

Transformer Fundamentals with GPT-1

Jacob Danner

Language Modeling

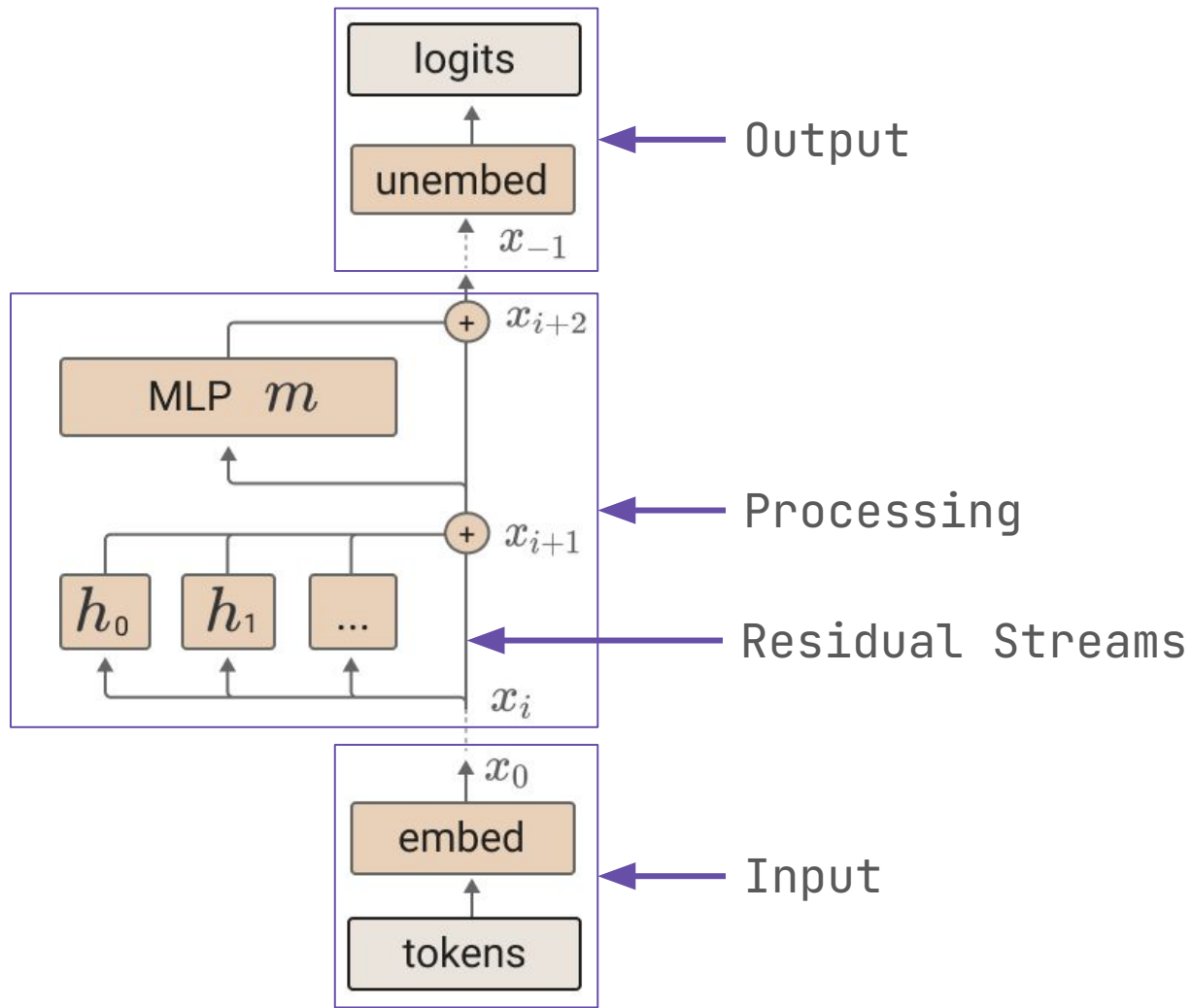
- **Defn:** A language model is a system that assigns probabilities to sequences of words, enabling it to generate, predict, or evaluate text.
- Old approaches:
 - N-gram models
 - Hidden Markov models
 - Feedforward neural networks
 - Recurrent neural networks
- *"Attention Is All You Need"* (2017, Google)
 - Encoder-Decoder Transformer
 - Translation

The GPT Family

- *"Improving Language Understanding by Generative Pre-Training"* (2018, OpenAI)
 - Generative
 - Pre-trained
 - Transformer (decoder)
 - Specs:
 - 40 GB of text (~1 billion tokens)
 - 117 million parameters
- Descendents:
 - ChatGPT
 - Claude
 - Gemini
 - etc.

GPT: What to Know

- Input
 - Tokens
 - Embed
- Processing (transformer blocks)
 - Attention (multi-head self attention)
 - QK (scoring)
 - OV (mixing)
 - MLP
 - Residual stream (the information highway)
- Output
 - Unembed
 - Sampling

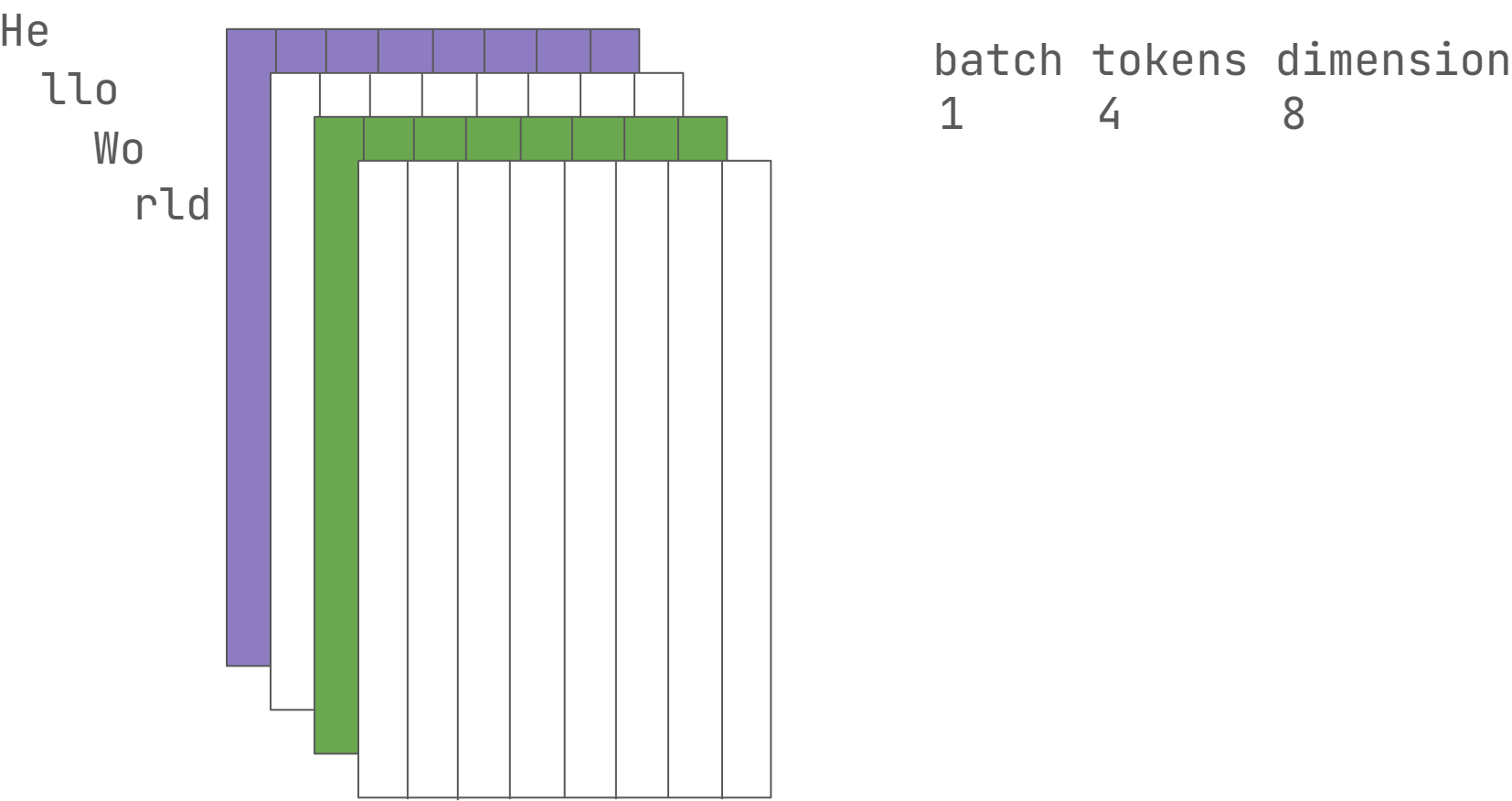


$$h_0 = UW_e + W_p$$

$$h_l = \text{transformer_block}(h_{l-1}) \forall i \in [1, n]$$

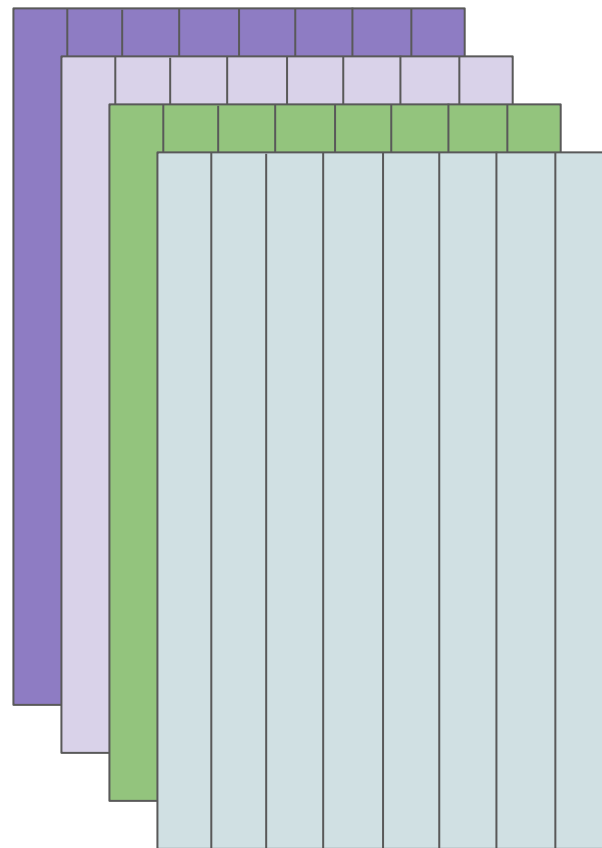
$$P(u) = \text{softmax}(h_n W_e^T)$$

Initial Residual Streams



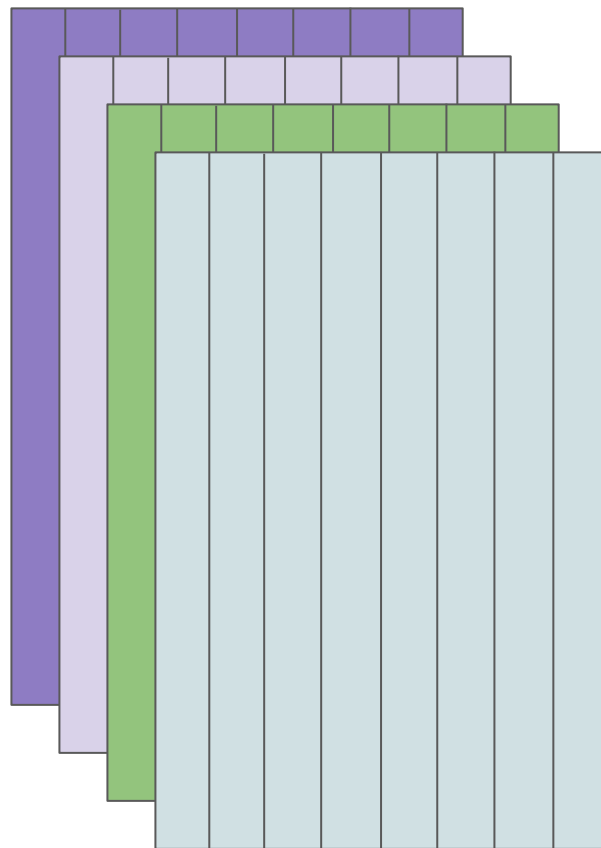
2 attention heads

	He	llo	Wo	rld
He	1	-	-	-
llo	.7	.3	-	-
Wo	.2	.2	.6	-
rld	.2	.2	.2	.4

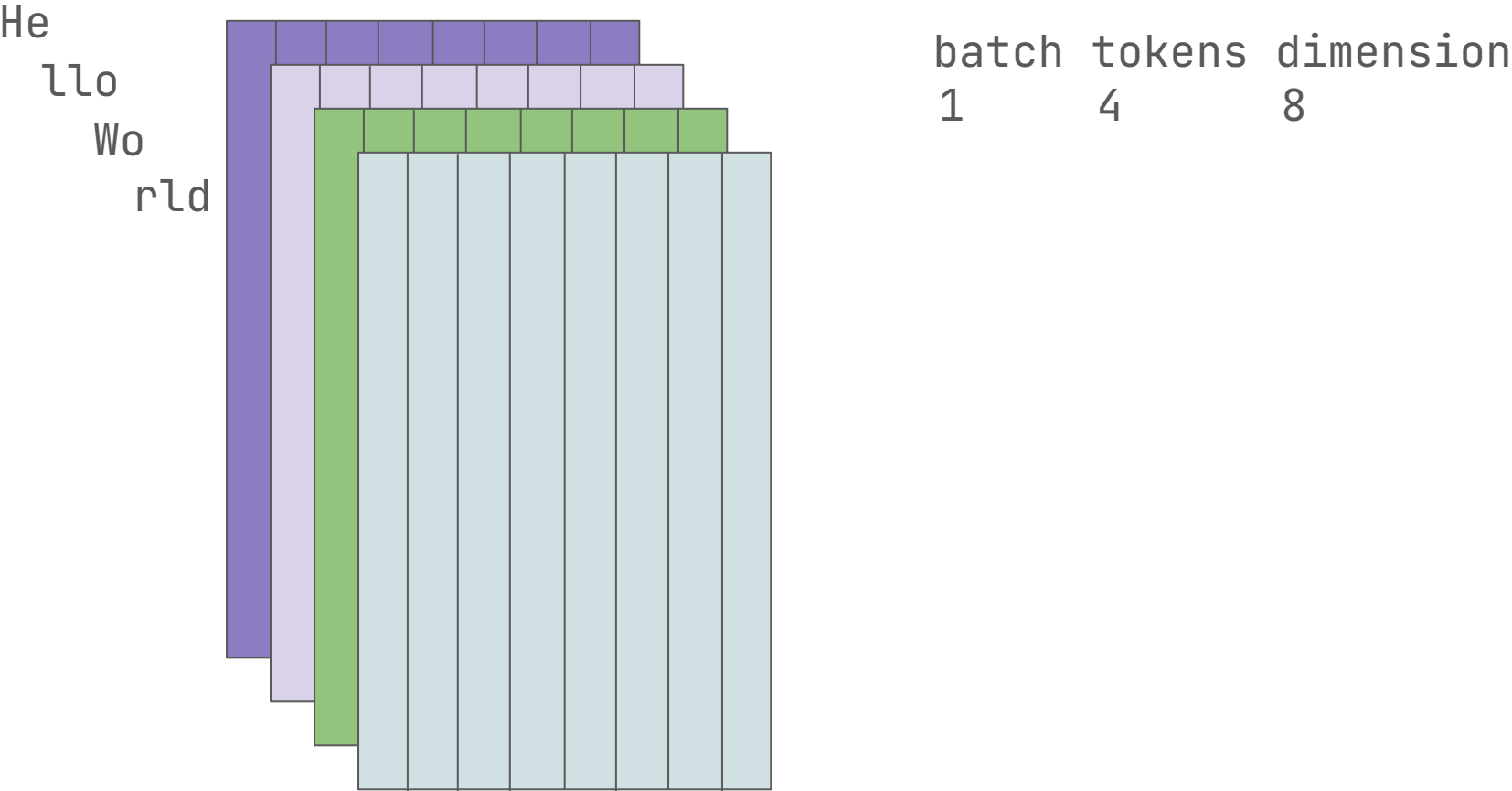


2 attention heads

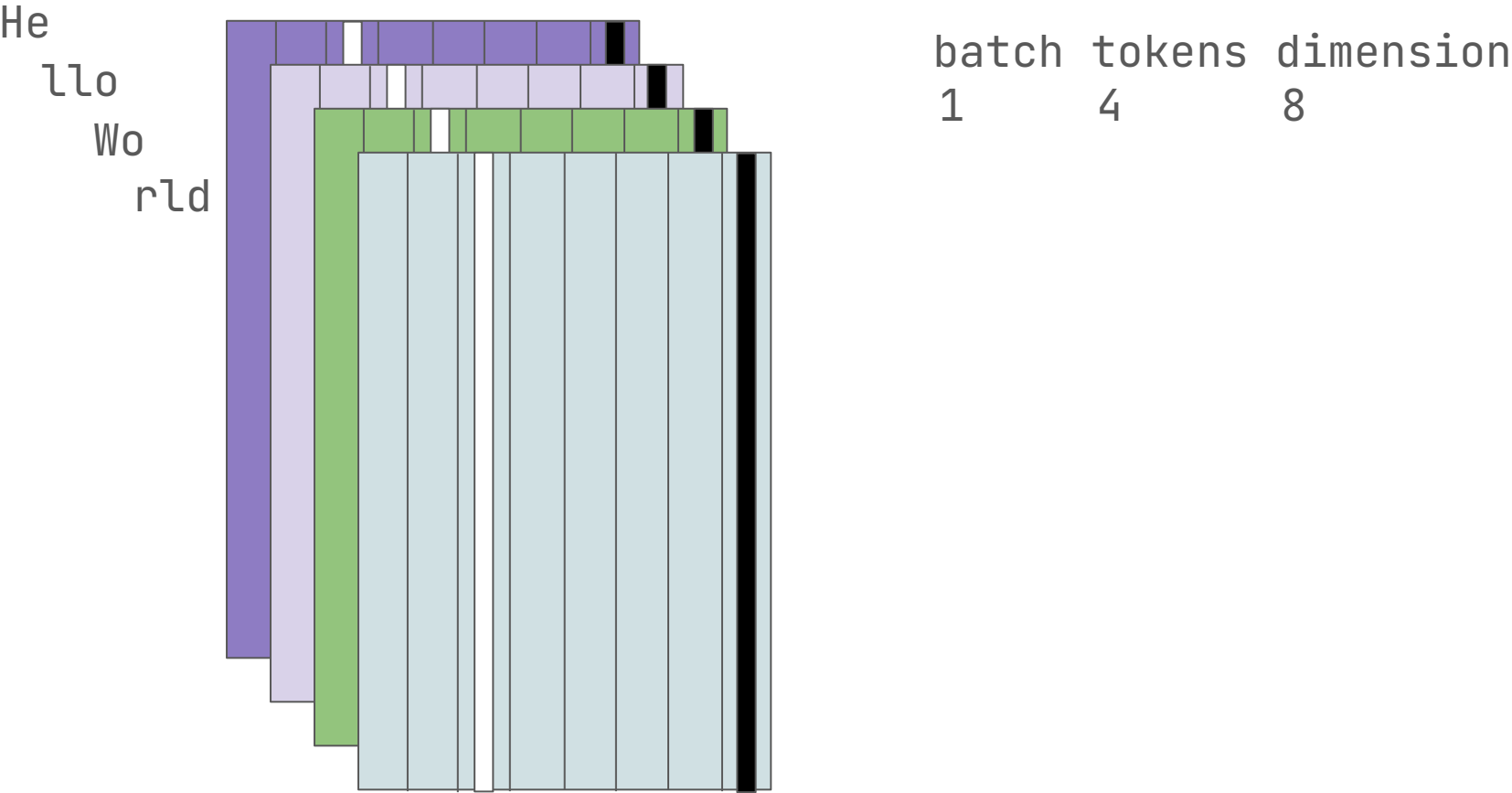
	He	llo	Wo	rld
He	1	-	-	-
llo	.7	.3	-	-
Wo	.2	.2	.6	-
rld	.2	.2	.2	.4



Residual Streams After Attention

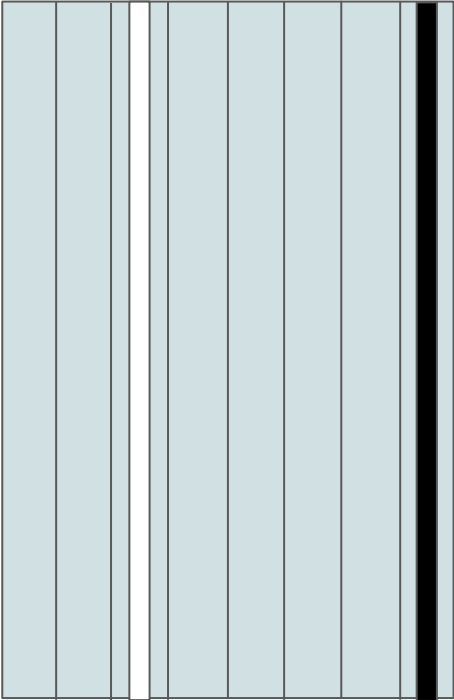


Residual Streams After Attention and MLP



Unembed

rld



.	60%
!	20%
,	8%
...	...

Resources / My path to this understanding

- **A Mathematical Framework for Transformer Circuits:** this paper has a section “Transformer Overview” that has influenced how I think about transformers more than anything else.
- **The papers mentioned in previous slides:** I found working through the GPT-1 paper really valuable. The “*Attention is All You Need*” paper is seminal and many people have created resources centering around it.
- **YouTube:** Serrano.Academy, Neel Nanda, Andrej Karpathy
- **Podcasts:** *Machine Learning Street Talk*, *The Cognitive Revolution*, *Latent Space*, + anything I can find with researchers from the frontier labs
- **Hands on:** I learned a lot from fine-tuning GPT-1
- **Claude and ChatGPT**