Build Your Own Chatbot

September 17, 2025

Why care about local?

- Privacy

- Privacy
- Environmental

- Privacy
- Environmental
- Cost

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

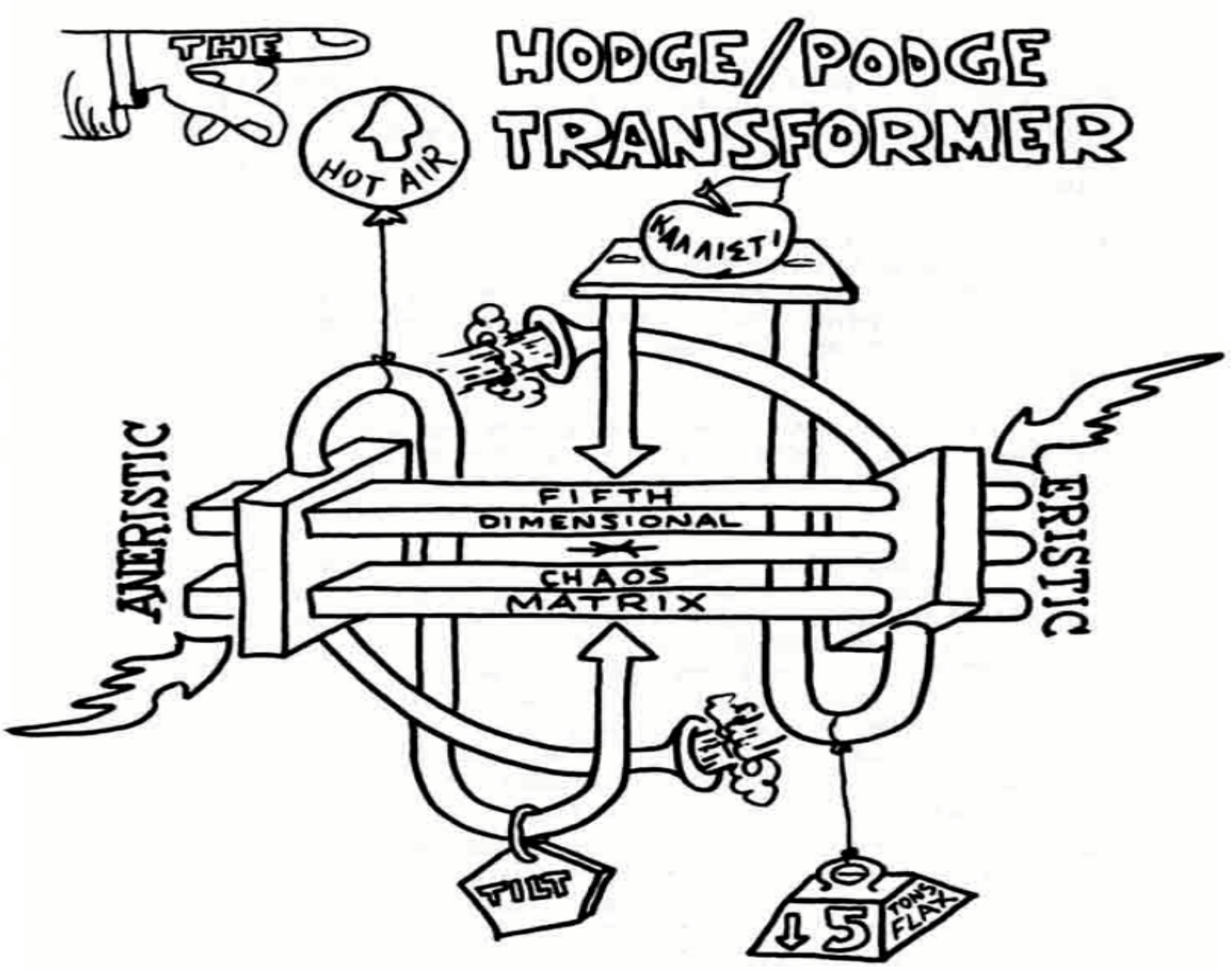
Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com


THE
HOT AIR
HODGE/PODGE
TRANSFORMER
ΚΑΛΛΙΣΤΙ
AERISTIC
FIFTH
DIMENSIONAL
CHAOS
MATRIX
ERISTIC
FILT
15
TOMS
FLAX

Important concepts for transformer-based models

- Context Window and Tokens

Important concepts for transformer-based models

- Context Window and Tokens
- Temperature

Important concepts for transformer-based models

- Context Window and Tokens
- Temperature
- Parameters

Important concepts for transformer-based models

- Context Window and Tokens
- Temperature
- Parameters
- Training

Important concepts for transformer-based models

- Context Window and Tokens
- Temperature
- Parameters
- Training
- The Random Seed.

Important concepts for transformer-based models

- Context Window and Tokens
- Temperature
- Parameters
- Training
- The Random Seed.
- The Prompt, aka "Programming for English Majors"

Andrej Karpathy
@karpathy

The hottest new programming language is English

8:14 PM · Jan 24, 2023 · **4.6M** Views

831 5.3K 33K 2.7K

- Context Window is the ”memory” of an LLM

- Context Window is the "memory" of an LLM
- And Tokens – words, roughly – fill up that "memory"

- Context Window is the "memory" of an LLM
- And Tokens – words, roughly – fill up that "memory"
- And the *response* also takes tokens.

temperature

- Is the ”entropy” of a model’s response

temperature

- Is the "entropy" of a model's response
- Low temperature tends to hew towards predictability and repetitiveness

temperature

- Is the "entropy" of a model's response
- Low temperature tends to hew towards predictability and repetitiveness
- High temperatures make models get...goofy.

Training

- Usually done on text corpuses

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- (cough) books3, others

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- (cough) books3, others
- And other terms like RLHF (Reinforcement Learning from Human Feedback)

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- (cough) books3, others
- And other terms like RLHF (Reinforcement Learning from Human Feedback)
- The larger the model, the more resources it takes to train or re-train.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.
- Which is why, say, GPT-5 requires this....

And you can run a 7B model on this....

And models get even smaller. Good (or at least amusing) results can be had from 1B and 0.5B parameters!

what is *quantization*

- Consider TIFF files vs jpegs or FLACs vs mp3s

what is *quantization*

- Consider TIFF files vs jpegs or FLACs vs mp3s
- It's a way to *drastically* reduce compute needs at the expense of some level of fidelity

what is *quantization*

- Consider TIFF files vs jpegs or FLACs vs mp3s
- It's a way to *drastically* reduce compute needs at the expense of some level of fidelity
- Without quantization, you pretty much need big GPUs – 16GB Nvidias seem to be the base level

what is *quantization*

- Consider TIFF files vs jpegs or FLACs vs mp3s
- It's a way to *drastically* reduce compute needs at the expense of some level of fidelity
- Without quantization, you pretty much need big GPUs – 16GB Nvidias seem to be the base level
- Quantized models *can* use GPUs, and indeed function better if they do, but they don't *need* them