Università degli Studi di Padova

The Dimensionality of Transformers

Unveiling GPT-2: Reconstructing the Geometry of the Embedding Space

Ginevra Beltrame Emanuele Coradin Ada D'Iorio Dario Liotta

July 17, 2024

Table of Contents

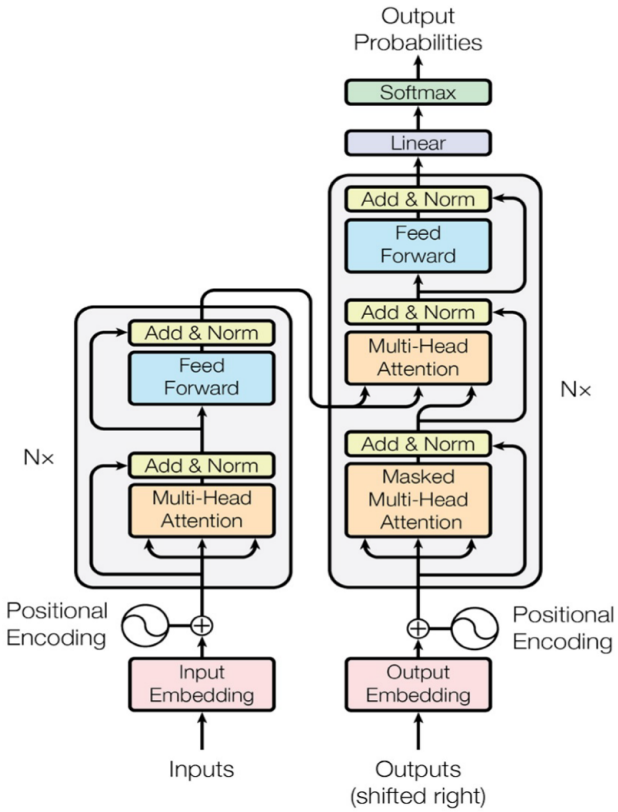
Table of Contents

Transformer internal structure

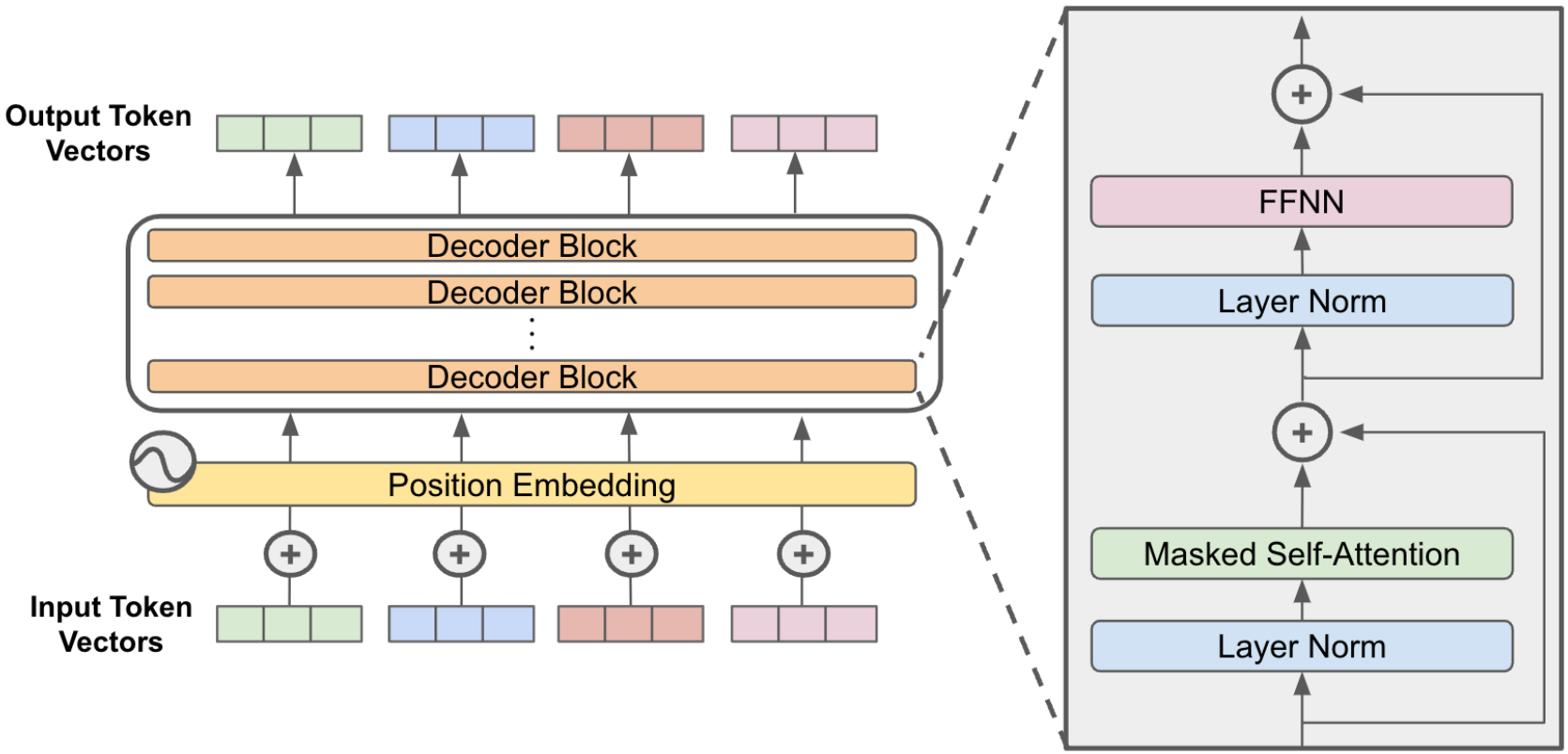
1. Positional embedding
   converting words to tokens and tokens to vectors

For each decoder

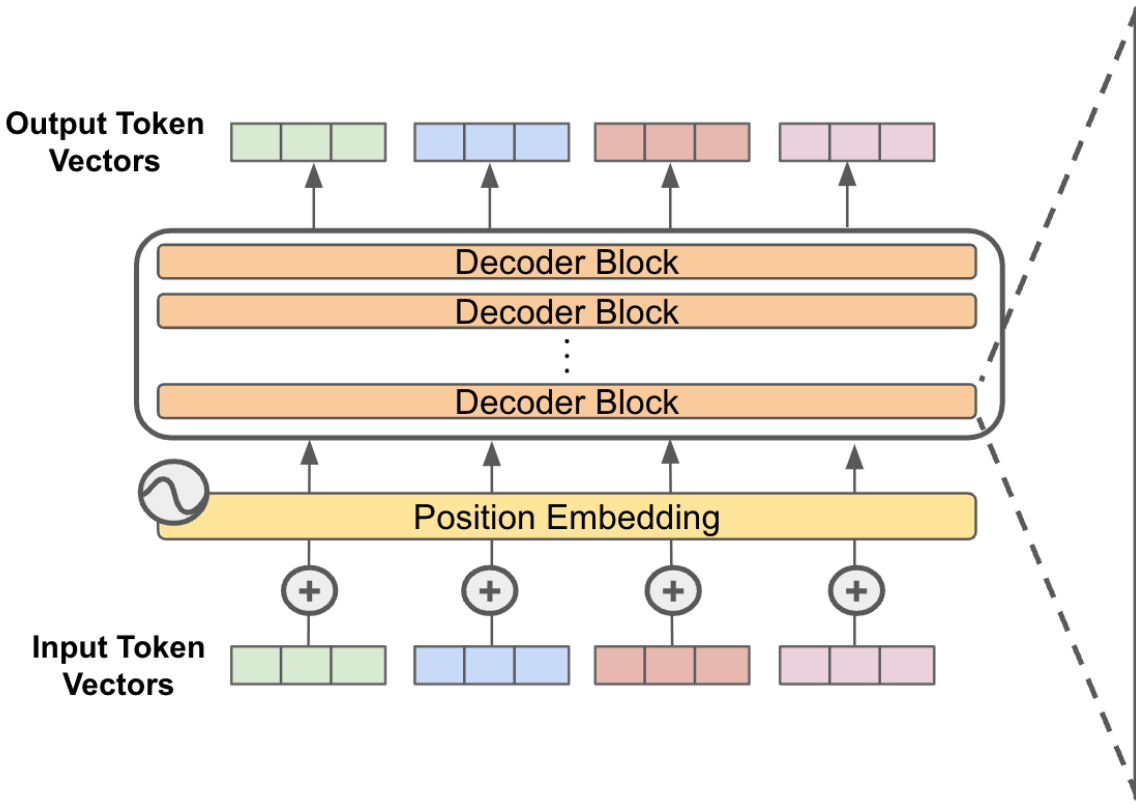
2. Attention Mechanism
   with twelve Attention Heads to capture the context
3. Addition & normalization
   to connect and stabilize each layer
4. Feed Forward Neural Network
   with one hidden layer

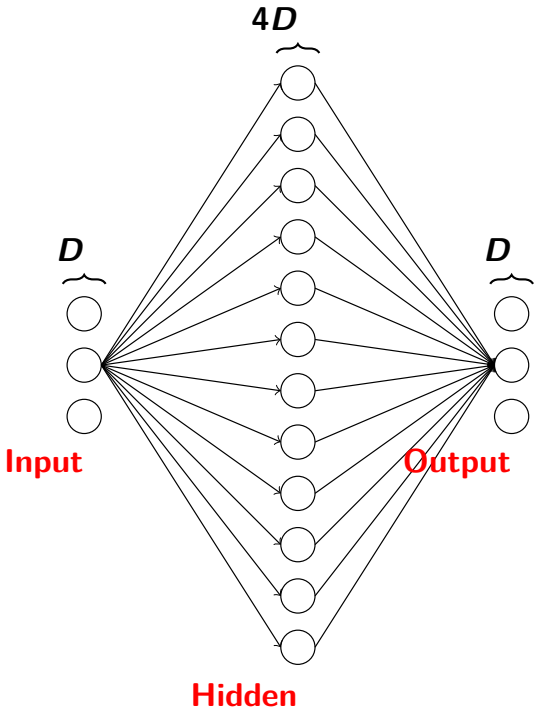


Transformer internal structure



Transformer internal structure





The attention mechanism

- Simulates human attention by assigning varying **degrees of correlation** among words in a sentence

- The dependencies can be related to semantic meaning and **context**
- Enhances the transformer's ability to capture **long-range dependencies**

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

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

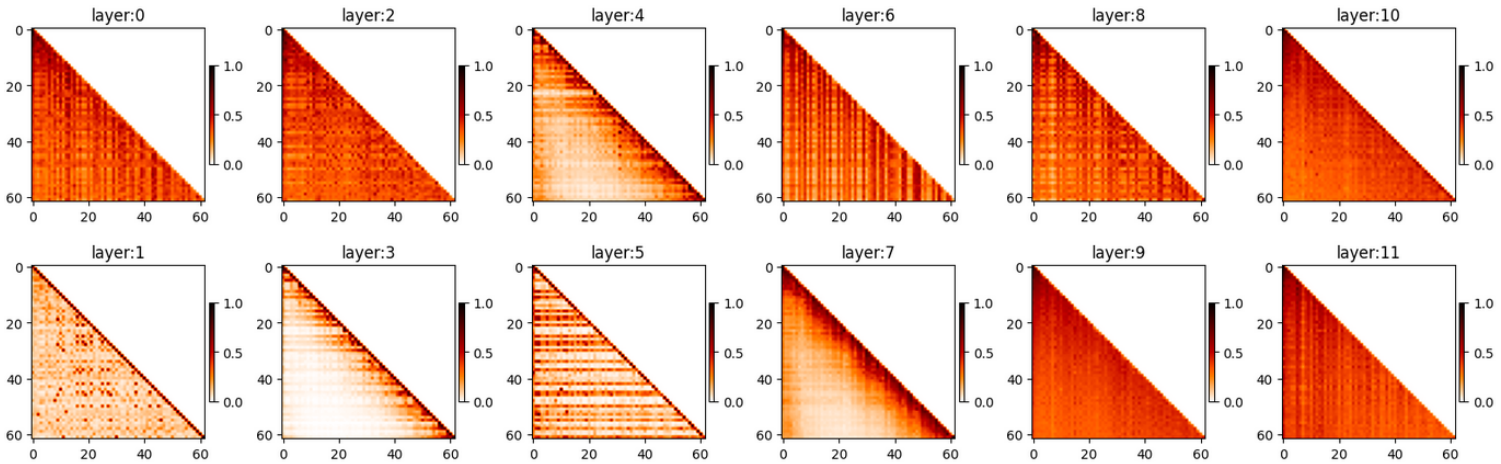
†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

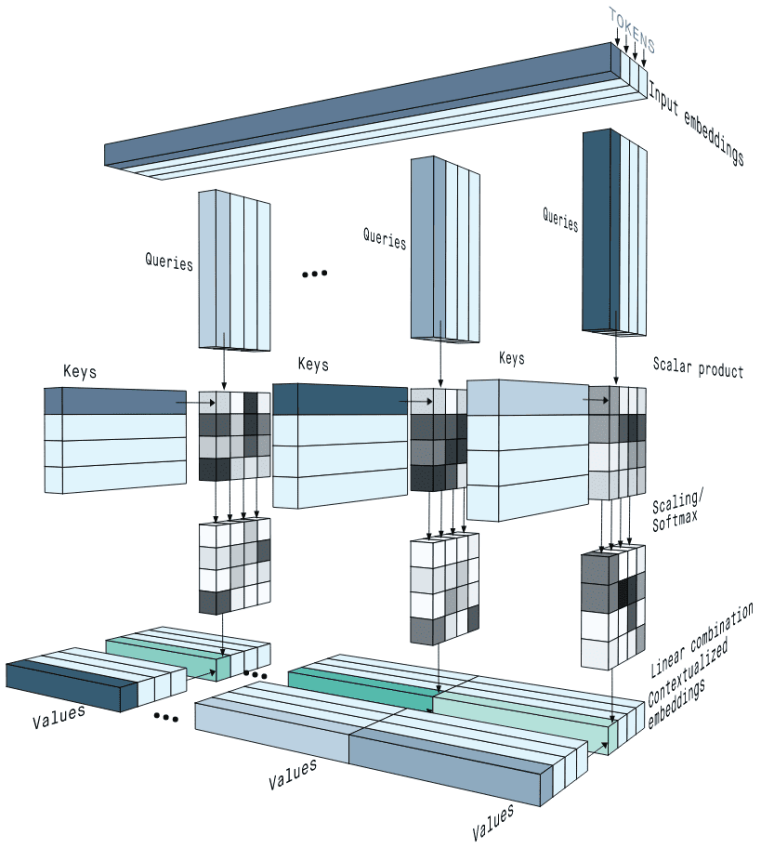
The attention mechanism

- The weights for the vectors in the embedding space are calculated by the **twelve attention heads**

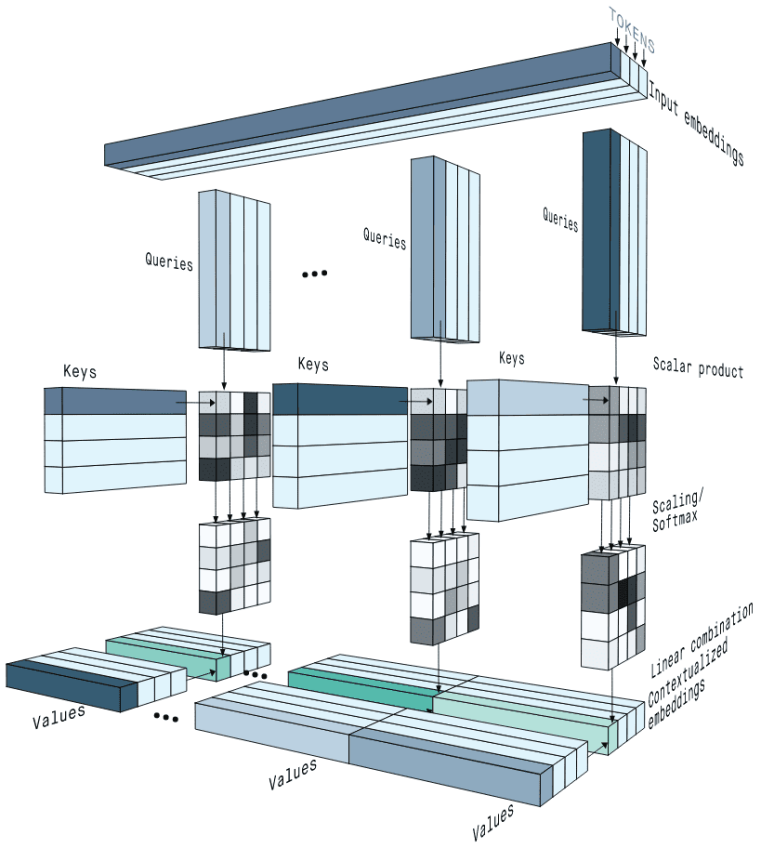


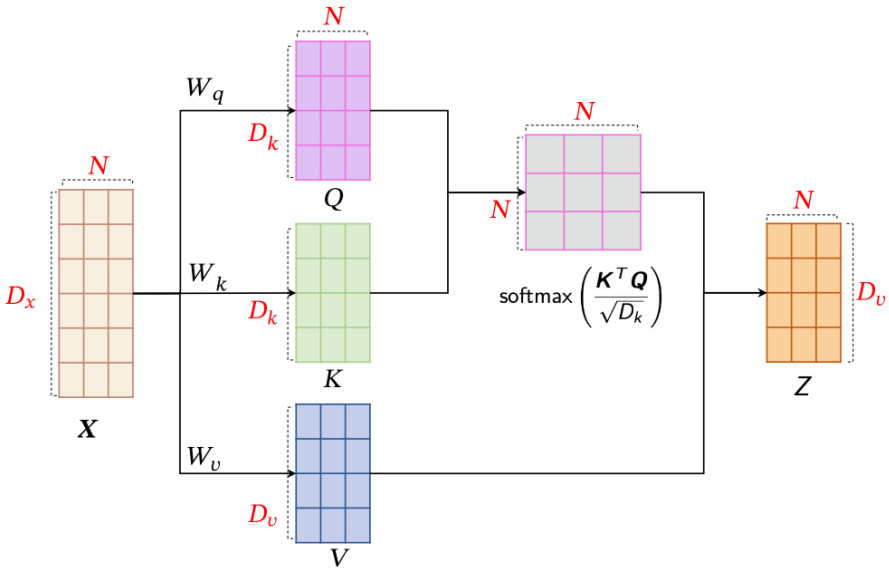
- Contextual **proximity matrices** are built and dynamically adjusted

The attention mechanism

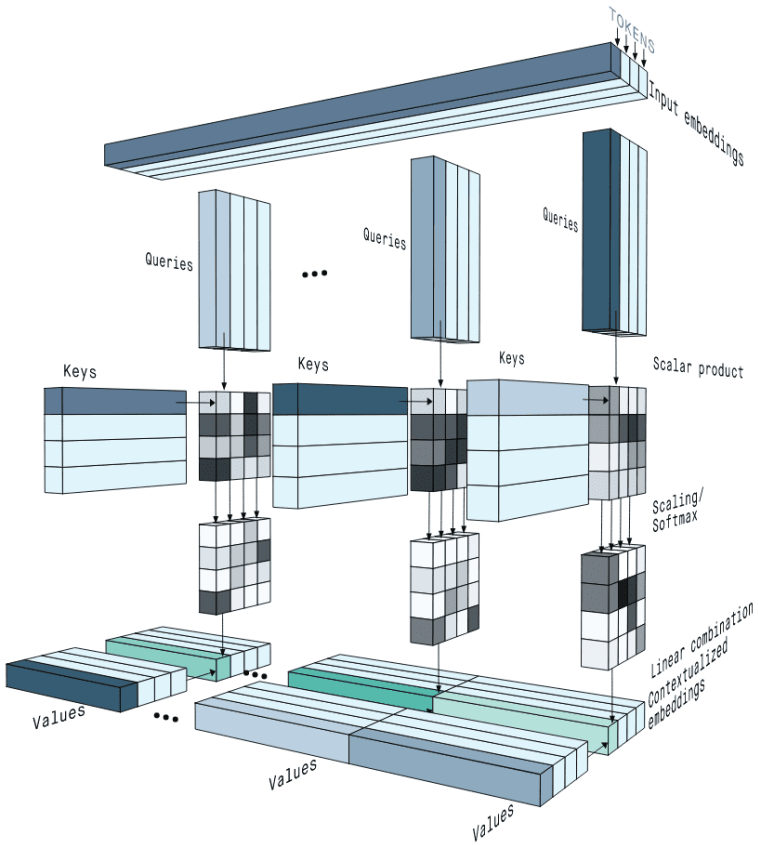


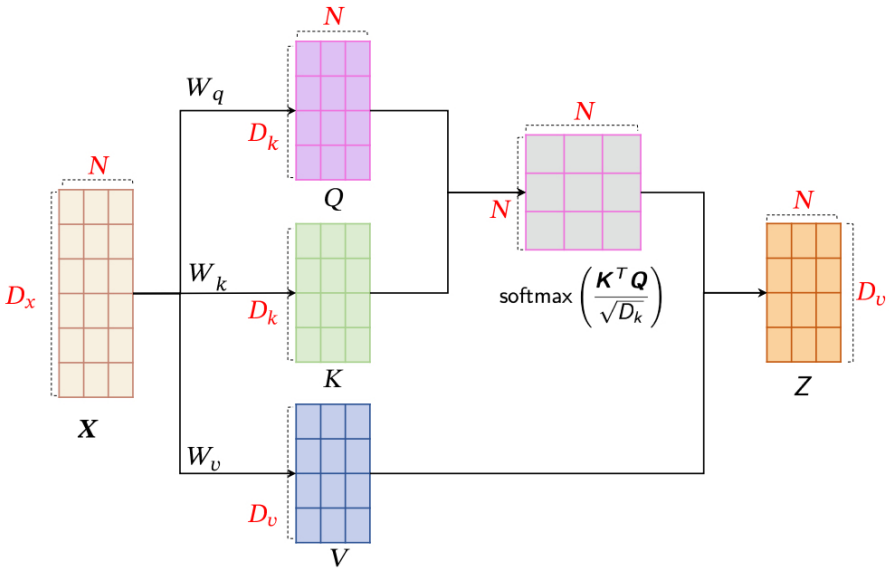
The attention mechanism

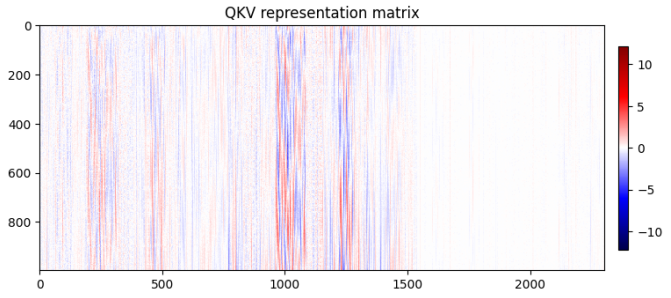




The attention mechanism





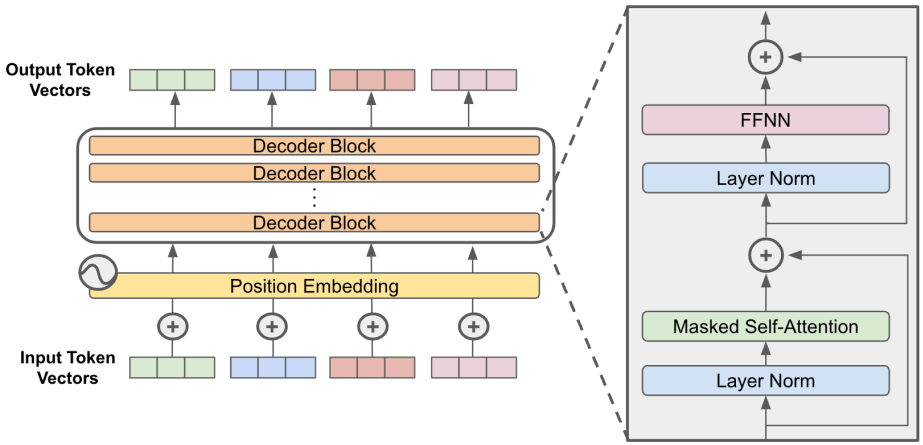


From one of our prompts

Table of Contents

Aims of the project

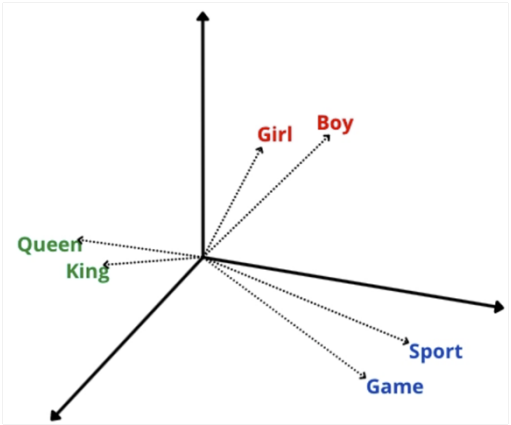
- Opening the **transformer black box**
  What exactly happens in there?
  How is the decoder structure layered?
  How is the prompt processed?

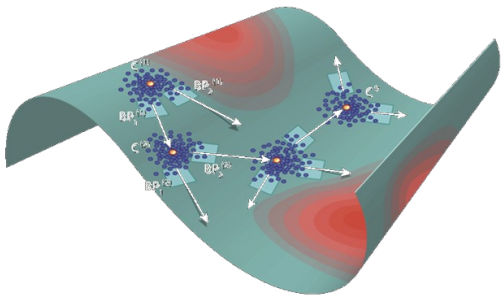


Aims of the project

- Opening the **transformer black box**
  What exactly happens in there?
  How is the decoder structure layered?
  How is the prompt processed?

- Inspecting the **embedding space**
  How does the geometry of the embedding space evolve?
  How does its dimensionality evolve?





Aims of the project

- Opening the **transformer black box**
  What exactly happens in there?
  How is the decoder structure layered?
  How is the prompt processed?

- Inspecting the **embedding space**
  How does the geometry of the embedding space evolve?
  How does its dimensionality evolve?

- Studying the mechanism behind the **last token** prediction
  How does the model produce each next-token output?
  How do the heads express the last token's dependencies on previous ones?

Table of Contents

Technical approach

The generative model we used is **GPT-2 small**

Technical approach

The generative model we used is **GPT-2 small**

117 million parameters

Technical approach

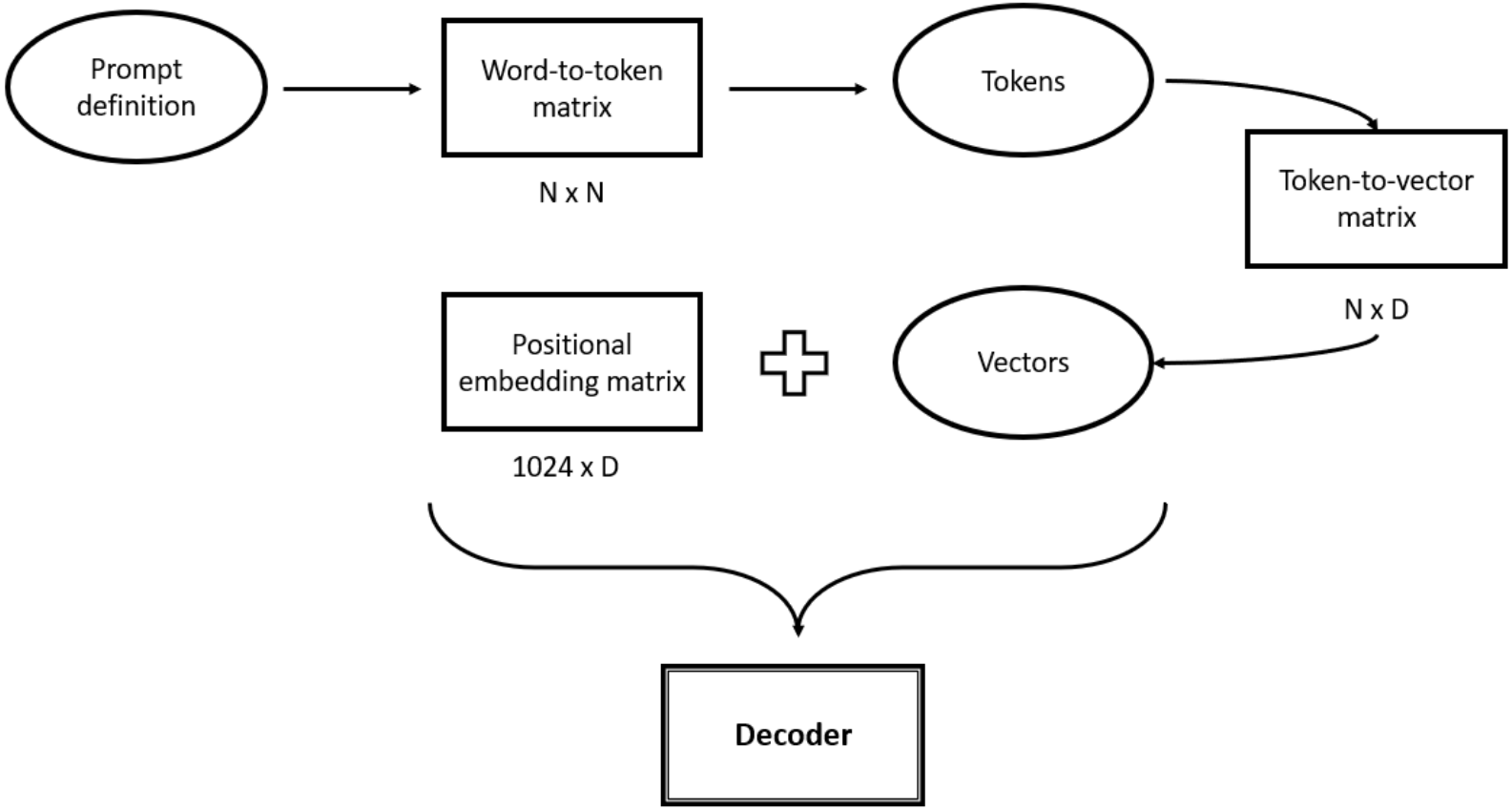
The generative model we used is **GPT-2 small**

117 million parameters

Our starting point: use of prompts with different **lenghts** and **semantic areas**

1. **Pre-processing**
   Encoding information inside the model
2. Study of the **dimensionality** of the embedding space
   Observing how the intrinsic dimension changes over the process
   Computing the relative distances between elements in the space, determining an appropriate metric
3. **Last-token analysis**
   The prediction of the last token depends only on the result of the last token in the prompt

Pre-processing

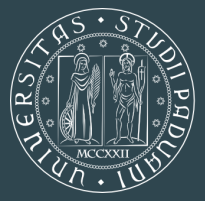


Dimensional analysis

Approaches:

- Basic clustering with **t-SNE**
- **Dadapy**
  - Dimensional analysis
  - Metrics comparison
- **PCA** with SVD
- Manifold volume with **Gram matrix**

DADApy - Gride algorithm

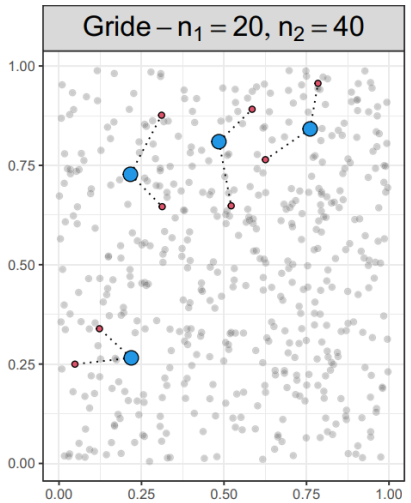


- Generalized Ratios Intrinsic Dimension Estimator: statistical method introduced in the DADApy package
- Allows **estimating the intrinsic dimension (ID) of a manifold** as an explicit function of the scale of the distances among data points

DADApy - Gride algorithm

- Generalized Ratios Intrinsic Dimension Estimator: statistical method introduced in the DADApy package
- Allows **estimating the intrinsic dimension (ID) of a manifold** as an explicit function of the scale of the distances among data points
- Method:

1. choose two nearest neighbors: the n_1-th and the n_2-th (typically $n_1 = 2n_2$)
2. exploit the maximum log-likelihood properties to estimate the ID based on n_1, n_2 and the radii of the correspondent hyperspheres of data points



Reference: https://doi.org/10.1038/s41598-022-20991-1

Prompts selection

Various **prompt sizes**

1. 50 words
2. 300 words
3. 950 words

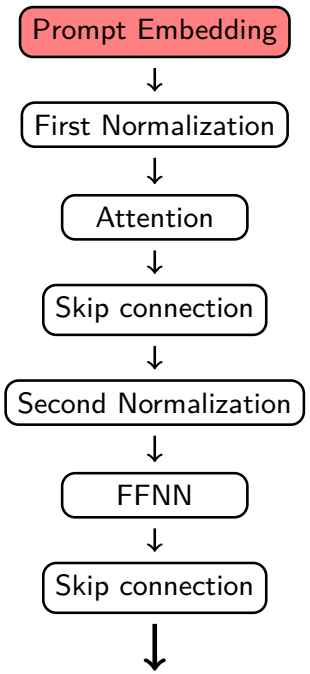
Different **semantic area** and **text structure**

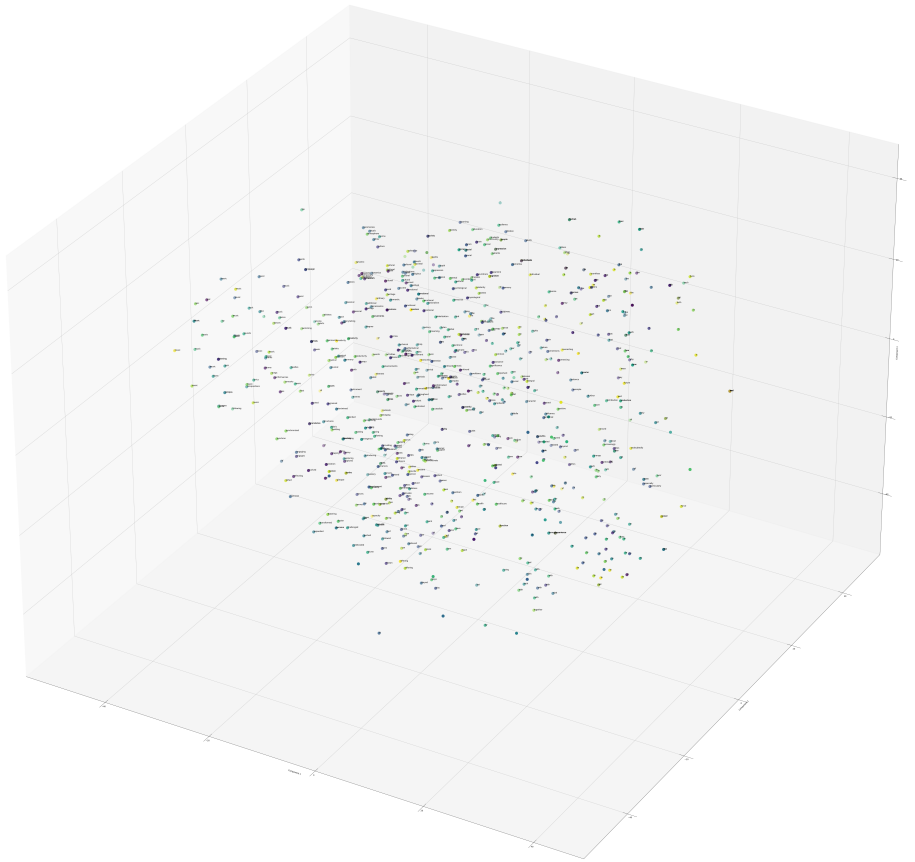
1. List
2. Poetry
3. Instructions
4. Plain text

	List	Poetry	Instructions	Text 950	Text 300	Text 50
Nature	trees forest river mountains wildlife ecosyste...	In nature's realm, where wonders bloom, A tape...	Taking care of biodiversity in a national park...	Nature, in its vast expanse and intricate deta...	Nature, with its breathtaking beauty and intri...	Nature encompasses Earth's landscapes, ecosyst...
Astronomy and Phyiscs	astronomy physics astrophysics cosmology unive...	In the velvet expanse where stars ignite, Astr...	Studying the phenomenon of dark energy and gal...	Astronomy and physics are two intertwined fiel...	Astronomy and physics stand as pillars of scie...	Astronomy explores the vastness of the univers...
Mathematics	addition subtraction multiplication division a...	In the realm where numbers weave their dance, ...	Solving differential equations involves a syst...	Mathematics is the language of patterns, struc...	Mathematics is the universal language of patte...	Mathematics is the language of patterns, shape...
Psychology	psychology cognition perception memory learnin...	In minds' depths, thoughts weave and sway, Emo...	Providing psychotherapy involves a structured ...	Psychology, as a field of study, delves into t...	Psychology is the scientific study of the mind...	Psychology delves into the complexities of the...
Music	melody harmony rhythm beat tempo dynamics pitc...	In the silence before the first note's birth, ...	Composing a musical piece is a creative endeav...	Music, a universal language, transcends cultur...	Music is a universal language that transcends ...	Music is a universal language that transcends ...

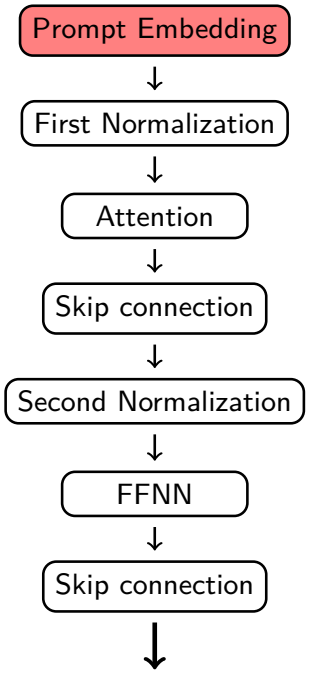
Table of Contents

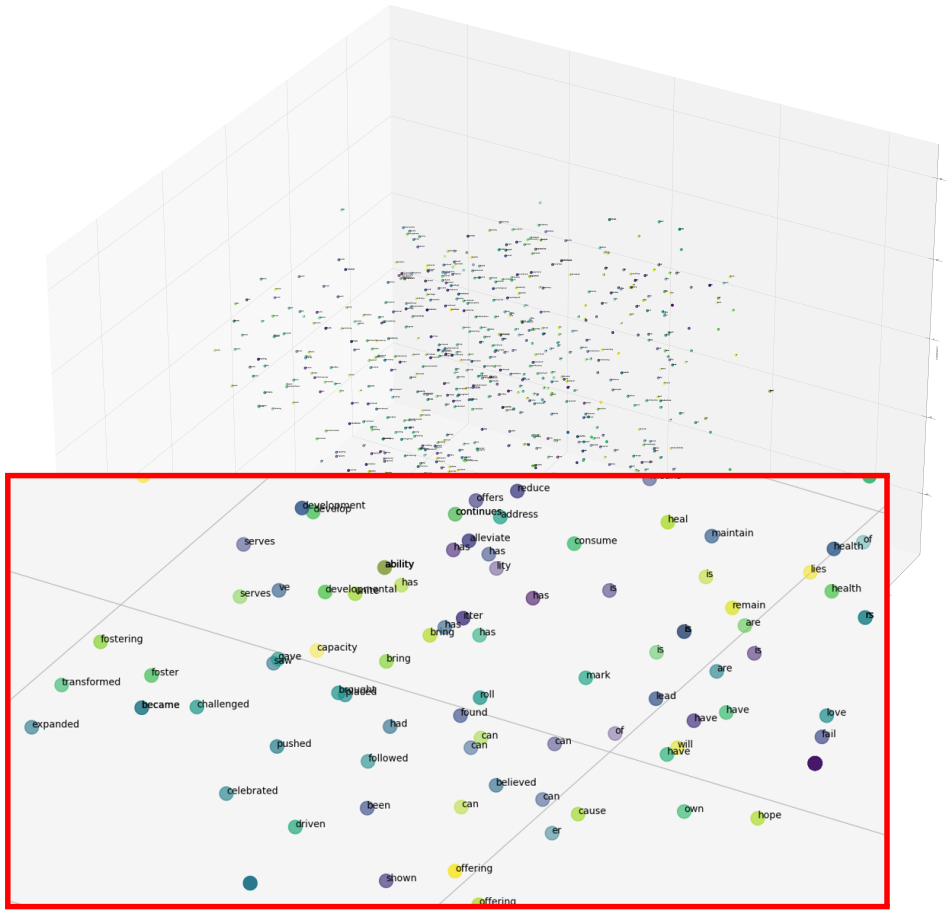
Semantic cluster visualization



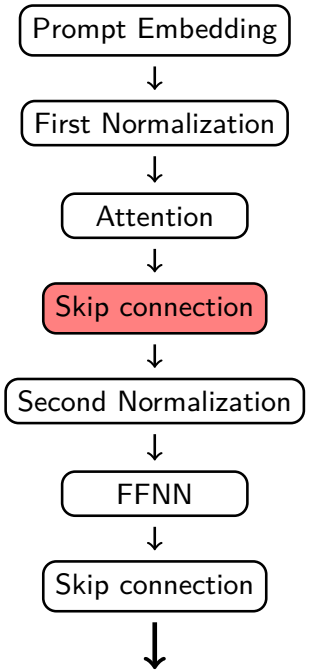


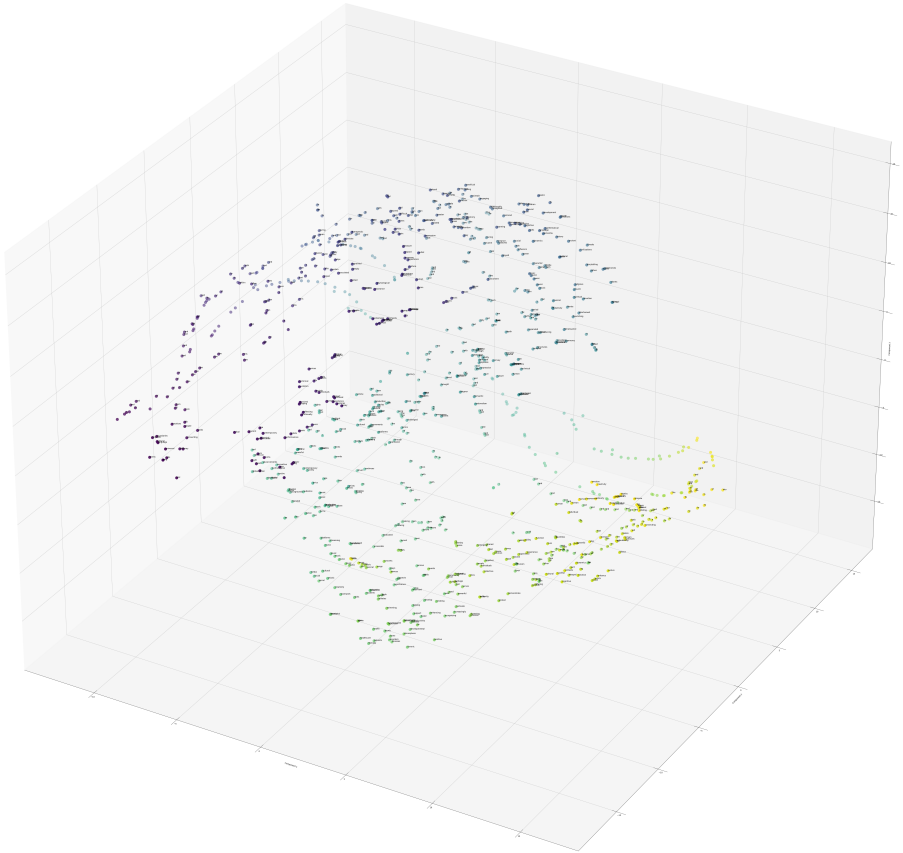
Semantic cluster visualization



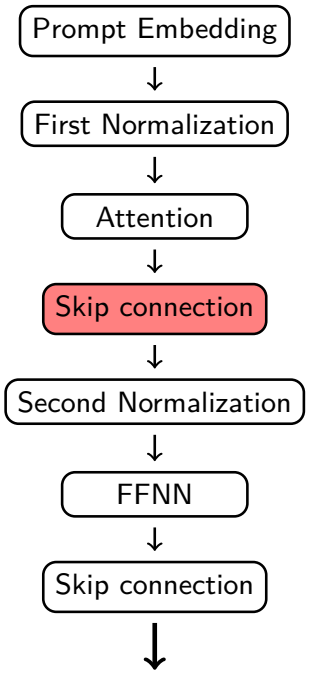


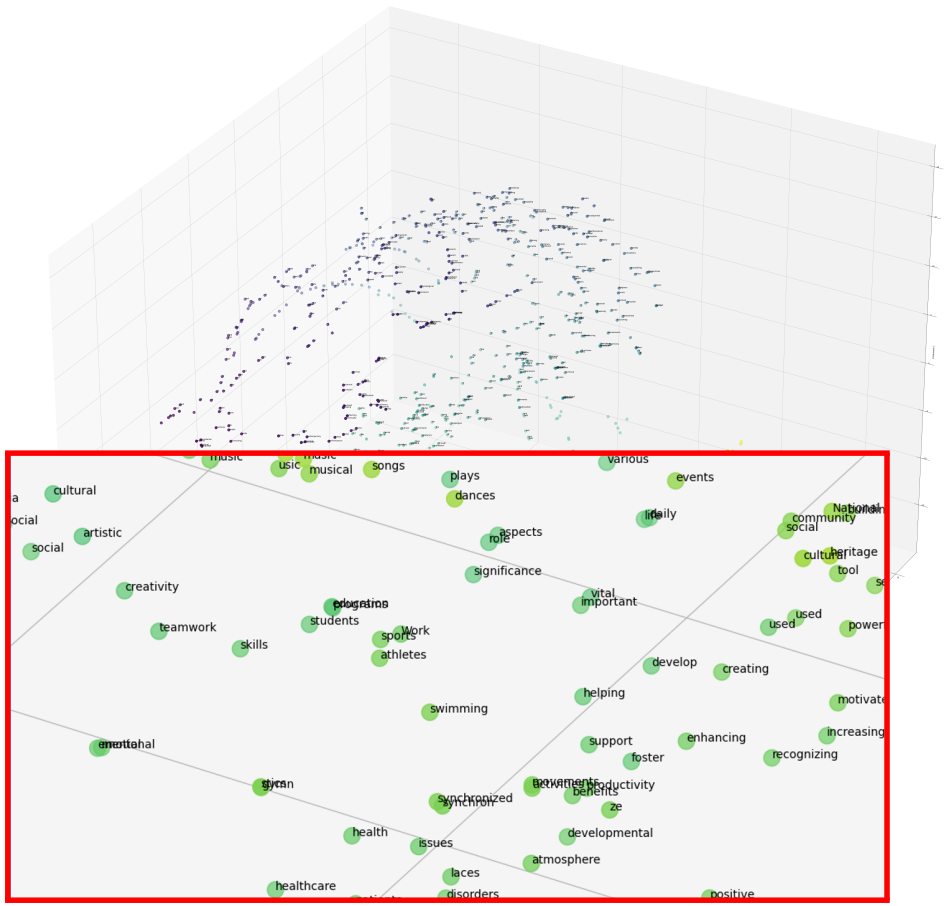
Semantic cluster visualization



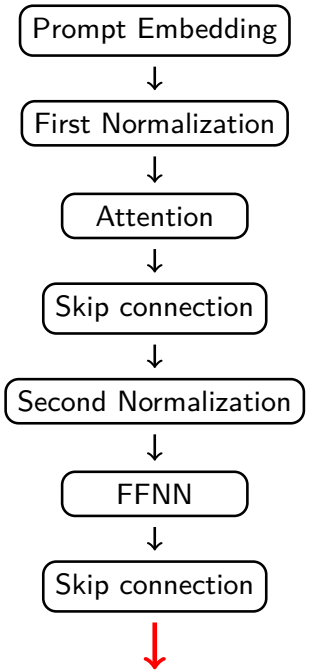


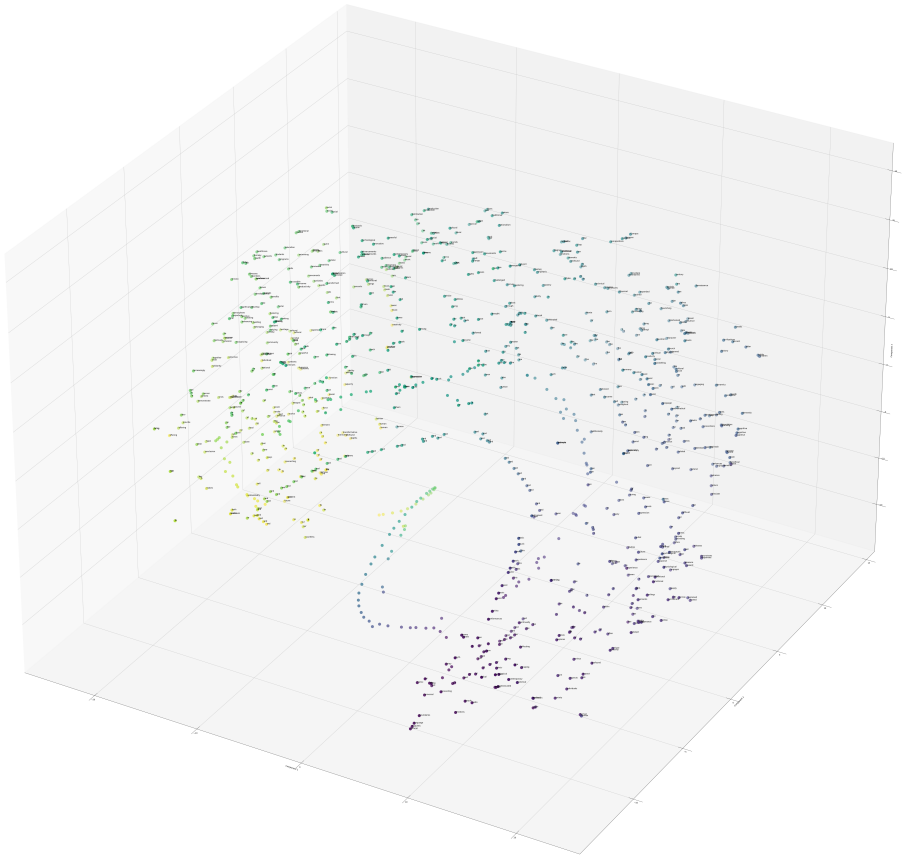
Semantic cluster visualization



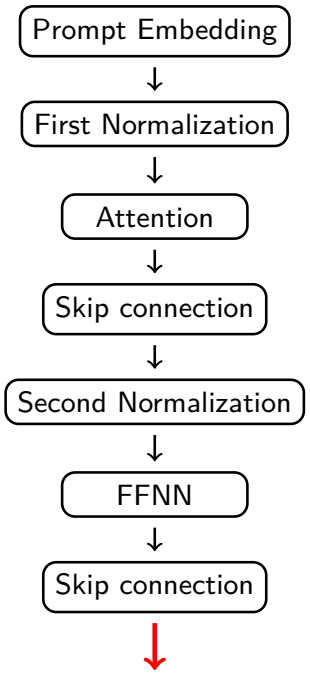


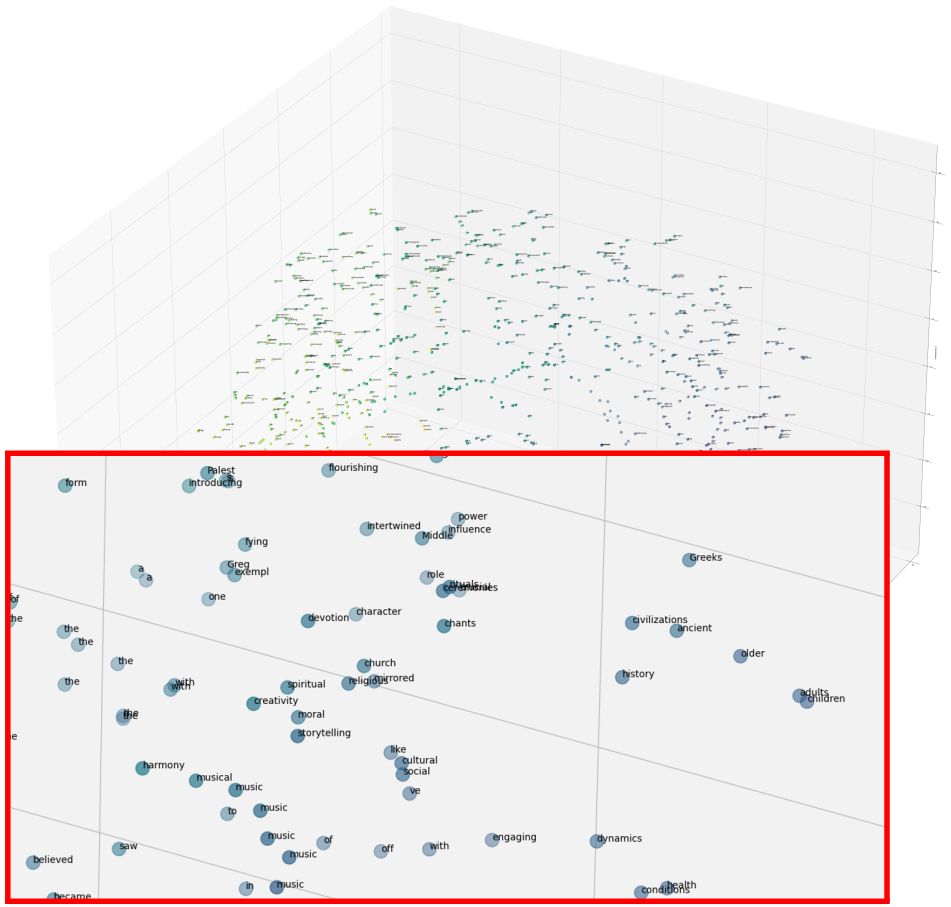
Semantic cluster visualization





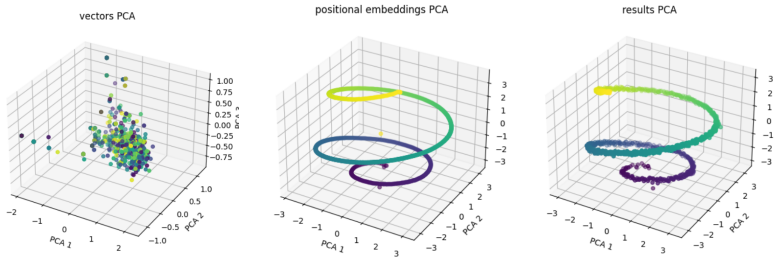
Semantic cluster visualization



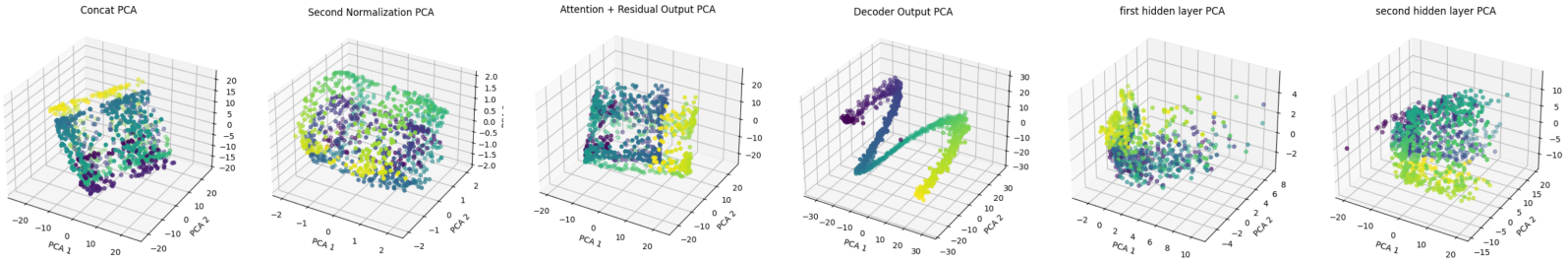


PCA analysis

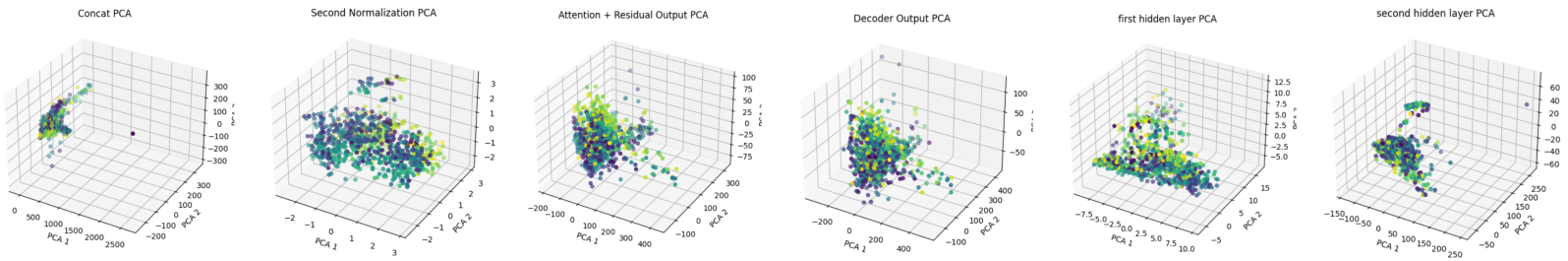
Positional Embedding:



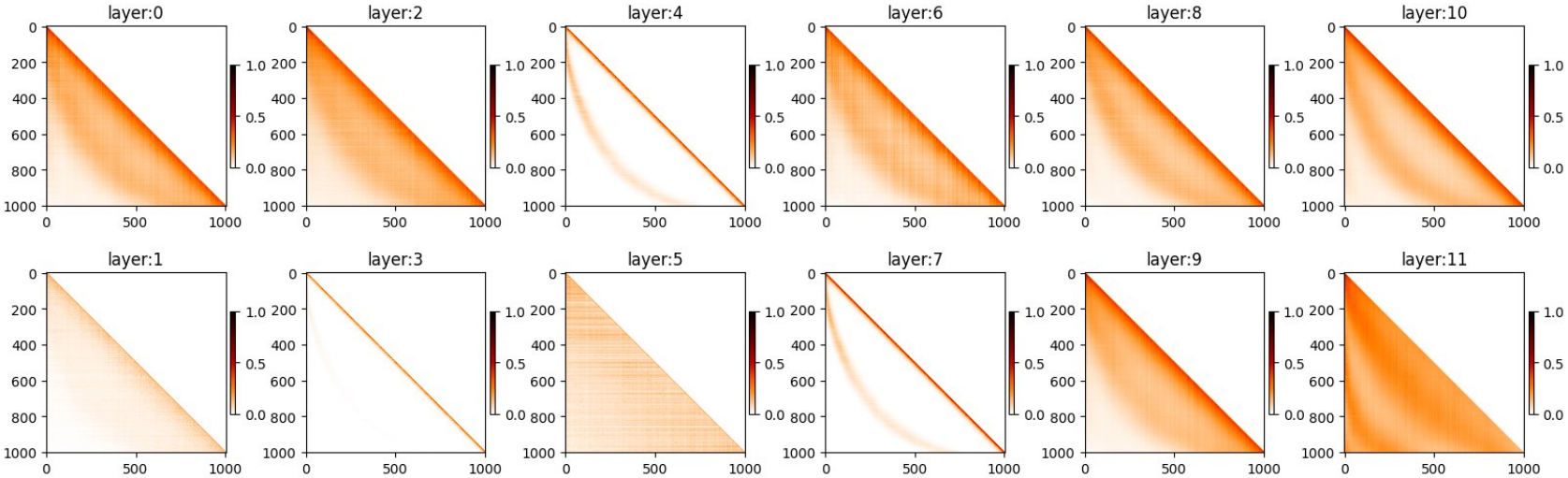
First Decoder:



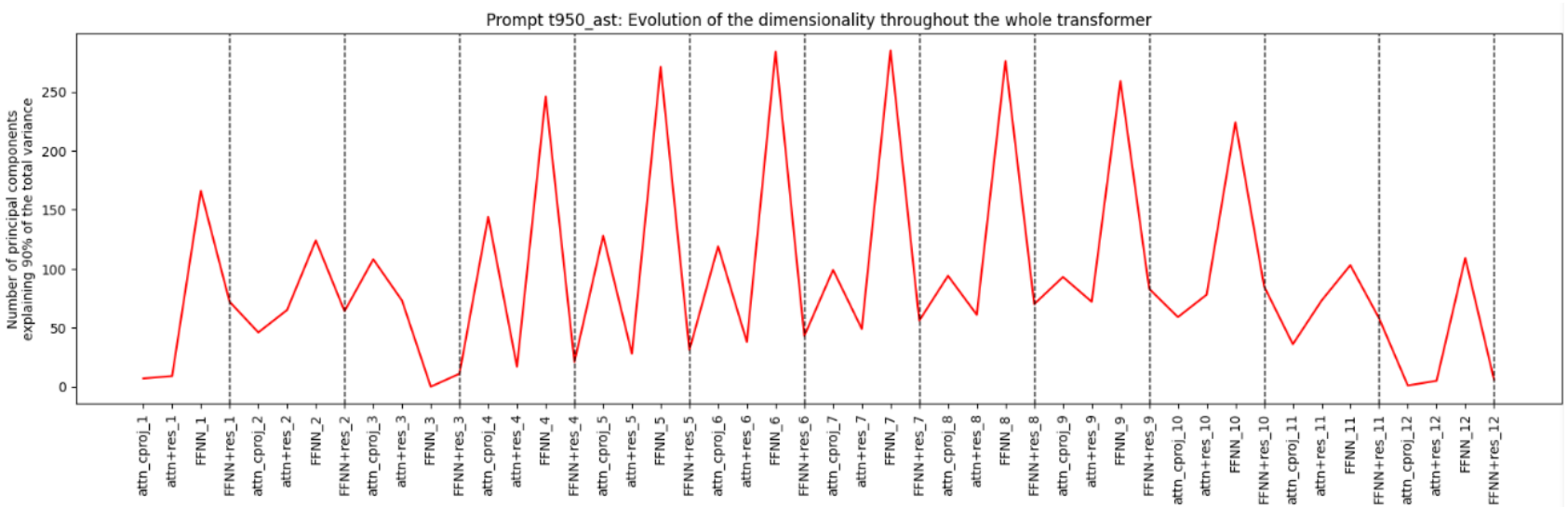
Last Decoder:



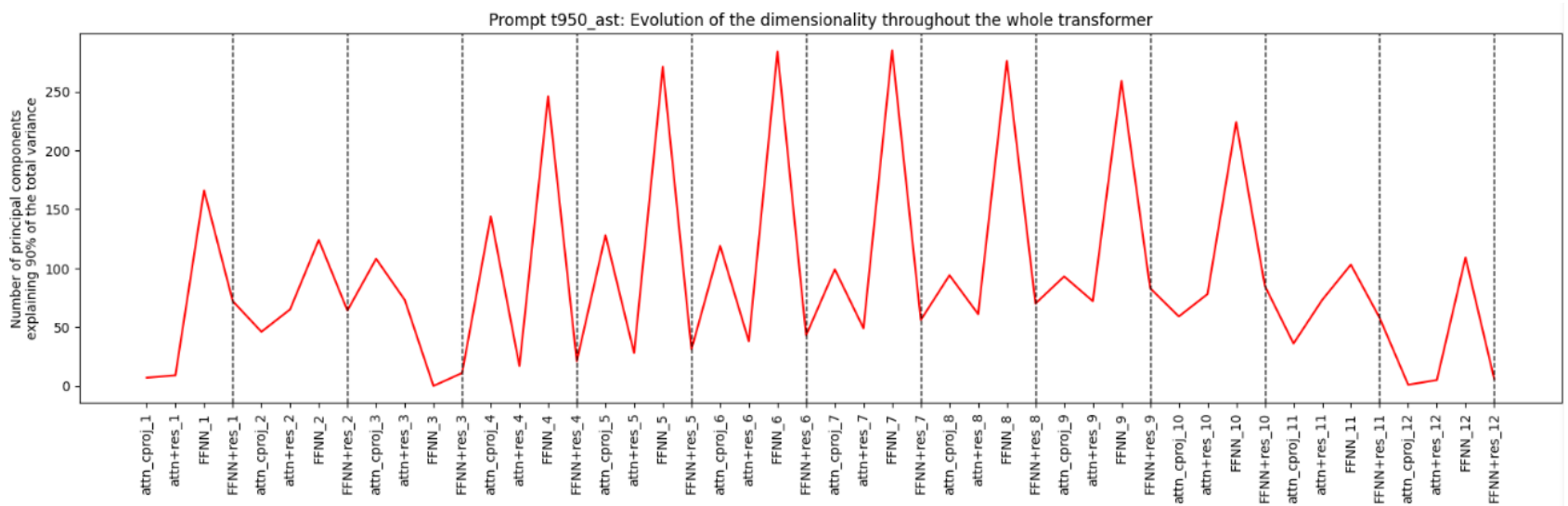
QKV matrices and attention heads

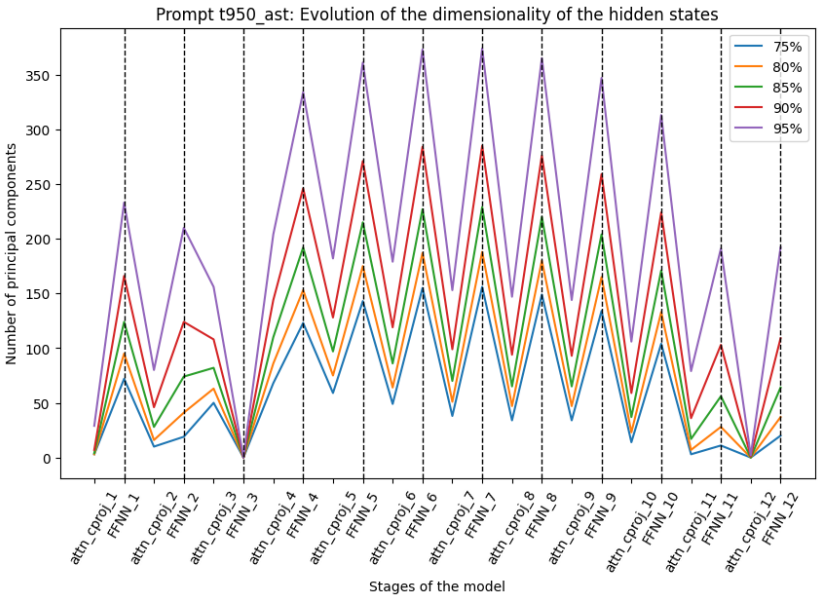


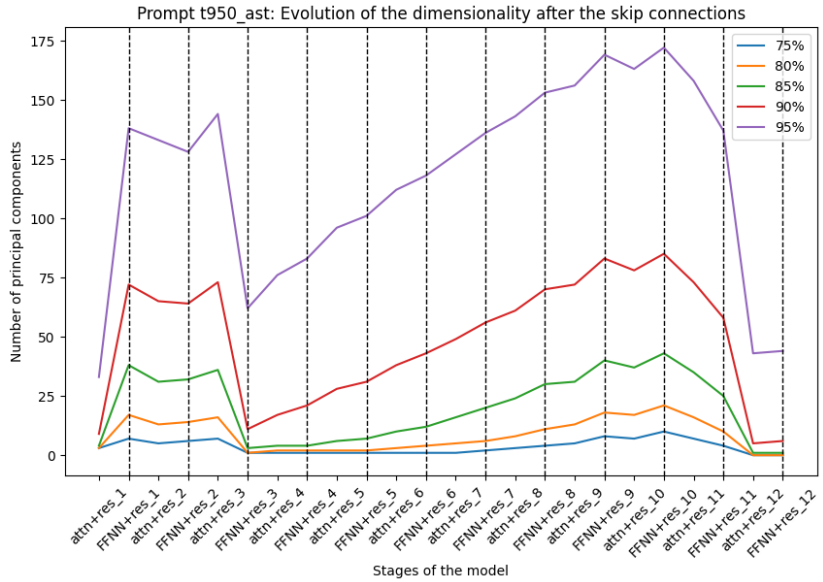
Dimensionality evolution



Dimensionality evolution



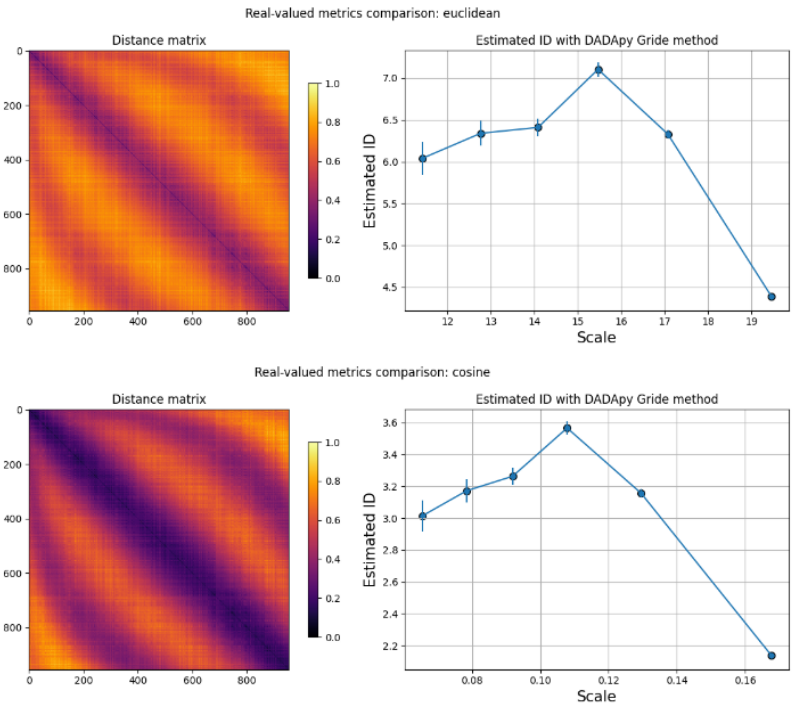


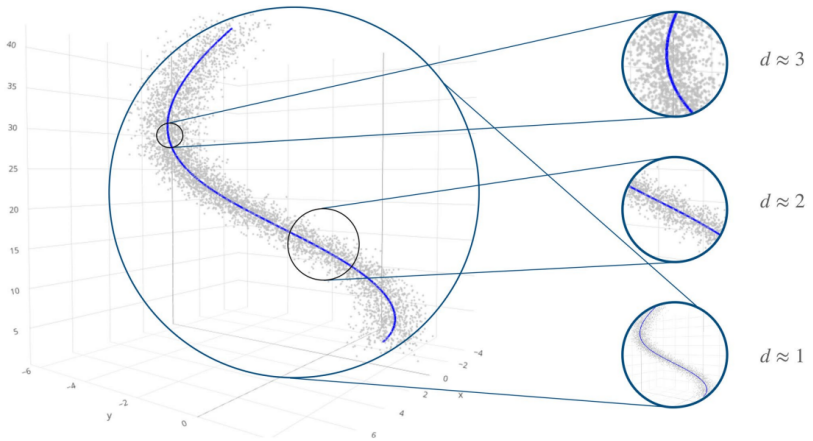


Intrinsic Dimension and Metrics

1000-token prompt

First decoder

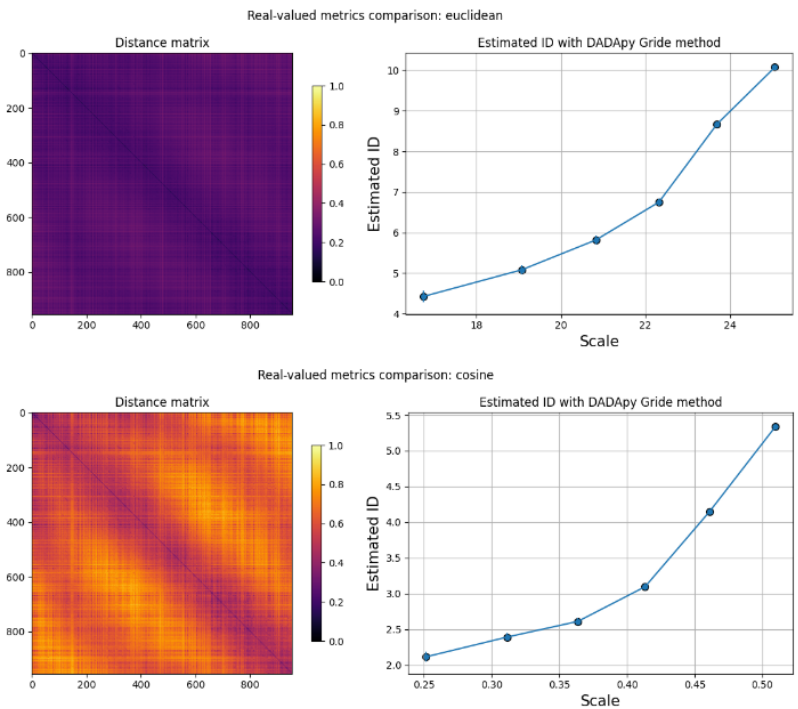




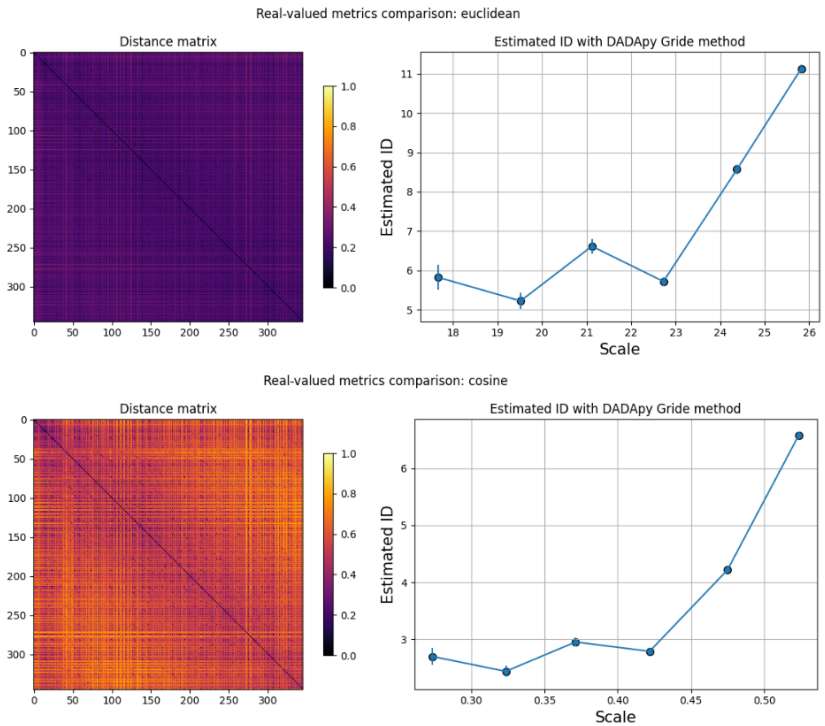
Intrinsic Dimension and Metrics

Università degli Studi di Padova

1000-token prompt

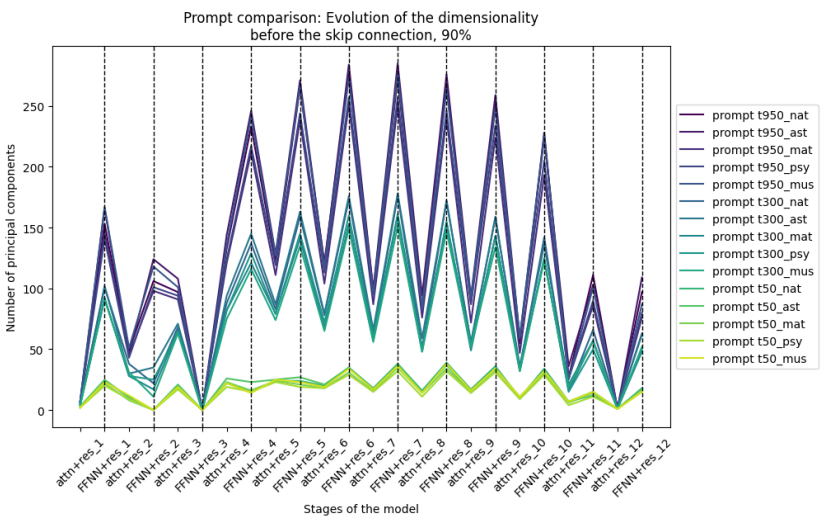


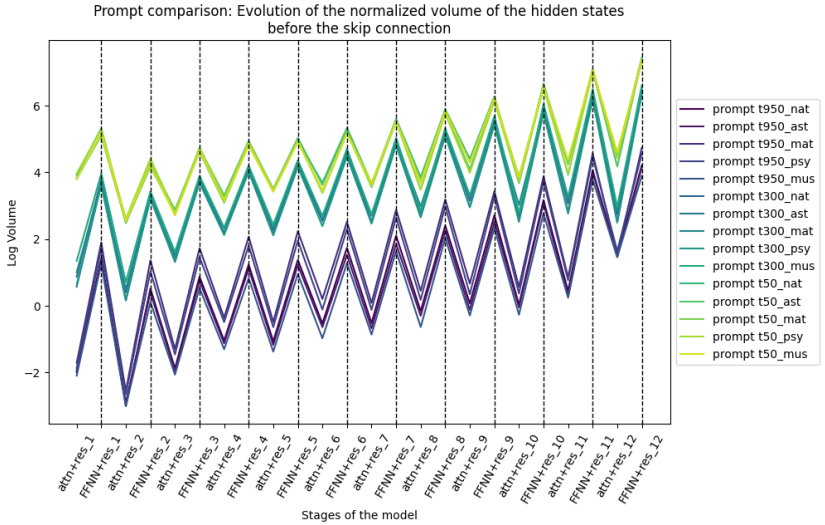
300-token prompt

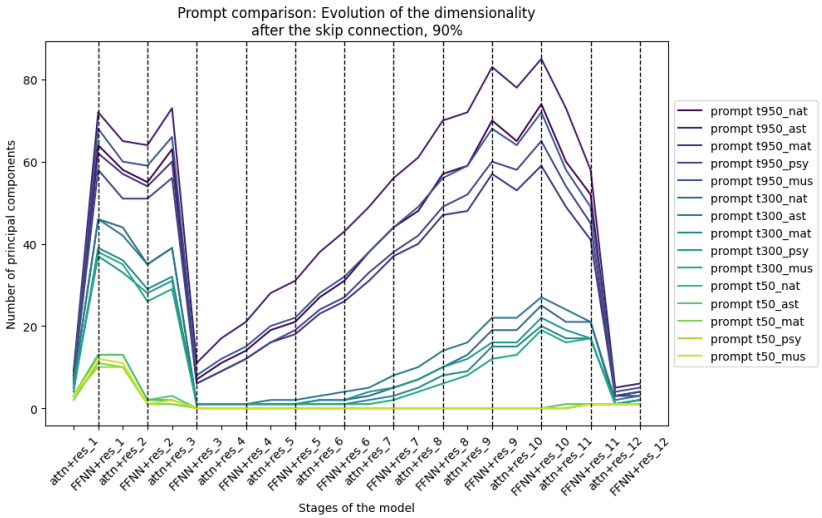


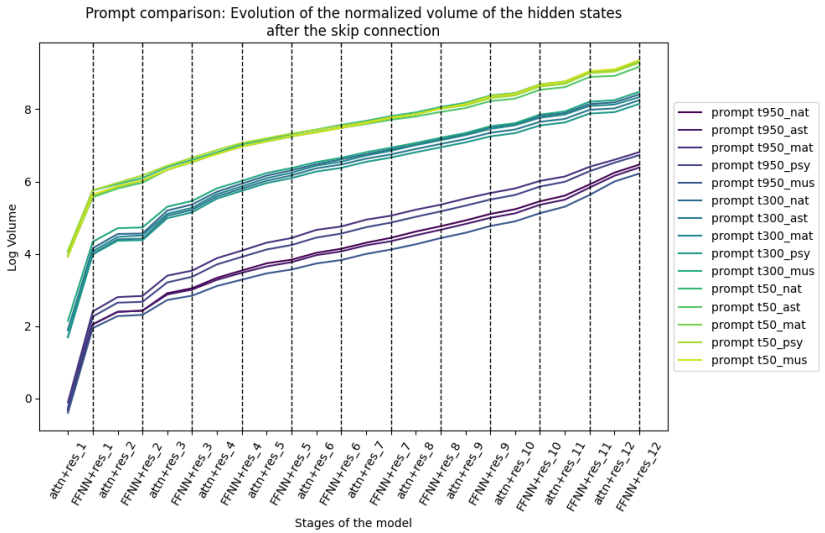
Statistical sample analysis

Prompt size comparison: [955, 980, 863, 889, 1003, 405, 434, 368, 409, 345, 63, 66, 58, 54, 62]



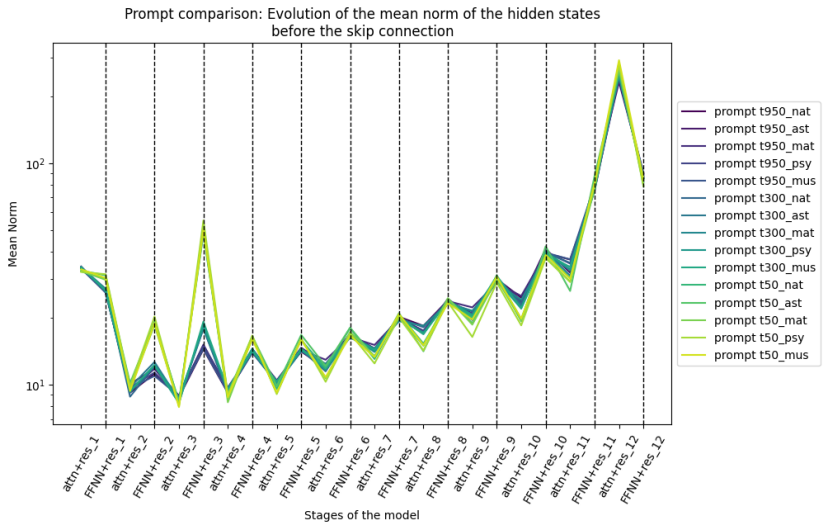


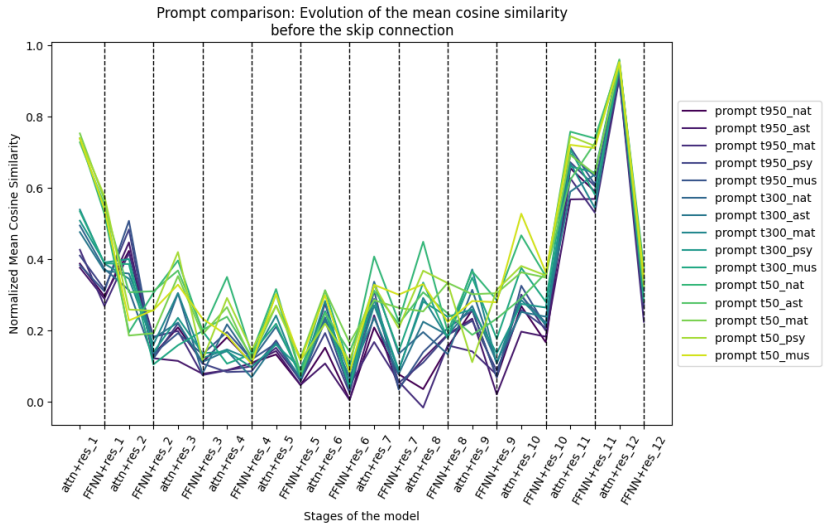


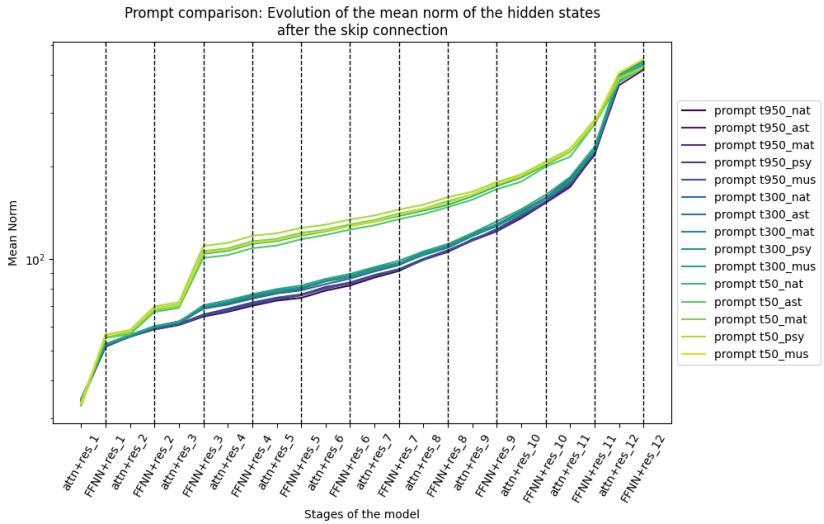


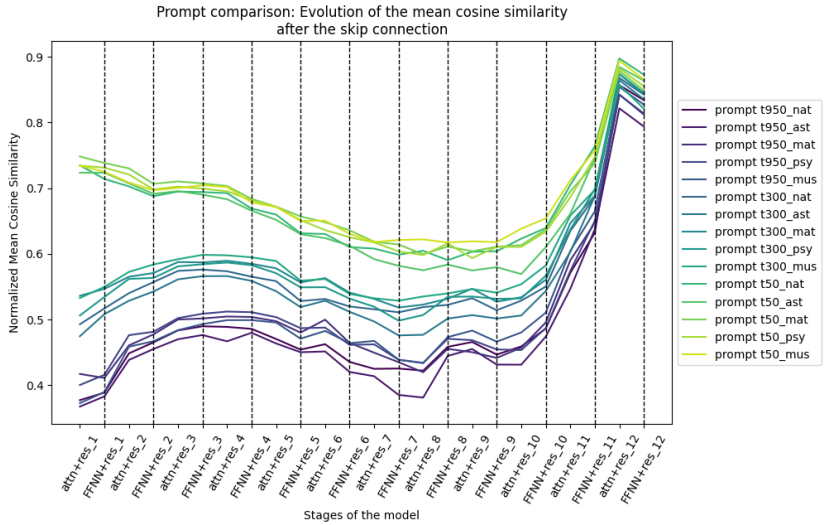
Statistical sample analysis

Prompt size comparison: [955, 980, 863, 889, 1003, 405, 434, 368, 409, 345, 63, 66, 58, 54, 62]



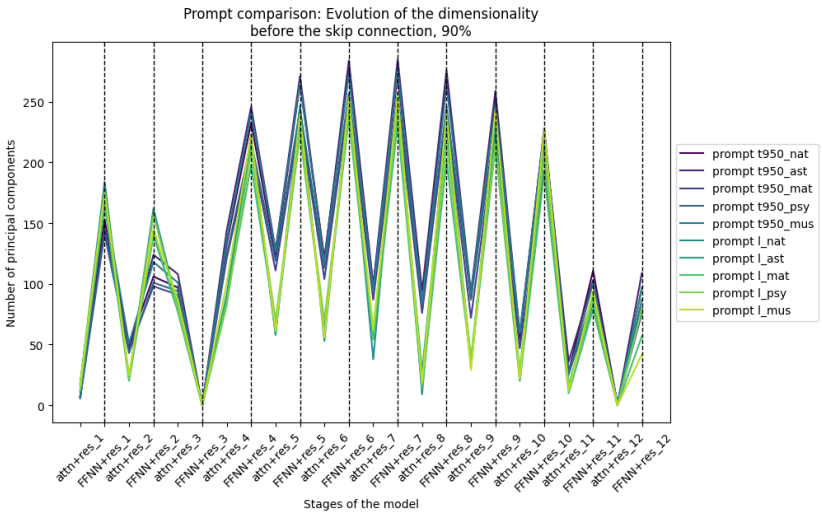


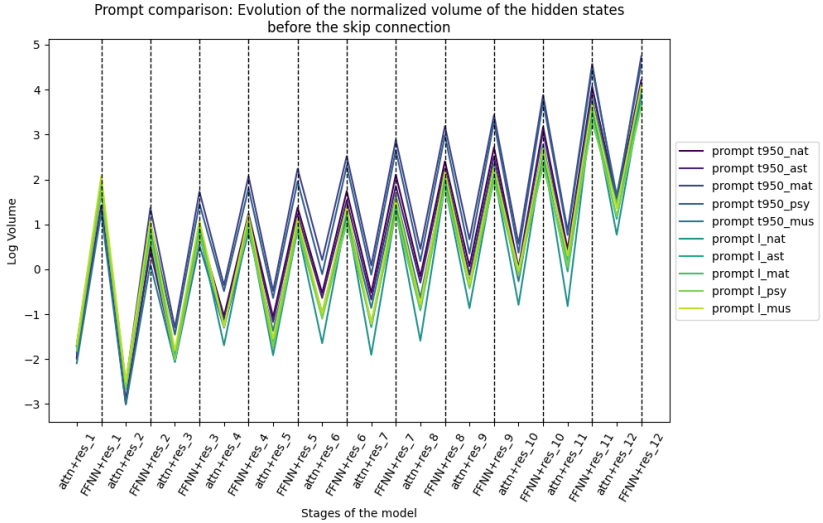


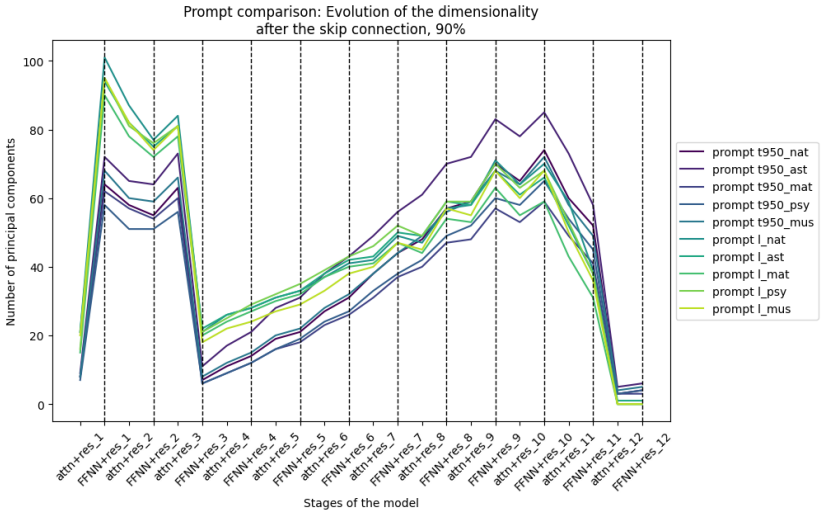


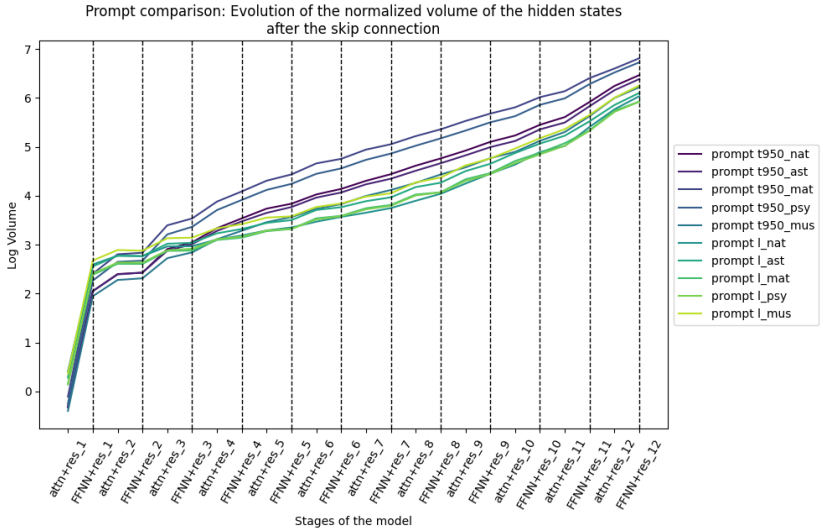
Statistical sample analysis

Prompt type comparison:



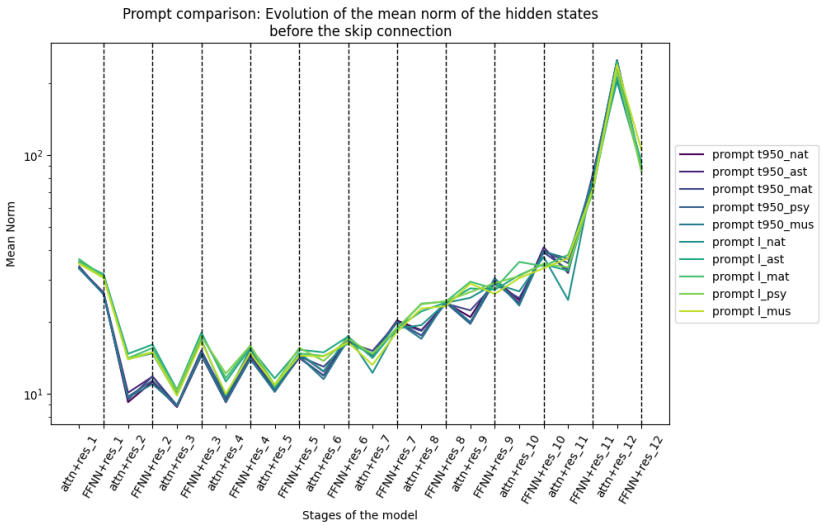


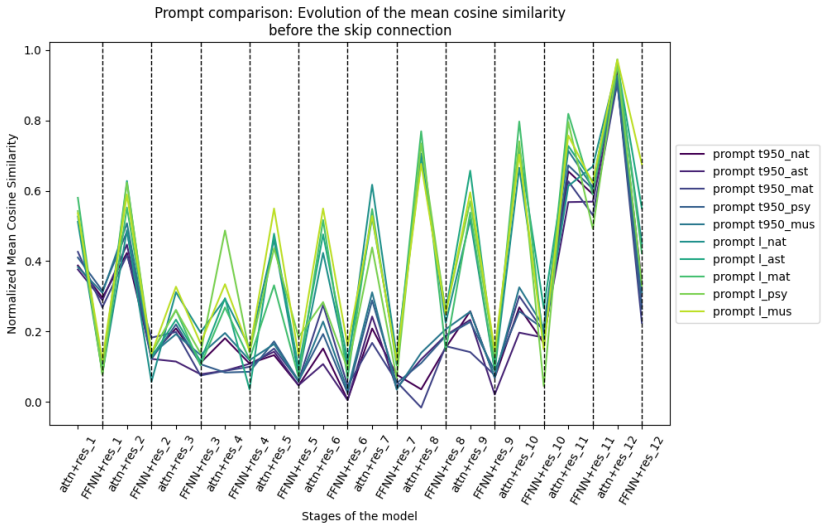


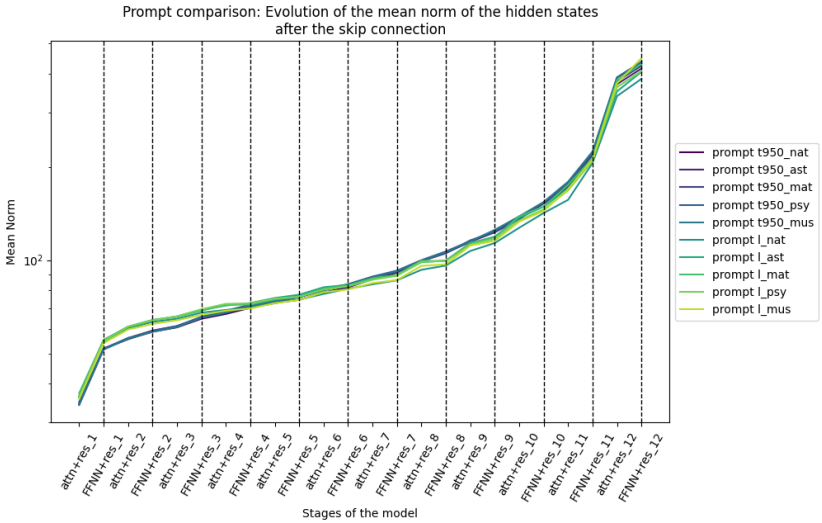


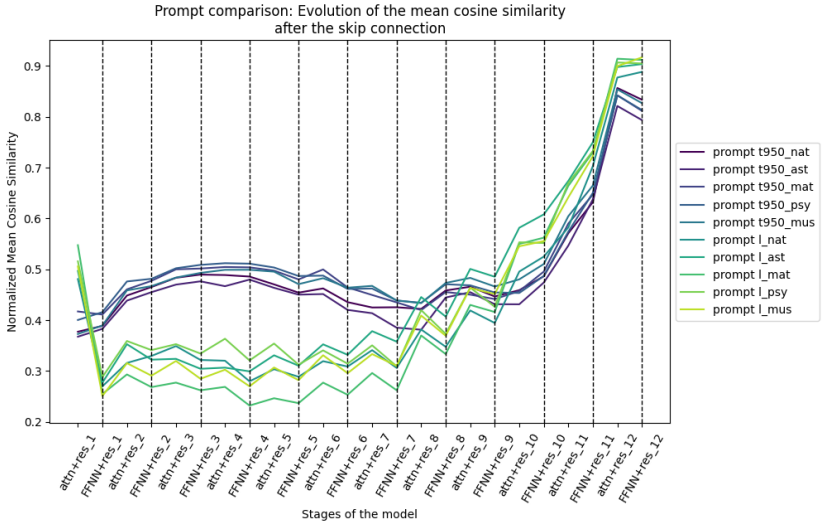
Statistical sample analysis

Prompt type comparison:

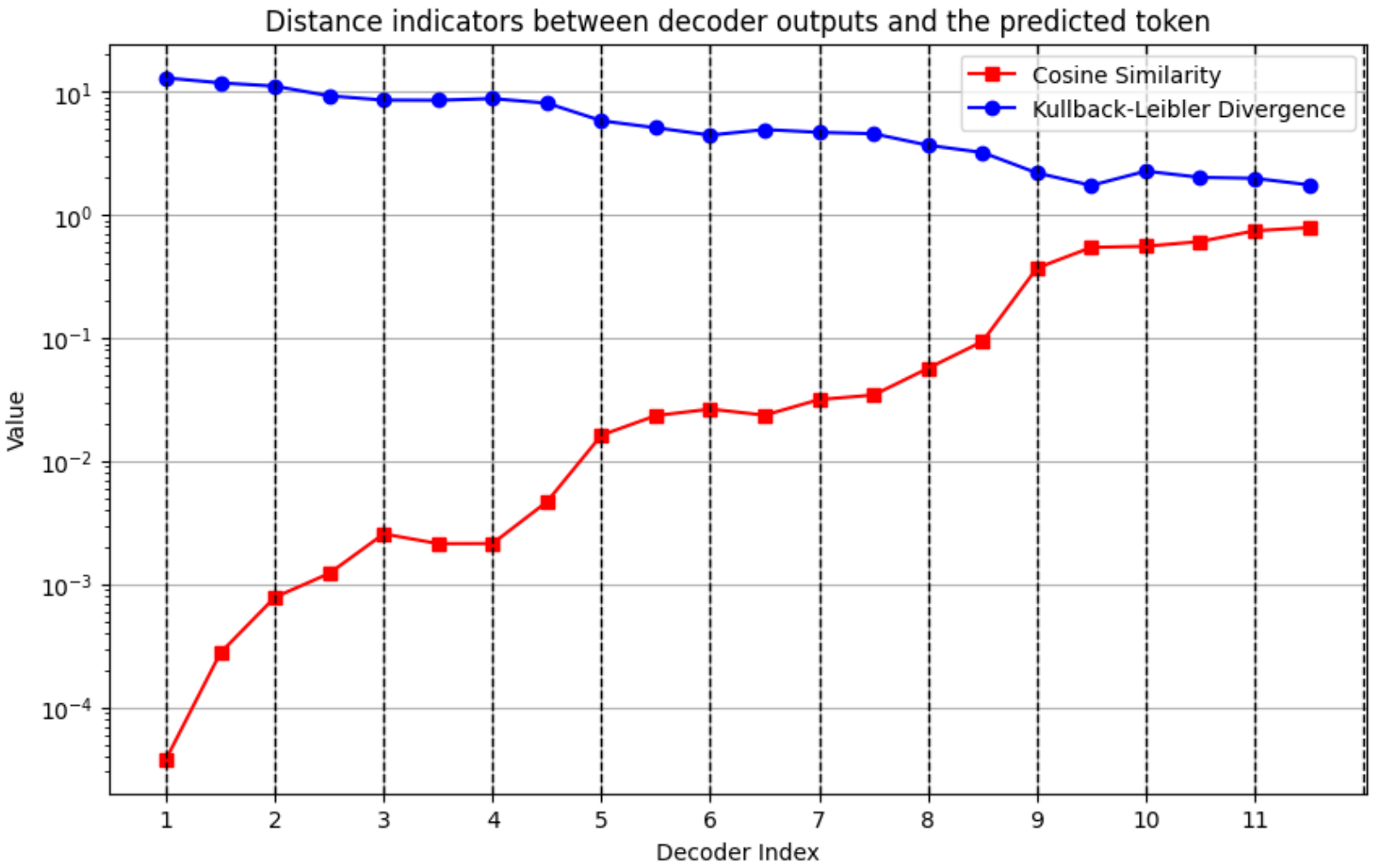




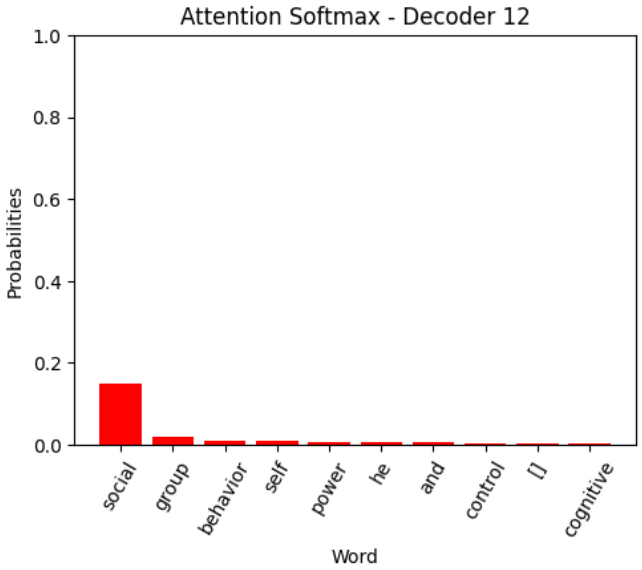


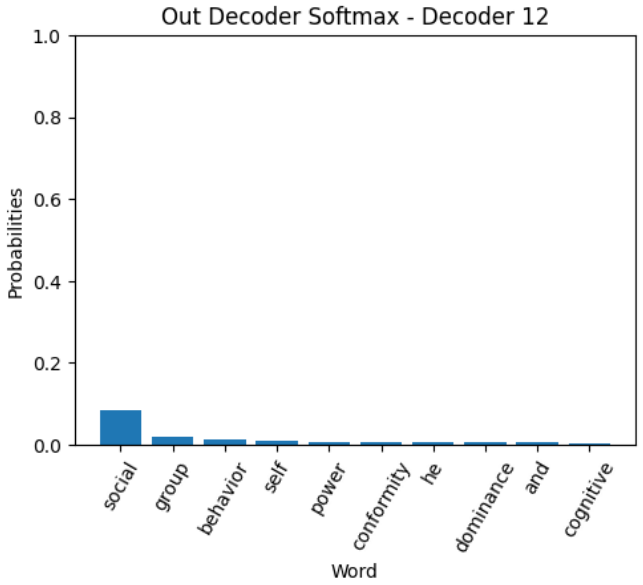


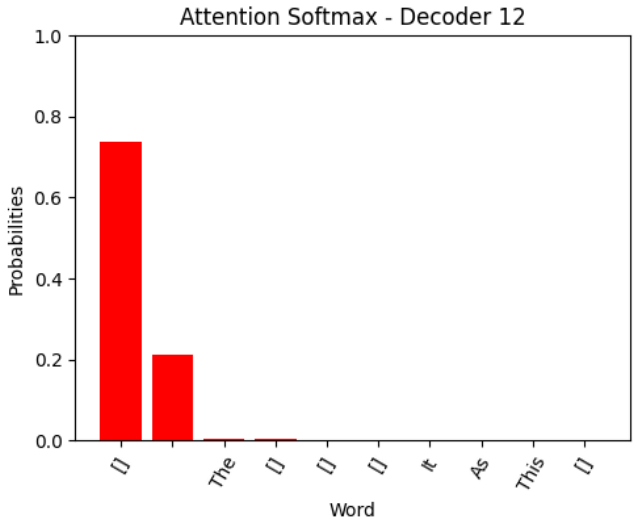
Last token analysis



Last token analysis







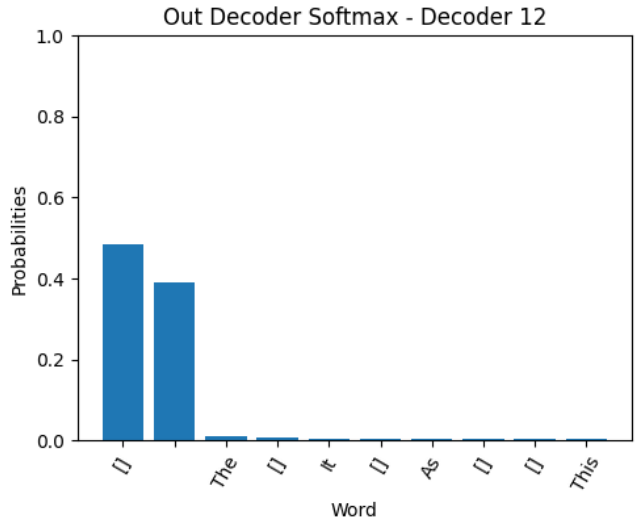


Table of Contents

Conclusions

- Opening the **transformer black box**
  - t-SNE and PCA plots show how the prompt is processed, layer-by-layer and decoder-by-decoder
- Inspecting the **embedding space**
  - Metrics and ID plots, the vectors are on a low-dim manifold (t-SNE)
  - Evolution of relevant quantities with statistical approach
- Studying the mechanism behind the **last token** prediction
  - GIF of the probability distribution evolution
  - Attention heads heatmaps express inter-dependencies between tokens


GPT-2
Thank you.