Code Llama: Open Foundation Models for Code

Code Llama is a family of large language models for code based on Llama 2 that provides state-of-the-art performance among open models. It offers infilling capabilities, support for large input contexts, and zero-shot instruction following ability for programming tasks.

Key Features:

- Provides state-of-the-art performance among open models for code generation
- Offers infilling capabilities based on surrounding content
- Supports large input contexts, up to 100k tokens
- Demonstrates zero-shot instruction following ability for programming tasks

Model Variants:

- Foundation models (Code Llama)
- Python specializations (Code Llama - Python)
- Instruction-following models (Code Llama - Instruct)

Model Sizes:

- 7B, 13B, and 34B parameters

Training Dataset:

- 500B additional tokens for Code Llama
- 100B additional tokens for Code Llama - Python

Performance:

- Code Llama reaches state-of-the-art performance among open models on several code benchmarks
- Scores of up to 53% and 55% on HumanEval and MBPP, respectively
- Code Llama - Python 7B outperforms Llama 2 70B on HumanEval and MBPP
- All models outperform every other publicly available model on MultiPL-E

Availability:

- Released under a permissive license that allows for both research and commercial use
- Inference code for both completion and infilling models provided in the accompanying repository

Benefits of Code Llama:

- Improved performance on code generation tasks compared to previous models
- Enhanced ability to handle large input contexts
- Increased proficiency in following instructions for programming tasks
- Wide range of applications, including code completion, debugging, and generating documentation

Limitations:

- While Code Llama demonstrates strong performance, it may still produce inaccurate or objectionable responses in some instances
- Developers should perform safety testing and tuning tailored to their specific applications before deploying any applications of Code Llama

Ethical Considerations:

- Code Llama is a new technology that carries risks with use
- Testing conducted to date has been in English and may not cover all scenarios
- Developers should be aware of these limitations and take appropriate precautions when using Code Llama

Environmental Impact:

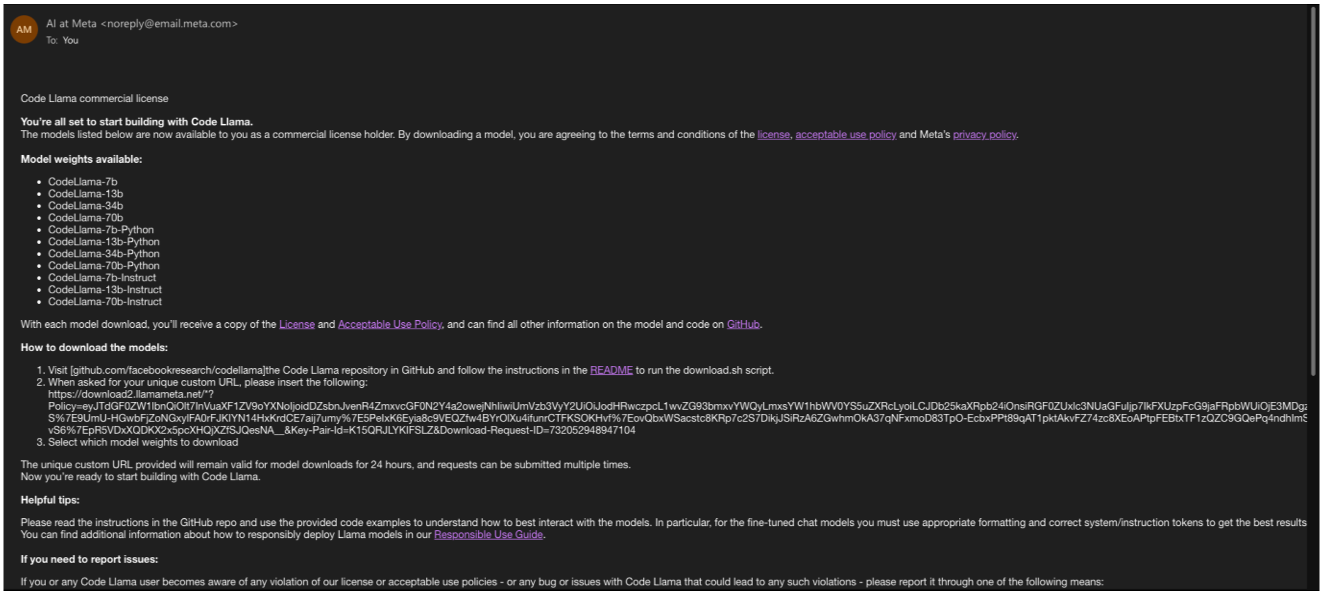
- Training all 9 Code Llama models required 400K GPU hours of computation on hardware of type A100-80GB
- Estimated total emissions were 65.3 tCO2eq, all of which were offset by Meta's sustainability program

Conclusion:

Code Llama represents a significant advancement in the field of large language models for code, offering state-of-the-art performance, infilling capabilities, support for large input contexts, and improved instruction following ability. While it carries some risks and limitations, careful testing and tuning can help ensure its safe and effective use in a wide range of applications.

GitHub:https://github.com/facebookresearch/codellama
Request Access to Llama:https://llama.meta.com/llama-downloads/
Blog: https://ai.meta.com/blog/code-llama-large-language-model-coding/
Research Paper: https://scontent-ord5-2.xx.fbcdn.net/v/t39.2365-6/369856151_1754812304950972_1159666448927483931_n.pdf?_nc_cat=107&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=DcjXd8LiGCkAX-IKFnX&_nc_ht=scontent-ord5-2.xx&oh=00_AfAN7EqHFC2q8Ew3ZENSouPwbA2nZoD9gdIFGE3AqPFknQ&oe=65D805CF

The email from Meta: (Link is no longer valid)



AI at Meta <noreply@email.meta.com>
To: You

Code Llama commercial license

You're all set to start building with Code Llama.
The models listed below are now available to you as a commercial license holder. By downloading a model, you are agreeing to the terms and conditions of the license, acceptable use policy and Meta's privacy policy.

Model weights available:

- CodeLlama-7b
- CodeLlama-13b
- CodeLlama-34b
- CodeLlama-70b
- CodeLlama-7b-Python
- CodeLlama-13b-Python
- CodeLlama-34b-Python
- CodeLlama-70b-Python
- CodeLlama-7b-Instruct
- CodeLlama-13b-Instruct
- CodeLlama-70b-Instruct

With each model download, you'll receive a copy of the License and Acceptable Use Policy, and can find all other information on the model and code on GitHub.

How to download the models:

1. Visit [github.com/facebookresearch/codellama]the Code Llama repository in GitHub and follow the instructions in the README to run the download.sh script.
2. When asked for your unique custom URL, please insert the following:
   https://download2.llamameta.net/*?
   Policy=eyJTdGF0ZW1lbnQiOlt7InVuaXF1ZV9oYXNoIjoidDZsbnJvenR4ZmxvcGF0N2Y4a2owejNhIiwiUmVzb3VyY2UiOiJodHRwczpcL1wvZG93bmxvYWQyLmxsYW1hbWV0YS5uZXRcLyoiLCJDb25kaXRpb24iOnsiRGF0ZUxlc3NUaGFuIjp7IkFXUzpFcG9jaFRpbWUiOjE3MDgz
   S%7E9UmU-HGwbFjZoNGxylFA0rFJKlYN14HxKrdCE7aij7umy%7E5PelxK6Eyia8c9VEQZfw4BYrOlXu4ifunrCTFKSOKHvf%7EovQbxWSacstc8KRp7c2S7DikjJSiRzA6ZGwhmOkA37qNFxmoD83TpO-EcbxPPt89qAT1pktAkvFZ74zc8XEoAPtpFEBtxTF1zQZC9GQePq4ndhImS
   vS6%7EpR5VDxXQDKX2x5pcXHQjXZfSJQesNA__&Key-Pair-Id=K15QRJLYKIFSLZ&Download-Request-ID=732052948947104
3. Select which model weights to download

The unique custom URL provided will remain valid for model downloads for 24 hours, and requests can be submitted multiple times.
Now you're ready to start building with Code Llama.

Helpful tips:

Please read the instructions in the GitHub repo and use the provided code examples to understand how to best interact with the models. In particular, for the fine-tuned chat models you must use appropriate formatting and correct system/instruction tokens to get the best results
You can find additional information about how to responsibly deploy Llama models in our Responsible Use Guide.

If you need to report issues:

If you or any Code Llama user becomes aware of any violation of our license or acceptable use policies - or any bug or issues with Code Llama that could lead to any such violations - please report it through one of the following means: