Generative AI Using Google's Gemma

Build with AI Hands-On Workshop

Baris Inandioglu itsbaris.com • December 5, 2024 • Northeastern University

Baris Inandioglu

Reach out to me!

- Email me at: inandioglu.b@northeastern.edu
- @baris-inandi on GitHub and LinkedIn
- Visit itsbaris.com for more information

What is Gemma?

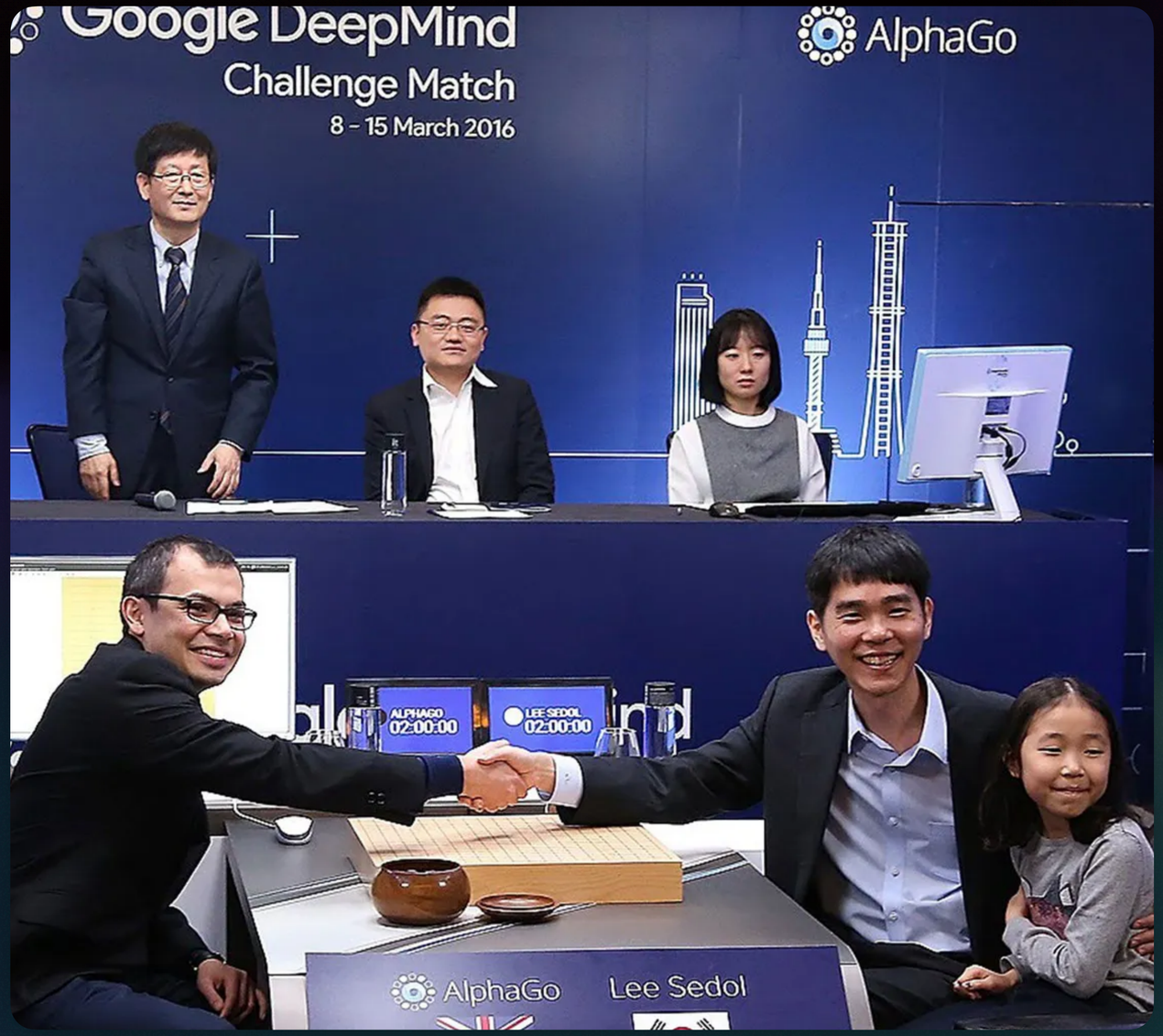
Google’s AI Language Model Released by DeepMind

- Gemma is a family of lightweight, state-of-the art open models built from the same research and technology used to create the Gemini models.
- More geared towards research and for developer use compared to Gemini.

https://developers.googleblog.com/en/gemma-explained-overview-gemma-model-family-architectures/

Need the slides? Go to **itsbaris.com/gemma**

- Same people that made AlphaGo, the deep learning model that beat Professional Go player Lee Sedol in Go.



https://www.wired.com/story/alphago-deepmind-google-wins-lee-sedol/

Need the slides? Go to **itsbaris.com/gemma**

What is Gemma?

The models released

- Gemma 1 (2B, 7B) - Transformer based text-to-text models.
- CodeGemma (2B and 7B) - A fine-tuned version of Gemma, optimized for code completion and generation.
- Gemma 2 (2B, 9B, 27B) - Updated text-to-text models trained with newer architecture with the 2B and 9B versions trained through distillation from larger models.
- RecurrentGemma (2B, 9B) - A model built on the novel Griffin architecture. This is a Gemma model that is not a transformer.
- PaliGemma (3B) - A vision-language model that can take in text and images and provide a text output.

https://developers.googleblog.com/en/gemma-explained-overview-gemma-model-family-architectures/

What are these numbers?

Number of parameters

- The numbers ending with “B”? Gemma 2B? Gemma 7B?
- This is the size of the models in billions of parameters. Larger models typically have more capacity for understanding and generating complex content.

- We will be working with the 2-Billion parameter version of Gemma

Gemma’s Features Are Perfect for Developers

- Google’s Responsible Generative AI Toolkit.
  - Read about it here: https://tinyurl.com/google-responsible-ai
- Works with Google’s Python Machine Learning library TensorFlow.
- Integration with Hugging Face (and alternatives)
  - We will make use of this in our demo!
- Can run on Google Cloud through Vertex AI
- Permits responsible use and distribution for everyone!

Gemma’s Performance Compared to Llama at 7B parameters

Llama is Meta’s Large Language Model

CAPABILITY	BENCHMARK	DESCRIPTION	Gemma 7B	Llama-2 7B	Llama-2 13B
General	MMLU 5-shot, top-1	Representation of questions in 57 subjects (incl. STEM, humanities and others)	64.3	45.3	54.8
Reasoning	BBH -	Diverse set of challenging tasks requiring multi-step reasoning	55.1	32.6	39.4
	HellaSwag 0-shot	Commonsense reasoning for everyday tasks	81.2	77.2	80.7
Math	GSM8K maj@1	Basic arithmetic manipulations (incl. Grade School math problems)	46.4	14.6	28.7
	MATH 4-shot	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	24.3	2.5	3.9
Code	HumanEval pass@1	Python code generation	32.3	12.8	18.3

Need the slides? Go to **itsbaris.com/gemma**

Let’s see Gemma in Action

https://tinyurl.com/try-gemma

Need the slides? Go to **itsbaris.com/gemma**

Let’s see Gemma in Action

This is the Google AI Studio

https://tinyurl.com/try-gemma

Using Gemma as a Developer

Need the slides? Go to **itsbaris.com/gemma**

Using Gemma as a Developer

Getting Started

- What we'll do:
  - Request access to Gemma through **Hugging Face**
  - Get an **access token**
  - Get access to Gemma in **Google Colab** using Python

Need the slides? Go to **itsbaris.com/gemma**

Create a Hugging Face Account

Confirm your email address by
clicking the link sent to your inbox

https://tinyurl.com/join-hugging-face

Request Access for Gemma

Don’t worry, it’s fast and easy.

https://tinyurl.com/access-gemma

Need the slides? Go to **itsbaris.com/gemma**

Claim Your Hugging Face Access Token

This token will give your code access to Gemma.

https://tinyurl.com/get-a-token

Now we are good to go.

huggingface Inbox - Personal 13:45

[Access granted] You have been granted access to the "Google's Gemma models family" gated group

To: Baris Inandioglu

Hi Baris Inandioglu,

You have been granted access to the "Google's Gemma models family" gated group (243 repositories) on huggingface.co.

To view all the gated repositories you have access to, please visit your settings.

Cheers,

The Hugging Face team

Need the slides? Go to **itsbaris.com/gemma**

Create a copy of the Colab Notebook

https://tinyurl.com/gemma-colab

Click **File** > **Save a copy in Drive**

Code is partially based on TirendazAcademy's publicly available GitHub repository at https://github.com/TirendazAcademy/Gemma-Tutorials.

Need the slides? Go to **itsbaris.com/gemma**

You can run Gemma in any application.

Now you know how to leverage Gemma in your own Applications!

Need the slides? Go to **itsbaris.com/gemma**

Thank you for Joining!

Reach out to me!

- Email me at: inandioglu.b@northeastern.edu
- @baris-inandi on GitHub and LinkedIn
- Visit itsbaris.com for more information

Baris Inandioglu

Need the slides? Go to **itsbaris.com/gemma**