed under the STABLE BELUGA NON-COMMERCIAL COMMUNITY LICENSE AGREEMENT
Contact: For questions and comments about the model, please email lm@stability.ai

Training Dataset

Stable Beluga 2 is trained on our internal Orca-style dataset. Training data is a synthetic dataset that was created to enhance the small model's reasoning abilities. The dataset comprises a diverse collection of tasks aimed at training AI models across various domains, focusing on cautious reasoning and alignment with ethical guidelines. It includes approximately 602,000 zero-shot queries grouped into 23 categories and 126 sub-categories, each sharing a common instruction format to promote consistency. The dataset also features 55,000 few-shot samples to encourage the model's ability to learn from context, around 160,000 math problems sourced from a variety of existing datasets, and 2,000 synthetically generated conversations between doctors and patients designed to test the model's specialized skills.

Intended Uses

The pretrained-only model can be used for prompting for evaluation of downstream tasks as well as text generation. In addition, the model can be fine-tuned on a downstream task. For all other checkpoints, please have a look at the model hub.

Summary of Model Risks by Example Uses

We identified **14** potential model risks and **18** mitigation strategies (Ⓜ) for **4** potential model uses

Applicable to the use
Not applicable to the use
Risk resulted in real-world harm*

Risks	Recommending personalized content (U1)	Creating personalized ad campaigns (U2)	Personalizing learning experiences (U3)	Assisting medical diagnoses (U4)
R1 Reflects offensive or biased content in model generated text despite finetuning	■	■	■	■
R2 Exhibits unreliable, unsafe or other undesirable behaviors	■	■	■	■
R3 Limits accuracy in areas underrepresented in the training dataset	■	■	■	■
R4 Facilitates the spread of misinformation by fabricating content making it unreliable for critical decisions	■	■	■	□
R5 Acts as 'black boxes' making it difficult to comprehend rationale behind outputs	□	■	■	■
R6 Carries biases present in the source data	■	■	□	■
R7 Exhibits limited real-world understanding	■	□	□	■
R8 Fails to cover all scenarios	□	■	■	□
R9 Harms economic interests of content creators by using their work without compensation	□	■	□	□
R10 Harms individuals' reputations by potentially revealing private medical information	□	□	□	■
R11 Increases psychological harm by exposing users to graphic and explicit content	■	□	□	□
R12 Infringes data protection laws by using sensitive data without proper authorization	□	■	□	□
R13 Undermines trust in AI systems by mishandling sensitive personal data	■	□	□	□
R14 Underperforms in non-English languages	■	□	□	□

Details of the Example Uses

U1 Purpose: **Recommending personalized content**
Capability: Analyzing preferences for suggestions
AI User: Streaming platforms
AI Subject: Content consumers
Domain: Recommender Systems and Personalization

U2 Purpose: **Creating personalized ad campaigns**
Capability: Analyzing user behavior from social media posts
AI User: Marketing agencies
AI Subject: Consumers
Domain: Marketing and Advertising

U3 Purpose: **Personalizing learning experiences**
Capability: Analyzing student performance and tailoring content
AI User: Educational platforms
AI Subject: Students
Domain: Education and vocational training

U4 Purpose: **Assisting in medical diagnoses**
Capability: Analyzing patient data and suggesting conditions
AI User: Healthcare professionals
AI Subject: Patients
Domain: Health and Healthcare

Mitigations for Risks

R1 Reflects offensive or biased content in model generated text despite finetuning
Capability risk · Representation and toxicity harms
- Ⓜ Perform safety testing
- Ⓜ Tune model to specific applications
- Ⓜ Prevent harmful responses
- Ⓜ Leverage various content moderation services
- Ⓜ Avoid using models unsuitable for your application

R2 Exhibits unreliable, unsafe or other undesirable behaviors
Capability risk · Information and safety harms
- Ⓜ Perform safety testing
- Ⓜ Prevent harmful responses
- Ⓜ Run your own suite of tests
- Ⓜ Avoid using models unsuitable for your application
- Ⓜ Exercise caution when using models in production systems
- Ⓜ Be mindful of potential issues in generated responses

R3 Limits accuracy in areas underrepresented in the training dataset
Capability risk · Representation and toxicity harms
- Ⓜ Correct through evaluation and fine-tuning prior to deployment
- Ⓜ Requires detailed studies for better quantification of risks
- Ⓜ Need additional analysis to assess potential harm or bias

R4 Facilitates the spread of misinformation by fabricating content making it unreliable for critical decisions
Human interaction risk · Misinformation harms
- Ⓜ Ensure outputs are not hallucinations
- Ⓜ Avoid treating model outputs as sources of truth
- Ⓜ Use the model responsibly
- Ⓜ Avoid using models for applications that may cause harm
- Ⓜ Perform more rigorous measurement, understanding and mitigations

R5 Acts as 'black boxes' making it difficult to comprehend rationale behind outputs
Capability risk · Human autonomy and integrity harms
- Ⓜ Requires detailed studies for better quantification of risks
- Ⓜ Acknowledge the important role of research and open source community

R6 Carries biases present in the source data
Capability risk · Representation and toxicity harms
- Ⓜ Perform safety testing
- Ⓜ Correct through evaluation and fine-tuning prior to deployment
- Ⓜ Leverage various content moderation services
- Ⓜ Perform more rigorous measurement, understanding and mitigations

R7 Exhibits limited real-world understanding
Capability risk · Misinformation harms
- Ⓜ Correct through evaluation and fine-tuning prior to deployment
- Ⓜ Requires detailed studies for better quantification of risks
- Ⓜ Perform more rigorous measurement, understanding and mitigations

R8 Fails to cover all scenarios
Capability risk · Information and safety harms
- Ⓜ Perform safety testing
- Ⓜ Correct through evaluation and fine-tuning prior to deployment
- Ⓜ Exercise reasonable caution when using models in production

R9 Harms economic interests of content creators by using their work without compensation
Systemic risk · Socioeconomic and environmental harms
- Ⓜ Acknowledge the important role of research and open source community
- Ⓜ Follow better regulations and standards from government and technology leaders

R10 Harms individuals' reputations by potentially revealing private medical information
Systemic risk · Information and safety harms
- Ⓜ Leverage various content moderation services
- Ⓜ Follow better regulations and standards from government and technology leaders

R11 Increases psychological harm by exposing users to graphic and explicit content
Human interaction risk · Representation and toxicity harms
- Ⓜ Leverage various content moderation services
- Ⓜ Avoid using models in applications causing harm or distress
- Ⓜ Perform safety testing
- Ⓜ Follow better regulations and standards from government and technology leaders

R12 Infringes data protection laws by using sensitive data without proper authorization
Human interaction risk · Information and safety harms
- Ⓜ Leverage various content moderation services
- Ⓜ Follow better regulations and standards from government and technology leaders

R13 Undermines trust in AI systems by mishandling sensitive personal data
Human interaction risk · Information and safety harms
- Ⓜ Leverage various content moderation services
- Ⓜ Follow better regulations and standards from government and technology leaders
- Ⓜ Benefit from safety guardrails

R14 Underperforms in non-English languages
Capability risk · Representation and toxicity harms
- Ⓜ Correct through evaluation and fine-tuning prior to deployment
- Ⓜ Requires detailed studies for better quantification of risks
- Ⓜ Need additional analysis to assess potential harm or bias

Glossary

Risks
Capability risk: emerges from the technical components of the model
Human interaction risk: emerges from the experience of people interacting with the model
Systemic risk: emerges from the impact of the system on the broader systems in which it is embedded, such as society, the economy, and the natural environment

Harms
* real-world harm sourced from the AI Incident Database [www.incidentdatabase.ai]
Malicious use: emerges when the model lowers costs and barriers for harmful actors to engage in illicit activities
Misinformation harms: emerge when the model under-, over-, or misrepresents certain groups, generating toxic, offensive, abusive, or hateful content
Information and safety harms: emerge when the model leaks, reproduces, generates, or infers sensitive, private, or hazardous information
Representation and toxicity harms: emerge when the model generates and spreads inaccurate or misleading information, causing people to develop false beliefs
Human autonomy and integrity harms: emerge when the model compromises human agency or circumvents meaningful human control
Socioeconomic and environmental harms: emerge when the model exacerbates inequalities or negatively impacts employment, innovation, or the environment