NTNU

Norwegian University of Science and Technology

# TDT4310, Lab 1

# Large language models and the latest developments

Tollef Jørgensen, January 16, 2024.

NTNU

Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.

Vinay Uday Prabhu, Abeba Birhane

# Outline

- Introduction to the labs/coursework
- Language models
- GPT and friends (or enemies)
- Latest developments
- Retrieval-augmented generation
- The future and risks

# About ~~me~~ you!

- Extremely smart
- Extremely motivated
- Extremely hyped to sign up for the reference group
  - *Make the course better for everyone involved! pls*

- Email gamback@ntnu.no / tollef.jorgensen@ntnu.no :-)

# Before we continue

- A few have mentioned that there was a bit of confusion regarding where to find information - as we didn't refer to any specific sources for the questions
- The main reason for this:
  - There's no real "perfect" source
  - Getting you to look up and understand the concepts for yourself was the goal, as one source (typically a blog or research paper) will not be enough on its own

# Introduction to the assignments

- **5 total - you need to pass all of them!**
- The first one, on LLMs, is more of a questionaire
  - Perhaps overwhelming to some, but this type of introduction was requested last year!
- **2 weeks per assignment**
  - Published on mondays. The next one will be available on the deadline date.
- All should be finished before easter, so the last week is left for preparations for the graded project.

# Introduction to the assignments

- Information about the course and assignments:
  - **Please mainly use the blackboard forum**
    - Can post anonymously as well
- Sensitive information/other questions:
  - Emails (**please add [TDT4310] in the subject**)
    - If related to an assignment (sick leave, etc.)
      - Me
    - Else:
      - Björn/me + CC TA's

# Language models

- ...Many definitions, but let's go with:
  - *A model that predicts the likelihood of a sequence of tokens (words) in a given context (sentence)*
  - We consider this to be a *probabilistic model*

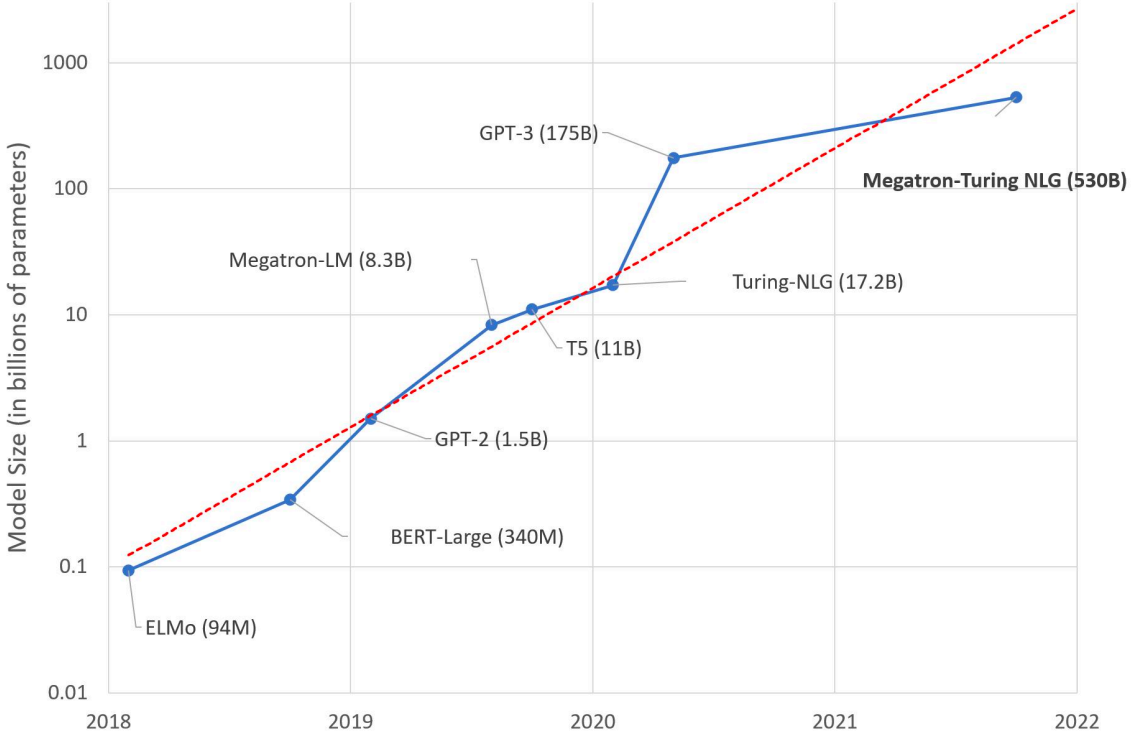
# Language models

- Early approaches:
  - Hidden Markov Models (transition probabilities for a word)
  - N-gram language models
    - "some words in a sentence"
      - (some, words), (words, in), (in, a), (a, sentence)
  - Context-free grammar: rules to generate valid strings
    - https://www.nltk.org/howto/generate.html

# Language models

- Modern:
  - Machine learning/statistical models
  - RNNs and LSTMs, Hierarchical networks, ...
  - Embeddings!
- Since 2018:
  - The transformer architecture and related models

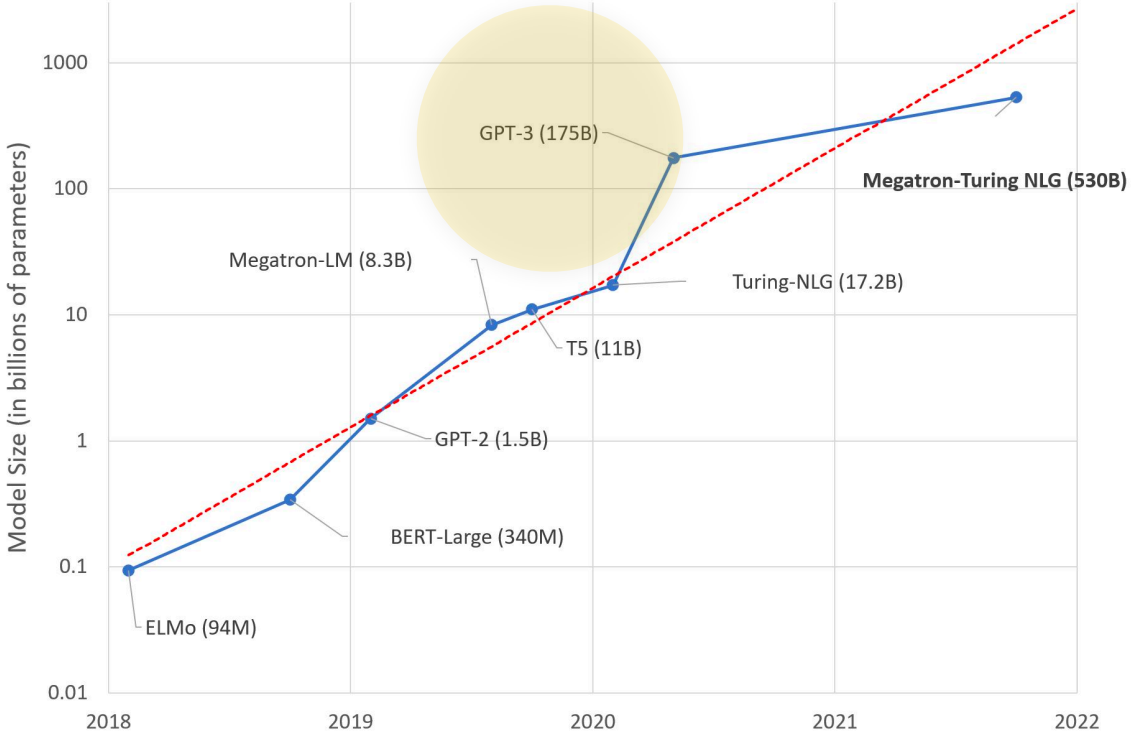
# Language models




https://huggingface.co/blog/large-language-models

# Language models

Similar to the base model of ChatGPT




https://huggingface.co/blog/large-language-models

# GPT and friends (or enemies)

- GPT, or *generative pre-trained transformer*, is nothing *new*, although the public mostly became aware of the term after the release of ChatGPT in 2022
- It was released in its first version back in 2018 by OpenAI
  - Limited data sources and size, although it was definitely big at the time!
  - Based on GPT3, in an updated "GPT 3.5" variant
- Model sizes were scaled up before the release of ChatGPT. This was due to the "scaling laws" we observed for transformer based models
  - More data + more parameters = better models...
  - This, however, is not so simple anymore!

# GPT and friends (or enemies)

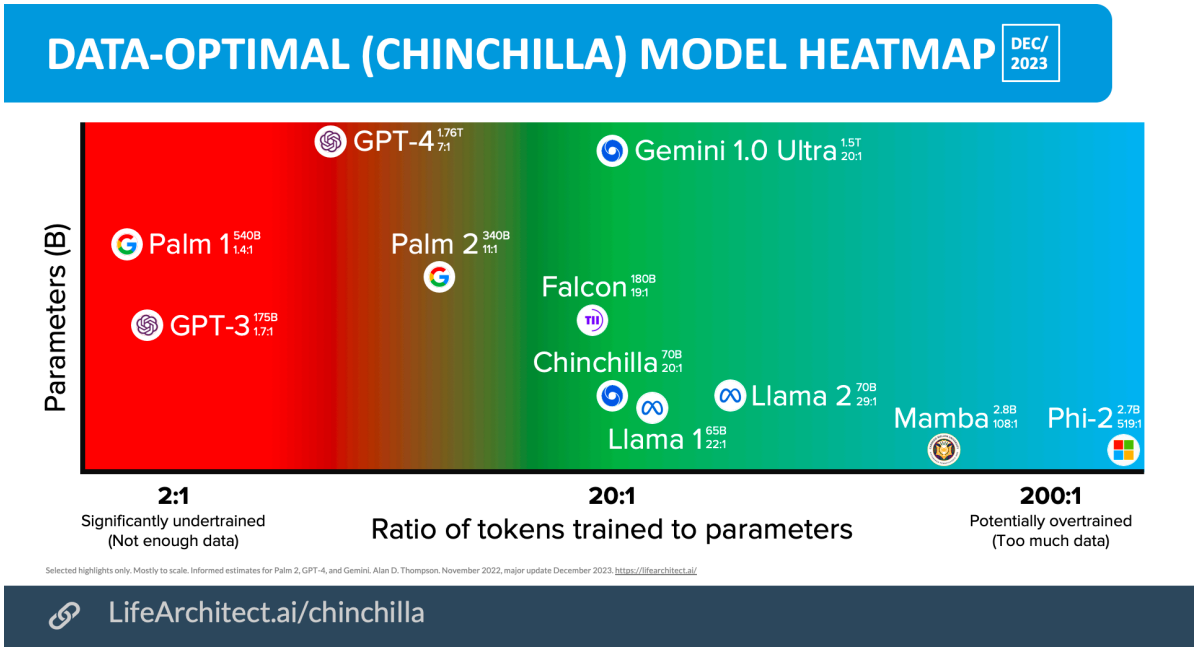
- DeepMind
  - Chinchilla family + Gopher, 2021 (gpt-2 with modifications)
    - from 44M to **280B** parameters
- Google
  - PaLM (2022) and PaLM-2 (2023) , 340-540B
  - Gemini (currently experimental)
- OpenAI
  - GPT series (117M, 1.5B, 175B, ???B) - Rumours say up to 1.7 trillion?!

# GPT and friends (or enemies)

- DeepMind
  - **Chinchilla** family + Gopher, 2021 (gpt-2 with modifications)
    - from 44M to **280B** parameters
- Google
  - PaLM (2022) and PaLM-2 (2023) , 340-540B
  - Gemini (currently experimental)
- OpenAI
  - GPT series (117M, 1.5B, 175B, ???B) - Rumours say up to 1.7 trillion?!

# Chinchilla optimality/scaling laws

- Claimed to find an optimal ratio:
  - Training tokens vs number of parameters



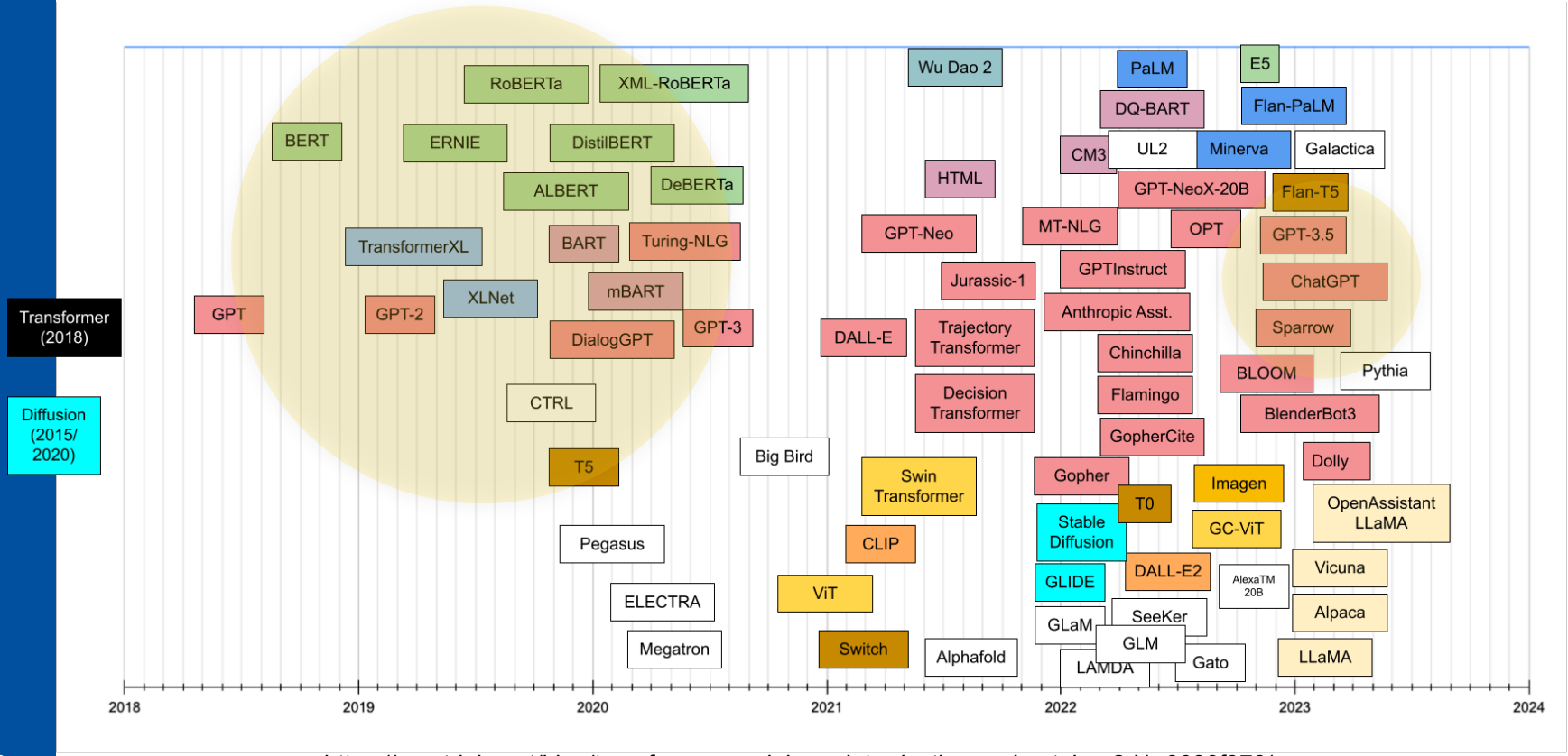
# Chinchilla optimality/scaling laws

- Claim:
  - We need to be using 11× more data during training than that used for GPT-3 and similar models.
  - Need to source, clean, and filter around **33TB** of text data for a **1T-parameter model**.

NTNU

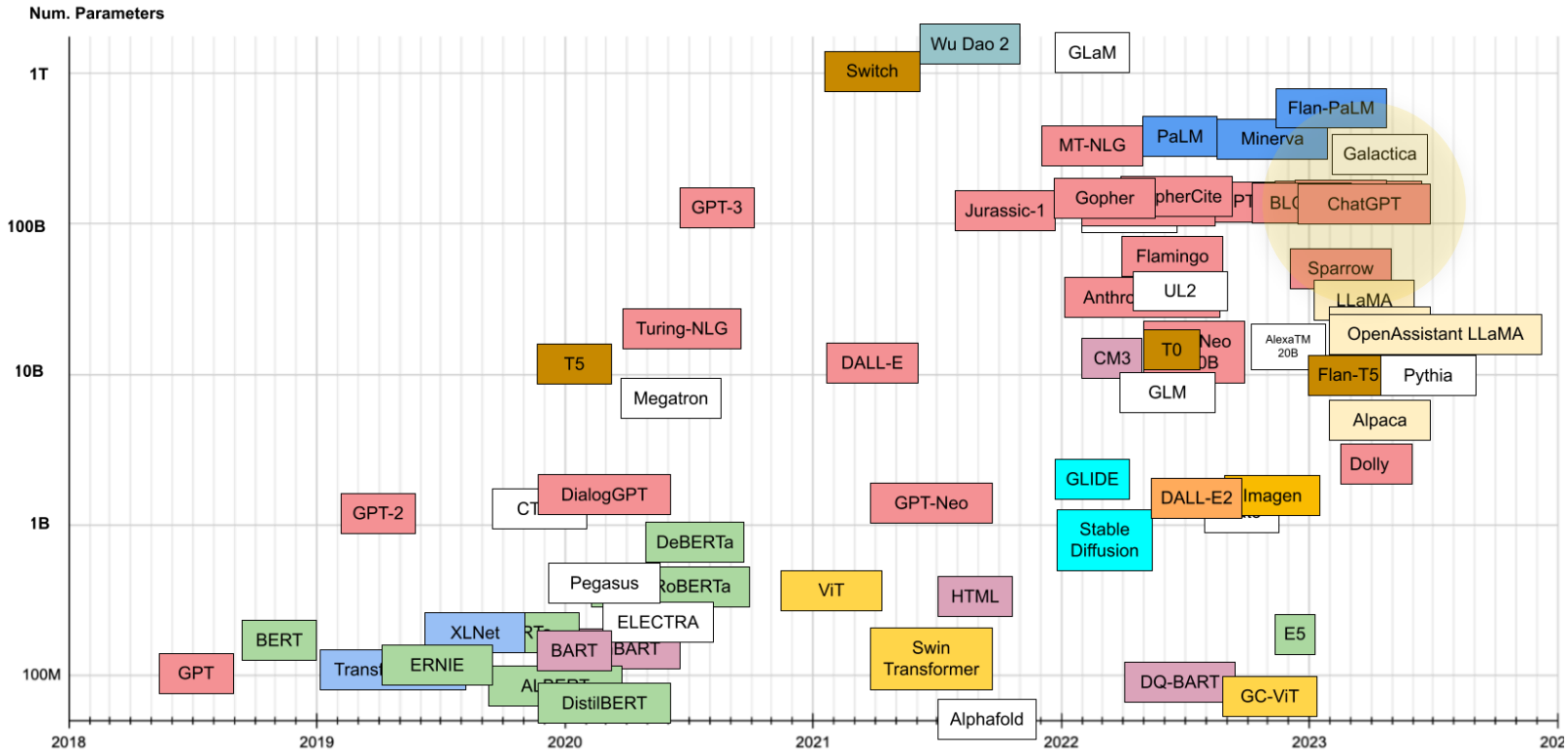
... But there's more ...

# Chronological timeline

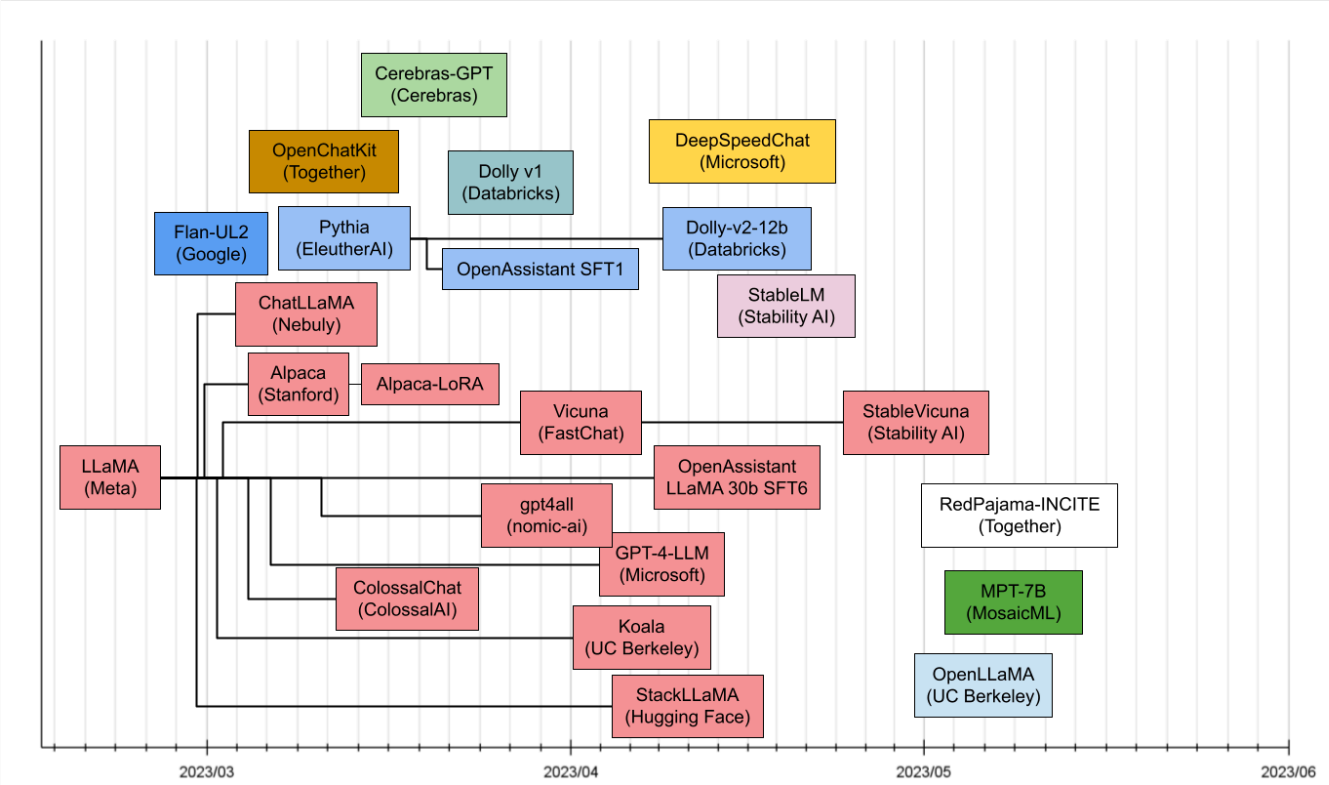



https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/

# Y-axis = N parameters



# The surge of open-sourced LLMs. Thanks, Meta 🤗



Learnable part of the system

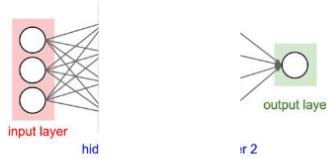
## Rule-based systems

Input → Hand-designed program → Output

## Classical machine learning

Input → Hand-designed features → Mapping from features → Output → Hand- designed loss function

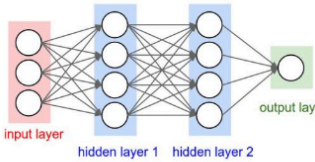
logistic regression




## Deep learning: (self-)supervised learning

Input → Learned features → Mapping from features → Output → Hand- designed loss function

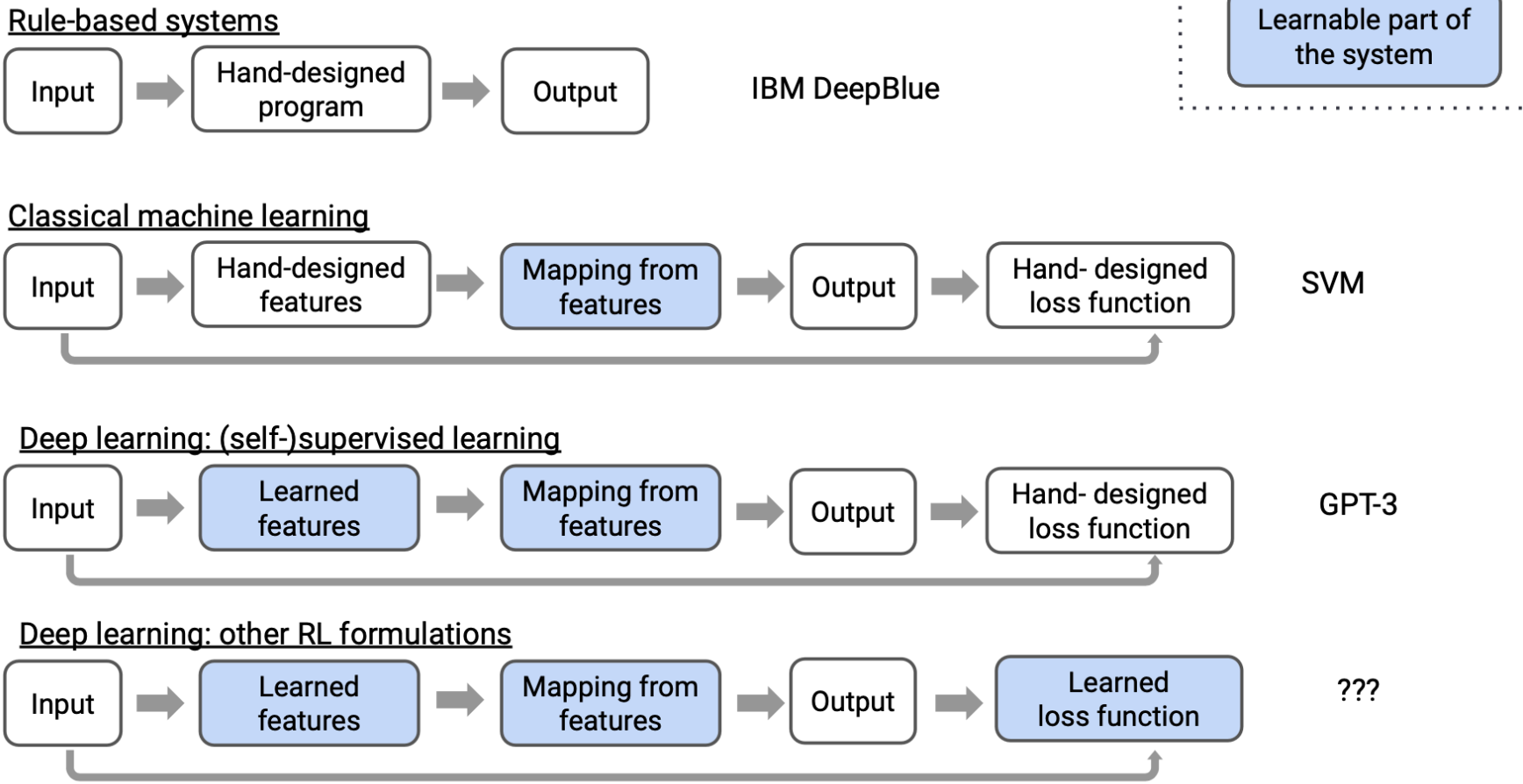
Feedforward neural net




## Deep learning: other RL formulations

Input → Learned features → Mapping from features → Output → Learned loss function

Hyung Won Chung, OpenAI



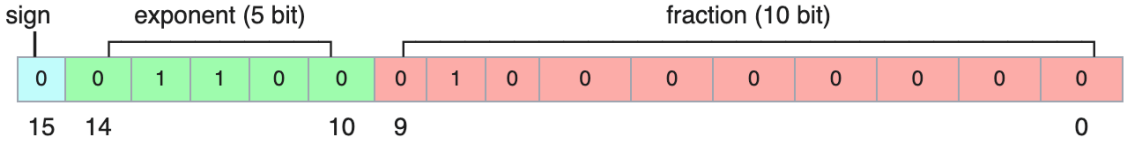

Hyung Won Chung, OpenAI

# A question

- What is the difference between a base model and what we think of when we hear "LLMs" today?
  - One of the questions to this weeks assignment

# The latest stuff

- Running models on consumer-grade hardware with post-training *quantization*
  - A 16-bit floating point requires 2 bytes. If we can reduce this, we reduce both the hardware required for inference + the space on disk.
    - A 70B model (such as LLaMA) requires 70B * 2 bytes
      - 140 GB!
  - Done by adjusting the dynamic range and precision of floating points



# Floating Point Formats

## bfloat16: Brain Floating Point Format

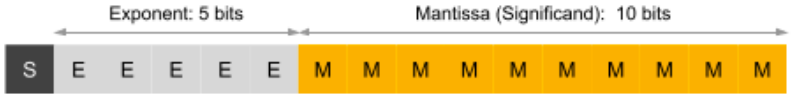
**Range: ~1e$^{-38}$ to ~3e$^{38}$**


## fp32: Single-precision IEEE Floating Point Format

**Range: ~1e$^{-38}$ to ~3e$^{38}$**


## fp16: Half-precision IEEE Floating Point Format

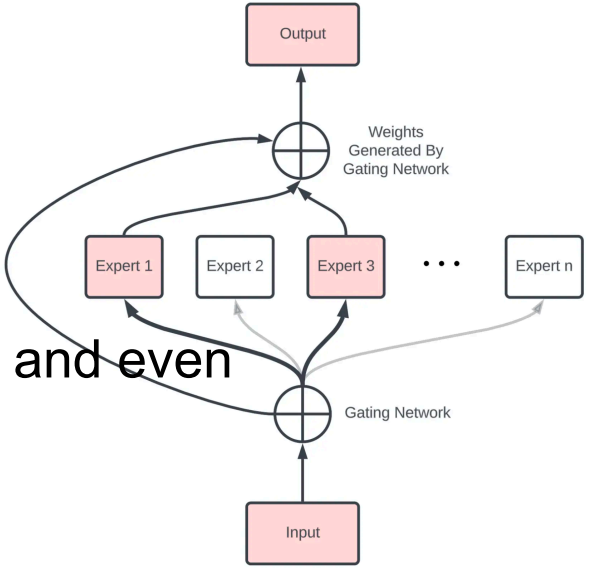
**Range: ~5.96e$^{-8}$ to 65504**




https://cloud.google.com/tpu/docs/bfloat16

# The latest stuff

- Mistral-7B is all the rage
  - A "small" model - only 7B parameters
  - Outperforms larger models like the 13B and even 30B+ models at certain benchmarks
- *Very* recently: Mixtral 8x7B
  - a *mixture-of-experts* model
  - Input is passed to a selection of expert sub-networks
    - One expert may be more suitable than another for a specific task.




MISTRAL AI_

https://mistral.ai/

# The latest stuff

- Even smaller:
  - TinyLlama 1.1B (published a Jan, 2024)
    - Everything open sourced! 🙌
- Still a way to go, but developments are rapid!
  - Open sourcing model weights lead to...
    - Instruction tuning on top of new datasets
    - Merged models
    - Different quantization methods
  - You end up with model names like...
    - OpenAssistant-Llama2-13B-Orca-8K-3319-GPTQ
      - Llama-2 fine-tuned on the orca-chat dataset, done by OpenAssistant, trained for 3319 steps, quantized with GPTQ

# The latest stuff

- Adding information to language models
  - Retrieval-augmented generation (RAG)

- Composing models as *agents*
  - Example: LangChain
    - *https://github.com/langchain-ai/langchain*

# Retrieval-augmented generation

- You've likely heard about the *knowledge cut-off*
  - E.g. in ChatGPT, they need to train or perform continued training on new data up to a certain point in time.
    - Essentially the entire web... Although this is not disclosed by OpenAI.
    - Currently cut-off to January, 2022.
- What is an obvious limitation to the way LLMs are trained?
  - Consider this: we are interacting with the existing weights of the model

# Retrieval-augmented generation

- Retrieval-augmented generation (RAG)
  - Add information into the context of the LLM
- Information can be
  - Knowledge bases (or just data from e.g. wikipedia or any other source)
  - Incorporate searches
  - ∞

# The future

- New architectures (Mamba is up-and-coming, moving away from the transformer)
- Ethical considerations
- Reliability and explainability
- Smaller models
  - Current LLMs are both expensive and energy-consuming
- Evaluation approaches

# The future - Evaluation

- Currently, most teams behind models evaluate on a selection of benchmarks
  - Monitored on LLM leaderboard:
    - https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

*A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...*

*A. rinses the bucket off with soap and blow dry the dog's head.*

*B. uses a hose to keep it from getting soapy.*

***C. gets the dog wet, then it runs away again.***

*D. gets into a bath tub with the dog.*

Example from *HellaSwag*

# The future - Evaluation

- More trustworthy/realistic:
  - Let users ask questions and rank answers anonymously
    - No inclusion of predefined questions
  - https://chat.lmsys.org/
  - https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

| Rank | Model | Arena Elo | 95% CI | Votes | Organization | License |
|---|---|---|---|---|---|---|
| 1 | GPT-4-Turbo | 1249 | +14/-13 | 23069 | OpenAI | Proprietary |
| 2 | GPT-4-0314 | 1190 | +14/-14 | 16237 | OpenAI | Proprietary |
| 3 | GPT-4-0613 | 1160 | +14/-12 | 20884 | OpenAI | Proprietary |
| 4 | Mistral Medium | 1150 | +15/-13 | 6586 | Mistral | Proprietary |
| 5 | Claude-1 | 1149 | +15/-13 | 16956 | Anthropic | Proprietary |
| 6 | Claude-2.0 | 1131 | +14/-13 | 11204 | Anthropic | Proprietary |
| 7 | Mixtral-8x7b-Instruct-v0.1 | 1123 | +15/-13 | 12469 | Mistral | Apache 2.0 |
| 8 | Gemini Pro (Dev) | 1120 | +18/-18 | 1898 | Google | Proprietary |
| 9 | Claude-2.1 | 1119 | +14/-12 | 20883 | Anthropic | Proprietary |
| 10 | GPT-3.5-Turbo-0613 | 1116 | +13/-13 | 26583 | OpenAI | Proprietary |

# Risks

- Environmental and financial costs
  - Minimal improvements lead to massive costs...
- Unmanageable training data
  - And where does the data come from?
  - Is it representative? (e.g. Reddit data)
- Research trajectories
  - Do we just want to improve on benchmark results?
  - *Real* use-cases

Source: Emily M. Bender, *On the dangers of stochastic parrots*

# Risks

- Potential harms of synthetic language
  - Stereotypes, hate speech
  - Training on synthetic data...
    - With more generated data, we keep training on the generated data as *true*
      - Could degrade the quality of LLMs at scale

Source: Emily M. Bender, *On the dangers of stochastic parrots*

NTNU

Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.

Vinay Uday Prabhu, Abeba Birhane