



UNIVERSITÀ  
DI TORINO

Università degli Studi di Torino

*Corso di Laurea Magistrale in Informatica*

## Concrete Numeric Representations in LLM Embeddings

Tesi di Laurea

**Relatore**

Prof. Di Caro Luigi

**Correlatore**

Dr. Torrielli Federico

**Candidato**

Gentiletti Emanuele  
900831

Anno Accademico 2024/2025



# Contents

<b>Introduction</b>	<b>1</b>
<b>Background</b>	<b>3</b>
The inductive bias of Tokenization . . . . .	3
Strategies for mathematical improvements through embeddings . . . . .	4
Savant syndrome and spatial representations . . . . .	5
Platonic representation hypothesis . . . . .	6
Dimensionality reduction and Embedding Visualization . . . . .	6
<b>Implementation</b>	<b>7</b>
Storage layer . . . . .	8
Extracting and Sampling . . . . .	10
Analysis . . . . .	12
User interface . . . . .	12
<b>Embeddings Analysis</b>	<b>15</b>
Methodology . . . . .	15
OLMo-2-1124-7B . . . . .	16
Linear analysis . . . . .	16
Non-linear analysis . . . . .	18
Correlation with mathematical properties . . . . .	20
Llama-3.2-1B-Instruct . . . . .	22
Linear analysis . . . . .	22
Non-Linear analysis . . . . .	25
Correlation with mathematical properties . . . . .	28
<b>Conclusions</b>	<b>29</b>
Key Findings . . . . .	29
Structured Numerical Embeddings . . . . .	29
Limitations . . . . .	29
Future Directions . . . . .	29
<b>Bibliography</b>	<b>31</b>



# Introduction

As above, so below. As within, so without.  
As the universe, so the soul.

---

*Emerald Tablet, misattributed*

This thesis takes a look at LLMs from the perspective of their embeddings, in particular their numerical ones. There are several reasons why I came to be interested in this topic, the first and naive one being that tokenization schemes, as naively implemented with the BPE algorithm, would leave a lot of space for improvement in numerical tasks, and it's interesting to explore how.

Better performance in LLMs has been sought through the lenses of scale, and looking for emergent properties as training time and resources increase. There are different reasons for this, one of them being The Bitter Lesson (Sutton, 2019), a heuristic principle that states that general methods that better leverage computation are better than methods that seek to use human domain-specific knowledge to inform the implementation. This has been observed, for example, in the domain of chess, where the strategies being put forward hard-coding human domain-specific knowledge were ultimately beaten by deep search.

After the big success story of scaling in LLMs, the main reach has been towards increasing model size and training on bigger datasets, and in unlocking the emergent capabilities that would come along those. While this approach has given results, albeit with some inconsistencies hard to reconcile from an epistemological perspective, such as the difficulty of actually designing good benchmarks for those abilities that actually verify they go beyond memorization (Skalse, 2023). What is proposed here is that we might be able to see this in more detail by seeing if there are established semantic links beyond relative measures through different representations in the same space, by checking the representation of numbers and the numbers that constitute that representation, and by seeing if there are semantic links that cross this barrier.

Along with this, we investigate here is the presence of structures that come to be through learning numerical representation. Given that we can naturally arrange numbers in a sequence, it comes natural to see what the disposition of those sequences form when arranged in the space of LLM representations. By taking inspiration from the Savant mode of human cognition, that comes through exploiting spatial arrangements as a means to perform calculations, we look for similar arrangements in LLMs, take a look at relevant research, and make hypotheses on why they come to be.



# Background

## The inductive bias of Tokenization

Modern LLMs are mostly autoregressive models built on the Transformer architecture (Vaswani et al., 2023). Transformers are a deep learning architecture based on attention, a mechanism that relates words in different positions in a sentence by computing weighted relationships between all input tokens, allowing the model to capture long-range dependencies and contextual relationships that sequential models like RNNs struggle with. The first step in most Transformer models is tokenization, which operates by converting input text into sequences of discrete tokens that are then mapped to high-dimensional vector representations. This initial step creates an inductive bias that shapes how the model processes information (Ali et al., 2024; Singh & Strouse, 2024), with significant implications for the application of numerical data to arithmetical tasks.

The most used algorithm for tokenization is currently Byte-Pair Encoding (Radford et al., 2019), which, given a fixed vocabulary size, starts with individual characters and iteratively merges the most frequently occurring pairs of adjacent tokens until the vocabulary limit is reached. This process naturally creates longer tokens for common substrings that appear frequently in the training data. For numbers, this means that frequently occurring numerical patterns like "100", "2020", or "999" might become single tokens, while less common numbers get broken into smaller pieces. The result is an idiosyncratic and unpredictable tokenization scheme where similar numbers can be tokenized completely differently based purely on their frequency in the training corpus. While GPT-2 used to have a purely BPE tokenizer, the successive iteration of GPT and generally more recent models either tokenize digits separately (so as '1234' → [1, 2, 3, 4]), or tokenize clusters of 3 digits, encompassing the integers in the range 0-999.

Most of the tokenizers right now do L2R (left-to-right) clustering (Millidge, 2023), meaning that a number such as 12345 would be divided in two tokens, 123 and 45. It has been shown (Singh & Strouse, 2024) that this kind of clustering leads to worse arithmetic performance, as this brings misalignment in digit positions and, as a consequence, in positional significance.

An even more surprising development is that forcing the R2L token clustering of numbers in models already trained with L2R clustering through the use of commas in the input (ex. 12, 345) leads to big improvements in arithmetic performance (Millidge, 2024; Singh & Strouse, 2024). Despite the model learning representations adapted to work with a L2R token clustering strategy, forcing a R2L clustering at inference time shows substantial improvements in arithmetic tasks, which means that despite being learned through an unfavorable tokenization approach, the numeric representations retain the properties that allow for the performance to improve when the digit clustering scheme is corrected.

This could be happening for different reasons, for example:

- Arithmetic operations would still work locally in the 0-999 range, which allows for a correct reading on them and possible generalization on a larger scale, which counteracts the unintuitive clustering scheme.
- The forced tokenization also happens in the data, as numbers are often separated by punctuation in clusters of 3 digits, right to left, for legibility reasons (Singh & Strouse, 2024).

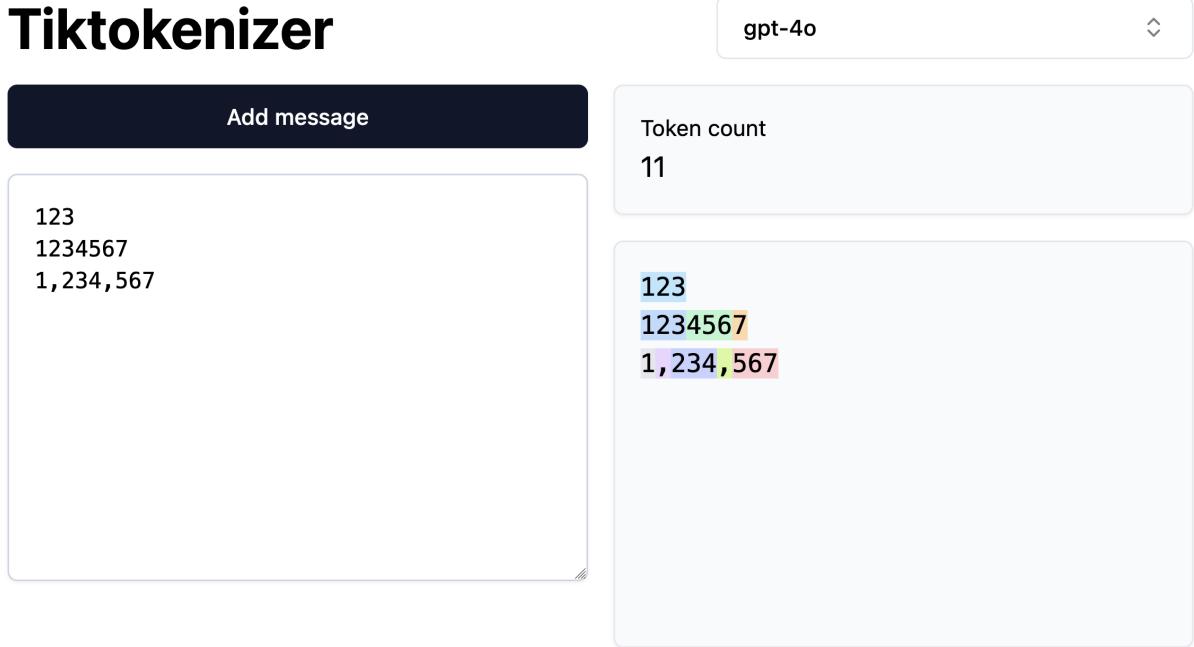


Figure 1: GPT-4o tokenization of different numerical quantities, displaying the L2R clustering and the comma trick to force R2L clustering.

- There's a geometric bias towards the right mode of operation given by the structures that form in the space to compute mathematical operations. This is the hypothesis that will be explored.

At the very least, the data being biased towards a R2L representation (in the form of using the Arabic number system and adopting legibility rules that accommodate right to left calculations) leads to embeddings that maintain that bias even when learned in a L2R fashion. This can be a possible hint towards the optimality of certain representations compared to others, given the resilience in preferring a certain tokenization scheme over the one the model is trained on.

Model	Strategy
LLaMA 1 & 2	Single digit
LLaMA 3	L2R chunks of 3 digits
OLMo 2	L2R chunks of 3 digits
GPT-2	Pure BPE
GPT-3.5 / GPT-4	L2R chunks of 3 digits
Claude 3 / Claude 4	R2L chunks of 3 digits

Table 1: Language models with their respective tokenization strategy for numbers.

## Strategies for mathematical improvements through embeddings

Beyond improving tokenization, there have been other, more comprehensive approaches to the improvement of the representation of numeric values. xVal is a notable one, as its approach encompasses real numbers beyond just integers and does away with learning different representation for each number.

The idea is maximizing the inductive bias in the representation by having embeddings that are computed based on the number to be represented (Golkar et al., 2023). Numerical values represented by a single

embedding vector associated with the [NUM] special token, that gets scaled on the basis of the numerical value to represent. There is an assumption made so that this works: the semantics of magnitude work as they pass from the represented object to the structure of the representation.

This fits very well with the idea of reification, which will be described later (Murray, 2010): the embedding, beyond its qualities as representational object, becomes an entity with features that actively aid in the calculation process.

The model uses two separate heads for number and token predictions. If the token head predicts a [NUM] token as the successor, the number head gets activated and outputs a scalar. The rest of the weights in the transformer blocks are shared, allowing the learning of representations that are useful for both discrete text prediction and continuous numerical prediction. This means the model develops number-aware internal representations throughout all its layers, not just at the output. The shared weights force the model to learn features that work for both linguistic and mathematical reasoning simultaneously.

The approach is shown to improve performance over a series of other techniques, mostly using a standard notation to represent numbers (Golkar et al., 2023). This very thesis has been inspired by the xVal paper, with one of its initial goals being to find good representations for computed numerical embeddings.

There are several different approaches to improving math performance in LLMs that don't necessarily come directly from training, but from a better understanding of how representations work. Other approaches include giving models better positional information about digits within numbers, such as Abacus Embeddings (McLeish et al., 2024), which encode each digit's position relative to the start of the number and can improve arithmetic performance substantially.

While on one hand particular modes of numerical cognition can be explored through explicitly reifying the representation (through the xVal approach), what the rest of the analysis hinges on is whether LLMs develop such structures on their own, by looking at their embeddings, as this could hypothetically inform us on how to build these structures ourselves in a more direct way than training.

## Savant syndrome and spatial representations

Savant syndrome is a rare condition in which people show exceptional proclivity towards certain specific activities, usually accompanied by great impairments in other areas of their lives. Savants can have exceptional abilities in math, art, music and other fields, as well as instant calculation abilities that don't seem to come through algorithmic processing.

A case study of a Savant patient DT (Murray, 2010) reveals a mathematical cognitive architecture with the following characteristics:

- has sequence-space synesthesia with a “mathematical landscape” containing numbers 0-9999
- each number possesses specific colors, textures, sizes, and sometimes movements or sounds
- prime numbers have distinctive object properties that distinguish them from other numbers
- arithmetic calculations happen automatically
- solutions appear as part of his visual landscape without conscious effort
- fMRI studies showed that even unstructured number sequences had coherent visual structure for DT.

Murray argues that savants possess highly accessible concrete representations of abstract concepts, for which she uses the term reification - the conversion of abstract concepts into concrete, spatial entities that can be directly “inspected” rather than computed.

Sequence-space synesthesia is the spontaneous visualization of numerical sequences in organized spatial arrangements. The remarkable mathematical abilities of savants with this condition suggest that their specialized perceptual representations confer significant computational advantages over normal human numerical calculation abilities.

Given that the spatial arrangement confers advantages in numerical calculation to the subject, we can pose the question: are there specific spatial arrangements that enable advantageous numerical calculations, and

are those present or replicable in LLMs? The spatial idea is easily translatable from the perceptive sphere to the representational one, by considering LLM embeddings. If these come through because of geometric properties of the structures, and going with the assumption that those structures are replicable in the high-dimensional vector spaces we're working with, it would follow that strict optimization through gradient descent could be a possible way to make them come about.

## Platonic representation hypothesis

According to (Huh et al., 2024) AI models, particularly deep networks, are converging. The central hypothesis is that different models are converging toward a shared statistical model of reality, akin to Plato's concept of an ideal reality. This representation is termed the Platonic representation.

This convergence appears to be driven by several selective pressures: larger models have more capacity to find optimal representations; models trained on more diverse tasks are constrained to find solutions that work across multiple domains; and deep networks have implicit biases toward simpler solutions.

For the investigation of numerical representations, this suggests that if there are indeed optimal geometric structures for mathematical reasoning, different models might naturally converge toward them during training. The shape suggested (the helix) has properties on an information-theory basis that make its use as a learning geometry more likely (Kantamneni & Tegmark, 2025). In particular, their self-similarity can be a useful error-correcting property.

## Dimensionality reduction and Embedding Visualization

To visualize and analyze the high-dimensional embedding spaces that LLMs use for numerical representations, we need techniques that make the underlying structure evident. For this reason, we employ the following dimensionality reduction techniques:

- **SVD (Singular Value Decomposition)** is a fundamental matrix factorization that decomposes any matrix  $A$  into three component matrices:  $A = U\Sigma V^T$ , where  $U$  and  $V$  contain orthogonal vectors (left and right singular vectors respectively) and  $\Sigma$  contains the singular values on its diagonal. SVD reveals the underlying structure of the matrix by identifying the principal directions of variation and their relative importance through the singular values.
- **PCA (Principal Component Analysis)** emerges as a specific application of SVD. By applying SVD to a centered data matrix (where each variable has been mean-centered), the right singular vectors  $V$  become the principal components - the directions of maximum variance in the data. The singular values in  $\Sigma$  are directly related to the eigenvalues of the covariance matrix.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding)** (Maaten & Hinton, 2008) converts similarities between data points in high-dimensional space into probabilities, then uses gradient descent to minimize the divergence between these probabilities and those of points in a low-dimensional embedding. It excels at preserving local neighborhood structure, making clusters very distinct in the visualization. However, t-SNE can distort global structure and distances between distant clusters become less meaningful, making it primarily useful for identifying local groupings in numerical embeddings.
- **UMAP (Uniform Manifold Approximation and Projection)** (McInnes et al., 2020) also preserves local structure like t-SNE, but additionally maintains more of the global structure through its foundation in topological data analysis. UMAP constructs a topological representation of the data in high dimensions, then uses stochastic gradient descent to optimize a low-dimensional representation to have similar topological properties. This makes it better suited for analyzing how models organize numerical concepts across different scales - both local clusters of similar numbers and global relationships between distant numerical regions.

# Implementation

The aim of the project was the realization of a framework to enable the analysis and visualization of embeddings. To achieve this, we used three parts:

- an extraction and sampling layer, used to load models from HuggingFace and extracting the parts of the embeddings layer of interest for this inquiry.
- a storage layer, meant to store the selected samples in an easy to retrieve manner. In particular, during the course of this project we focused on sampling integers and some random embeddings, to have a comparison to see whether the structures formed were artifacts of the dimensionality reduction technique used.
- a CLI, to provide a user interface for the download, extraction and sampling of embeddings from HuggingFace models
- an analyzer, meant to compute base statistics about the embeddings and to extract results from the PCA analysis and <?> give insights on variance by component
- a visualizer, to create plots that give a visual intuition of the structures underneath the embedding data.
- a dashboard, provided through a Marimo notebook to show interactive visualization and to provide interactive data analysis features.

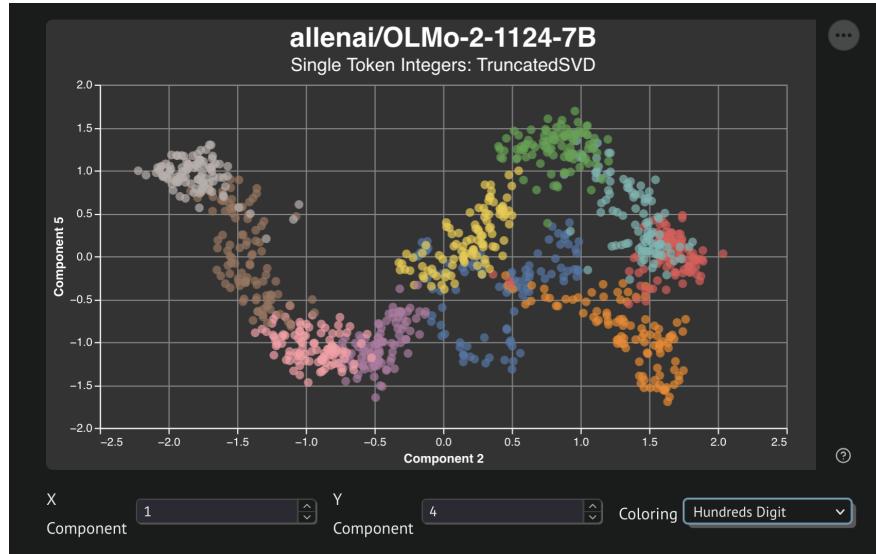


Figure 2: Plotting component for the Marimo dashboard.

## Storage layer

The necessity of a storage layer came out of a necessity to be able to work with the data in manageable way, as loading the whole model was very often impractically slow and a lot of the work had to be done in a resource-constrained environment.

The sample data was stored in instances of the `EmbeddingsSample` class `lst.0.1`, along with metadata reporting the source model ID and, for random samples, the seed used for replicability purposes.

---

**Listing 0.1** Container classes for embeddings samples and their metadata.

---

```
@dataclass
class IntegerSampleMeta:
    model_id: str
    tag: Literal["integers"] = "integers"

@dataclass
class RandomSampleMeta:
    model_id: str
    sample_size: int
    seed: int
    tag: Literal["random"] = "random"

@dataclass
class ReducedSampleMeta:
    original: EmbeddingsSampleMeta
    estimator: BaseEstimator
    tag: Literal["reduced"] = "reduced"

@dataclass
class EmbeddingsSample[M: EmbeddingsSampleMeta]:
    sample_id: int
    meta: M
    embeddings_df: pd.DataFrame = field(repr=False)
```

---

Initially, the storage of each sample was done in a dedicated Parquet file, an efficient file format that would have provided easy serialization of Pandas dataframes, which were the main data structure employed in the analysis. While initially adequate, this implementation didn't allow for easy sample metadata storage, and required an ad-hoc cataloguing system based on filesystem names to store and retrieve items on the basis of their metadata.

To address this, a choice was made to implement a more proper storage layer. It was realized using DuckDB, a database similar to SQLite that uses a single file to store data and provides vector functionality for the storage of the embeddings. This allowed to store and retrieve the embeddings, and their metadata, in a more robust way. DuckDB also offers facilities to work directly on dataframes using SQL queries, and exchanging data between dataframes and the database in this way, which revealed very useful for loading purposes.

---

**Listing 0.2** SQL schema for the storage layer.

---

```
CREATE OR REPLACE TABLE embeddings (
    model_id VARCHAR NOT NULL,
    token_id INTEGER NOT NULL,
    token VARCHAR NOT NULL,
    embeddings FLOAT[] NOT NULL,
    PRIMARY KEY (model_id, token_id),
);

CREATE OR REPLACE SEQUENCE embedding_id_seq;
CREATE OR REPLACE TABLE samples (
    sample_id INTEGER DEFAULT NEXTVAL('embedding_id_seq') PRIMARY KEY,
    model_id VARCHAR NOT NULL,
    meta JSON,
    created_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP,
);

CREATE OR REPLACE TABLE embedding_to_sample (
    model_id VARCHAR NOT NULL,
    token_id INTEGER NOT NULL,
    sample_id INTEGER NOT NULL,
    PRIMARY KEY (model_id, token_id, sample_id),
);
```

---

## Extracting and Sampling

Extraction is performed by downloading models using HuggingFace's Transformers library, which allows for download and deployment of popular open source models.

---

**Listing 0.3** Extraction class for embeddings

---

```
class HFEmbeddingsExtractor:
    """Extracts embeddings from a Hugging Face model."""

    def __init__(self, name_or_path: str):
        self.name_or_path = name_or_path

    @cached_property
    def embeddings(self):
        model = AutoModel.from_pretrained(self.name_or_path)
        model.eval()
        embeddings = model.embed_tokens
        return embeddings

    def extract(self, token_ids):
        with torch.no_grad():
            token_ids = torch.tensor(token_ids)
        return self.embeddings.forward(token_ids).squeeze().numpy()
```

---

LLMs that make use of a tokenization step receive their sentences in input as a list of token IDs, where each token ID corresponds to an embedding vector. It is the LLM's tokenizer responsibility to take sentences, split them at the appropriate token boundary, adding special tokens where necessary, and convert them into token IDs for the LLM processing.

The logic to do this is split between the `HFTokenizerWrapper` and `HFEmbeddingsSampler` classes. `HFTokenizerWrapper` invokes the tokenizer to get the token IDs that correspond to the embeddings of interest (avoiding special tokens, like `<Beginning of Sentence>` and such), while `HFEmbeddingsSampler` has the logic for integer and random selection.

The sampling happens by first picking the tokens of interest. For numbers, we first verify that the model uses a tokenization scheme useful for the numeric analysis intended, by trying to tokenize each integer from 0 onwards, up to a maximum ceiling of 10.000, until we find the first integer that gets tokenized using more than one token. The analysis is limited in scope to single-token integers in the range 0-999, as these parameters correspond to a multitude of open source models available as of today.

After the sampling process is completed, the results are returned as a dataframe, along with the corresponding provenance metadata.

For the random sampling of embeddings, token ids are picked by doing a random extraction of numbers between 0 and the vocabulary size of the model, and then proceeding in a similar way as described before `<?>`.

---

**Listing 0.4** HFTokenizerWrapper class, providing utility functions for tokenization.

---

```
class HFTokenizerWrapper:
    """Wrapper for Hugging Face tokenizers."""

    def __init__(self, tokenizer):
        self.tokenizer = tokenizer

    def tokenize(self, tokens) -> torch.Tensor:
        return self.tokenizer(
            tokens,
            add_special_tokens=False,
            return_attention_mask=False,
            return_tensors="pt",
        )["input_ids"]

    @classmethod
    def from_pretrained(cls, model_id):
        tokenizer = AutoTokenizer.from_pretrained(model_id)
        return cls(tokenizer)

    def token_ids_to_tokens(self, token_ids):
        tokens = self.tokenizer.convert_ids_to_tokens(token_ids)
        return [token if token is not None else "<unk>" for token in tokens]
```

---

---

**Listing 0.5** Code for HFEmbeddingSampler

---

```
class HFEmbeddingsSampler:
    def _single_token_integer_ids(self, max_value=10_000) -> Iterable[int]:
        for num in range(max_value):
            token_ids = self.tokenizer.tokenize(str(num)).squeeze()
            if token_ids.ndim == 0:
                yield token_ids.item()
            else:
                return
        warnings.warn(
            f"All integers from 0 to max_value={max_value} are single token."
            "There may be more single-token integers."
        )

    def single_token_integers(self) -> tuple[pd.DataFrame, IntegerSampleMeta]:
        token_ids = np.fromiter(self._single_token_integer_ids(), int)
        tokens = range(len(token_ids))
        embeddings = self.extractor.extract(token_ids)

        df = make_embeddings_df(token_ids, tokens, embeddings)
        meta = IntegerSampleMeta(model_id=self.model_id)

        return df, meta

    ...
    def _random_token_ids(self, sample_size, seed):
        rng = np.random.default_rng(seed)
        return rng.choice(self.tokenizer.vocab_size, size=sample_size, replace=False)
    ...
```

---

## Analysis

Most of the analysis is done in the `EmbeddingsAnalyzer` class, which reorganizes data and provides it in a format suitable for consultation and visualization.

---

**Listing 0.6** Initializing code for `EmbeddingsAnalyzer`

---

```
@dataclass
class EmbeddingsAnalyzer:
    embeddings_df: pd.DataFrame
    meta: EmbeddingsSampleMeta

    @classmethod
    def from_sample(cls, sample: EmbeddingsSample):
        """Initialize from an EmbeddingsSample."""
        return cls(
            embeddings_df=wide_embeddings_df(sample.embeddings_df),
            meta=sample.meta,
        )
```

---

The format used for the embeddings here is a dataframe with the columns `token`, `token_id` and the embeddings spread out in columns named `embeddings_{dimension index}`. Even though it's a little unwieldy, this allows for compatibility with most of the libraries operating with dataframe that assume mono-dimensional column indices with string column names. This class also provides the facility for dimensional reduction through the `run_estimator` method, which takes an estimator as input and returns a new `EmbeddingsAnalyzer` instance with the embeddings being fit through the estimator.

A notable feature implemented here is the analysis of the correlations between embedding features and mathematical sequences, done in the `feature_to_sequence_analysis_df` method. This is done by generating various mathematical sequences and then encoding them using either:

- Direct encoding, for the ones that don't grow too much in value, like  $\log_n$  or  $n$ . This simply means that the mathematical sequence vector the feature is tested against is produced by directly inserting the relative values, like  $[\log(0), \log(1), \log(2), \dots]$
- One-hot encoding, for faster growing sequences. The indices that correspond to a value contained in sequence get the value 1, the others get the value 0.
- Gaussian-smoothed one-hot encoding, where the values are passed through a Gaussian filter to check for smoother feature detection.

As a result of the analysis, a dataframe is produced that provides data about features and their respective correlations.

## User interface

The end user can make use of the tools by loading a model through the CLI, which was programmed using the `typer` library. It can be used to load model numerical and random samples into the database through the command `embcli load <hf_model_id>`, which can then be listed and consulted through the Marimo dashboard.

---

**Listing 0.7** Code for mathematical sequence encoding.

---

```
def one_hot_encode(sequence, size):
    return np.isin(np.arange(size), sequence).astype(int)

def one_hot_gaussian_smooth(binary, sigma=2.0):
    return gaussian_filter1d(binary.astype(float), sigma=sigma)

def make_encoded_sequences(max_token: int, sigma: float = 2.0):
    encoded_sequences = {}

    direct_sequences = direct_encoded_base_sequences(max_token)
    for name, seq in direct_sequences.items():
        encoded_sequences[name, "direct"] = seq

    binary_sequences = binary_encoded_base_sequences(max_token)
    for name, seq in binary_sequences.items():
        one_hot = one_hot_encode(seq, max_token)
        encoded_sequences[name, "binary"] = one_hot
        encoded_sequences[name, "gauss"] = one_hot_gaussian_smooth(one_hot, sigma=sigma)

    return encoded_sequences
```

---



# Embeddings Analysis

## Methodology

The analytic part of this work consists in the search for structures in LLM numerical embeddings. Two models are taken in consideration:

- OLMo-2-1124-7B (OLMo et al., 2025) is a model by AllenAI, which is favorable to research uses thanks to the full disclosure of training data, code, logs and checkpoints. This model was trained on 4.05T tokens using a two-stage curriculum: initial pretraining on 3.90T tokens of web-based data, followed by a specialized mid-training phase on 150B high-quality tokens including 10.7B synthetic mathematical data specifically designed to enhance numerical reasoning capabilities.
- Llama-3.2-1B-Instruct (Grattafiori et al., 2024), due to being a small and manageable model to do analysis with on limited hardware. This model derives from the Llama 3 family, which was trained on approximately 15T multilingual tokens with dynamic data mixing throughout training, including mathematical content integrated continuously rather than in separate phases.

Both models underwent single-pass training without data repetition, meaning their numerical representations developed through one-time exposure to mathematical content.

For each of these, dimensionality reduction is applied through PCA, SVD, t-SNE and UMAP, and the results are used to produce 2D and 3D visualization meant to show the geometric structure that the representation assume in the space.

Then, an analysis of correlation with mathematical sequences is done: for each dimension, we encode mathematically interesting sequences the embedding features might be detecting for, and measure their Pearson correlation coefficient along with their respective p-value. We adopt the following encoding for sequences:

- direct, meaning we test the correlation directly with the sequence. We do this with plain integers in sequence (0, 1, 2, 3) and their logarithm ( $\log(1)$ ,  $\log(2)$ , ...)
- one hot encoded, for the sequences where the growth would be too fast to have meaningful detection through direct encoding. We tried this for Fibonacci, triangular, and prime numbers
- Gaussian-smoothed one-hot encoded, same as last point but allowing a more gradual climb around the positions corresponding to the numbers by applying a Gaussian smoothing filter to the one-hot encoded vector.
- Fourier encoding, as  $\sin\left(2\pi \cdot \frac{n_i}{T}\right)$  and  $\cos\left(2\pi \cdot \frac{n_i}{T}\right)$  where  $n_i$  is the element of the sequence to be encoded and  $T \in \{1, 2, 5, 10, 100\}$  is a collection of possible periods of encoding, as suggested in (Kantamneni & Tegmark, 2025) as a possible way for features to encode numeric characteristics.

# OLMo-2-1124-7B

## Linear analysis

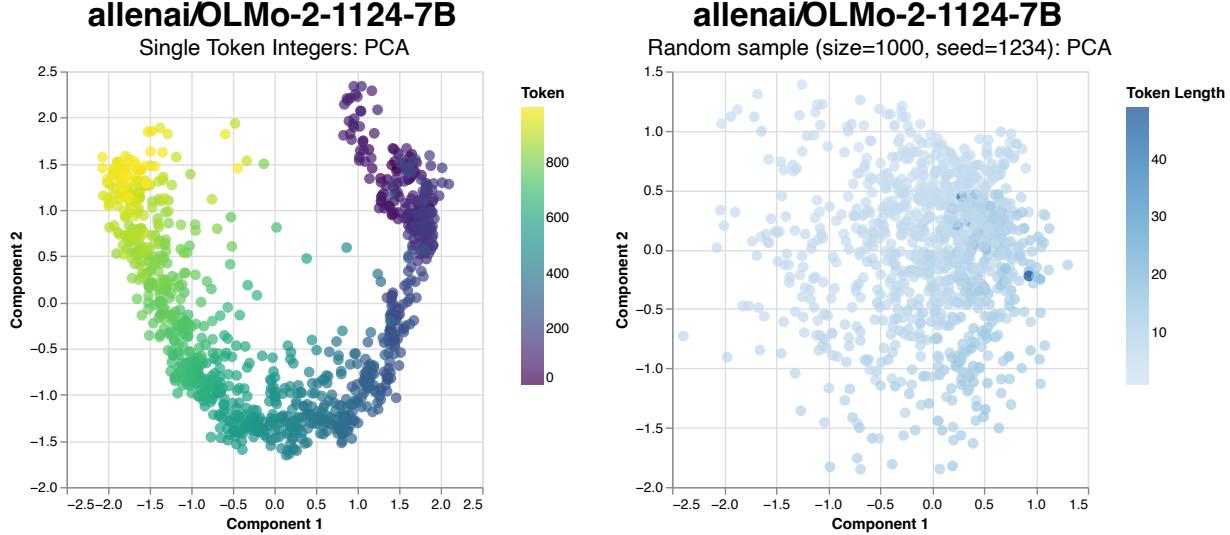


Figure 3: Principal components 1 and 2 of the OLMo model. Random embeddings sample for comparison.

In fig. 3, we plot the first two principal component of the numerical embeddings, color-grading them on the basis of their numerical value. The result shows that numerical tokens do not occupy the embedding space randomly: they follow a constrained path that preserves numerical relationships, suggesting that the model has learned to encode ordering properties of the numbers within its representation. The gradient is particularly smooth, suggesting that similar numbers maintain spatial proximity in the reduced space. It's also notable that in the top right of the curve, there seems to be a smaller curve forming; it's better visible through the SVD visualizations, but that part corresponds to single and double-digit integers replicating the bigger overall structure of the lower curve.

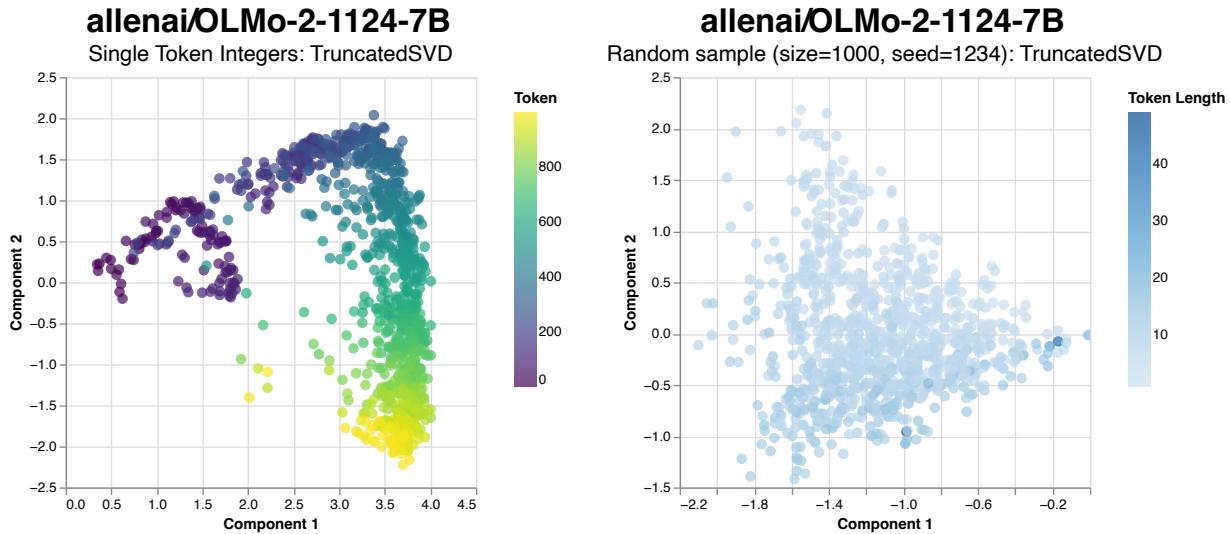


Figure 4: SVD for the two main components of the OLMo model, with random embeddings sample for comparison

The relationship and similarities are even more clear in fig. 4, which, lacking the data centering done in the

PCA, shows a much more consistent geometric structure, showing that the encoding of information likely happens in absolute distances rather than just with relative positioning between data points. This will also inform the strategies we use for reconstructing the datapoints with UMAP, as we'll do projections that make use of Euclidean distance as well as cosine similarity as a metric.

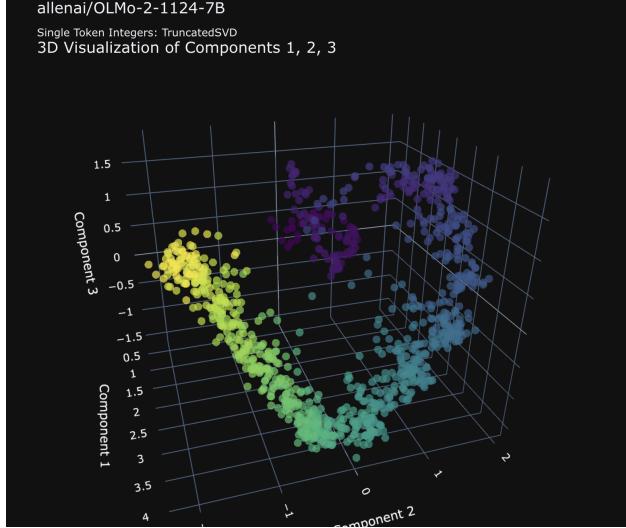


Figure 5: OLMo-2-1124-7B 3D visualization of SVD components .

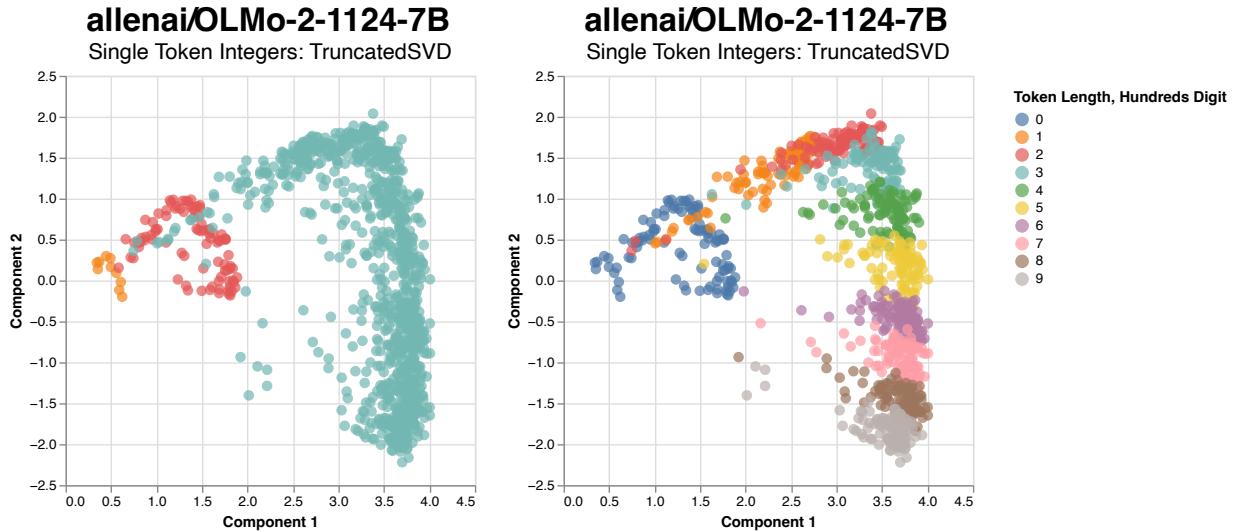


Figure 6: SVD coloring done by digit length and hundreds digit, highlighting the clustering properties of the embeddings.

Looking at fig. 6, the self-similar structure repeating through different integer lengths is striking. Models like Qwen-2.5 do away with tokenizing numbers outside the single digits from 0 to 9, and this picture can offer a compelling explanation on why that can be justified. In fact, it seems like the encoding of higher digit quantities brings along a lot of redundancy. On the other hand, as discussed for (Kantamneni & Tegmark, 2025), this same redundancy could be used by the models as an error correcting mechanism when it has to apply numerical operations, leading to possibly better performance.

## allenai/OLMo-2-1124-7B

Single Token Integers: PCA

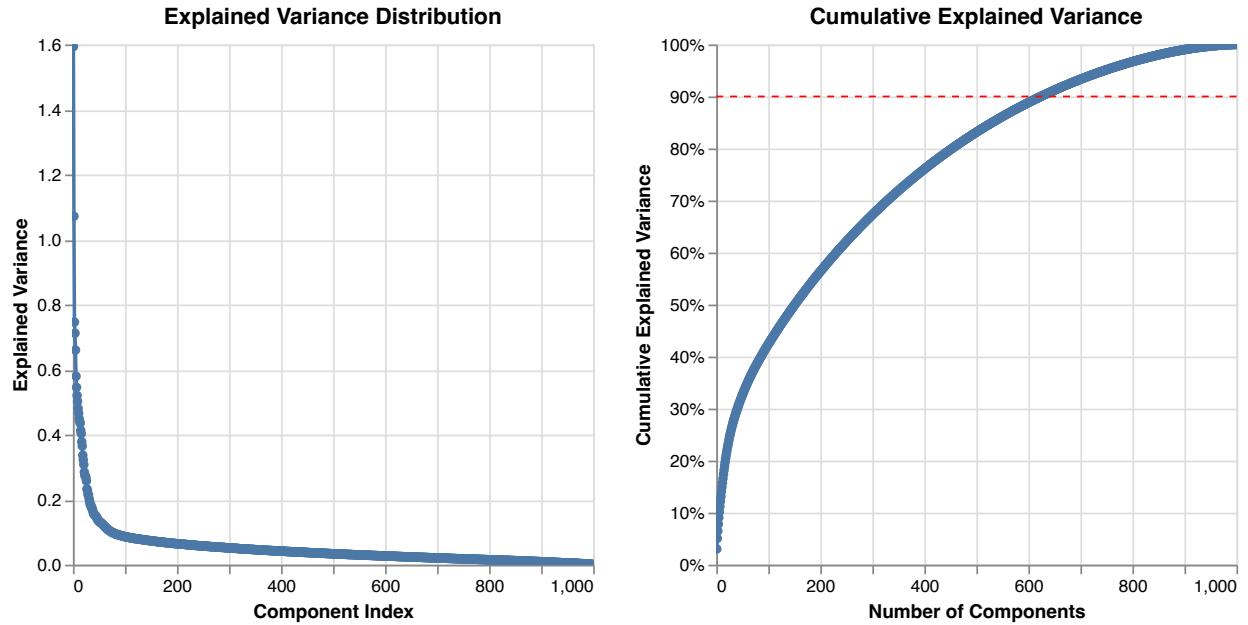


Figure 7: OLMo PCA - explained variance overview

### Explained variance

The explained variance by component plot (fig. 7, left) shows a sharp drop within the first few components, meaning that the first principal components capture dramatically more variance than subsequent ones. The cumulative explained variance (right) shows that approximatively 600 principal components are needed to reach 90% of explained variance.

By this we can conclude that the embeddings have a much lower intrinsic dimensionality than their full 4096 dimensions, and that they lie on a low-dimensional manifold in the full representation space. Only one-fifth of the total embedding space is necessary to capture 90% of the variance, and, as described earlier, the structures already encoded provide already a lot of redundancy.

### Non-linear analysis

Parameter	Value
perplexity	75
max_iter	3000
learning_rate	50
early_exaggeration	20
random_state	42

Table 2: t-SNE hyperparameters for the presented plots.

The t-SNE visualization in fig. 8 shows a distinctive branching pattern emanating from a central region, with low numbers at the center and higher ones radiating outward. The color progression follows these branches, indicating that numerical sequences are preserved along each arm. The gradient seems also to transition

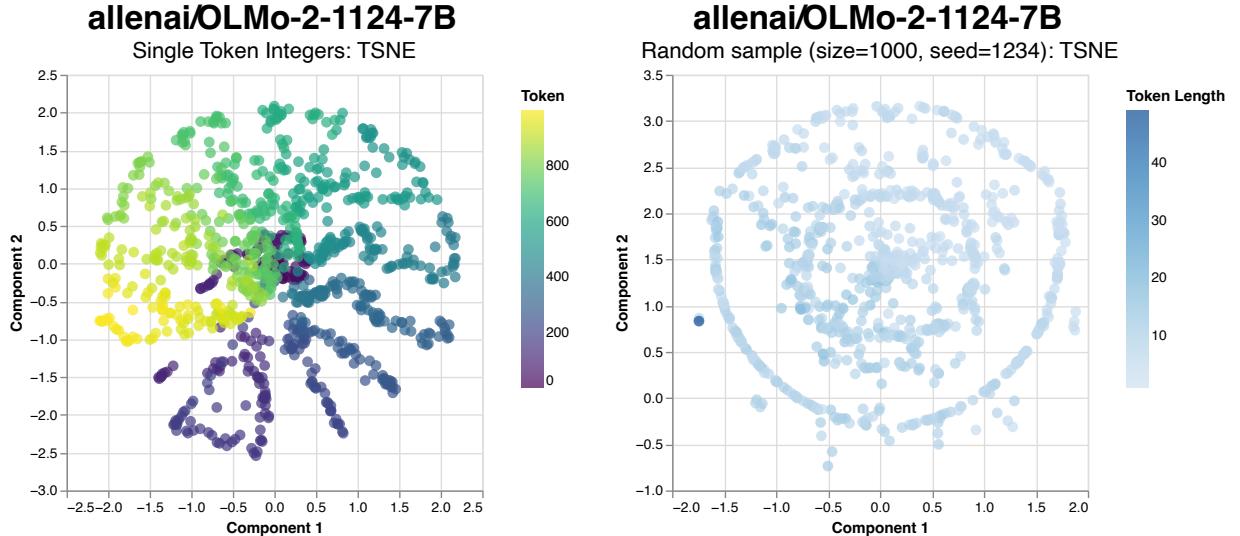


Figure 8: t-SNE visualization for OLMo embeddings.

circularly; branches with gradually increasing numbers turn around the center before abruptly getting back to the start. When interpreting the colors as indicators of depth, it can look like a spiral from a top-down perspective, giving a visual confirmation of what has been said about helical structures in (Kantamneni & Tegmark, 2025).

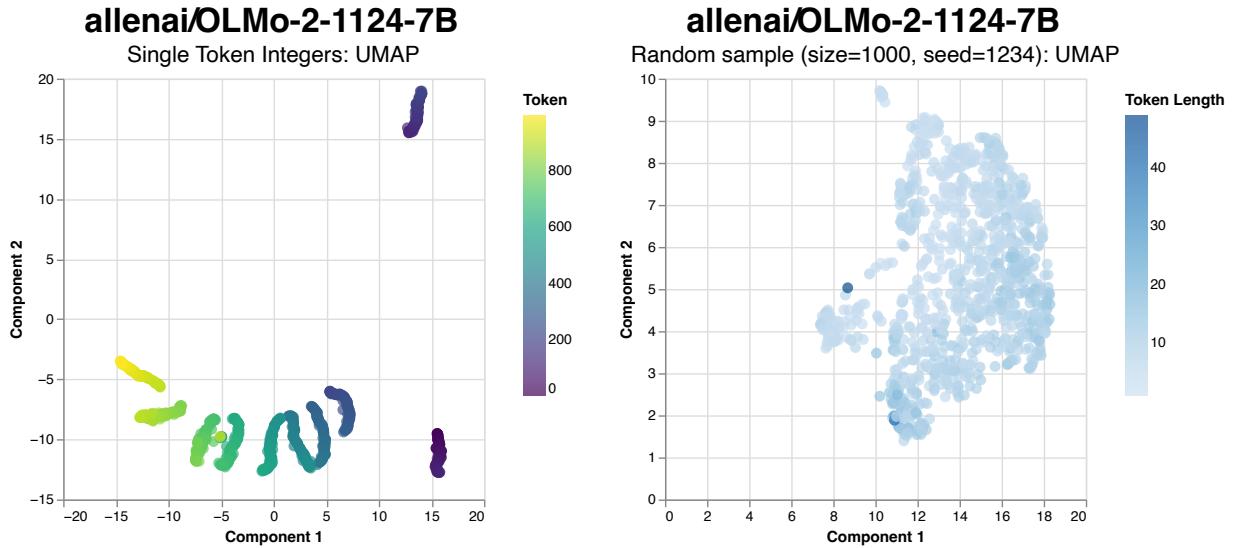


Figure 9: UMAP visualization with cosine distance

UMAP has been run using both Euclidean (fig. 10) and cosine distances (fig. 9), since the SVD visualization has shown that absolute distances can matter in this model. In the UMAP case we can observe a loss of shape similar to what happened in the PCA and SVD case. While the structure is congruent when using Euclidean distances, segregated clusters form when representing cosine similarity, with their predominant criterion of division being the hundreds' digit. Using Euclidean distances gives a picture similar to t-SNE, but projected and stretched and with more dispersion for numbers close to zero. The spiral-like conformation is also notable here.

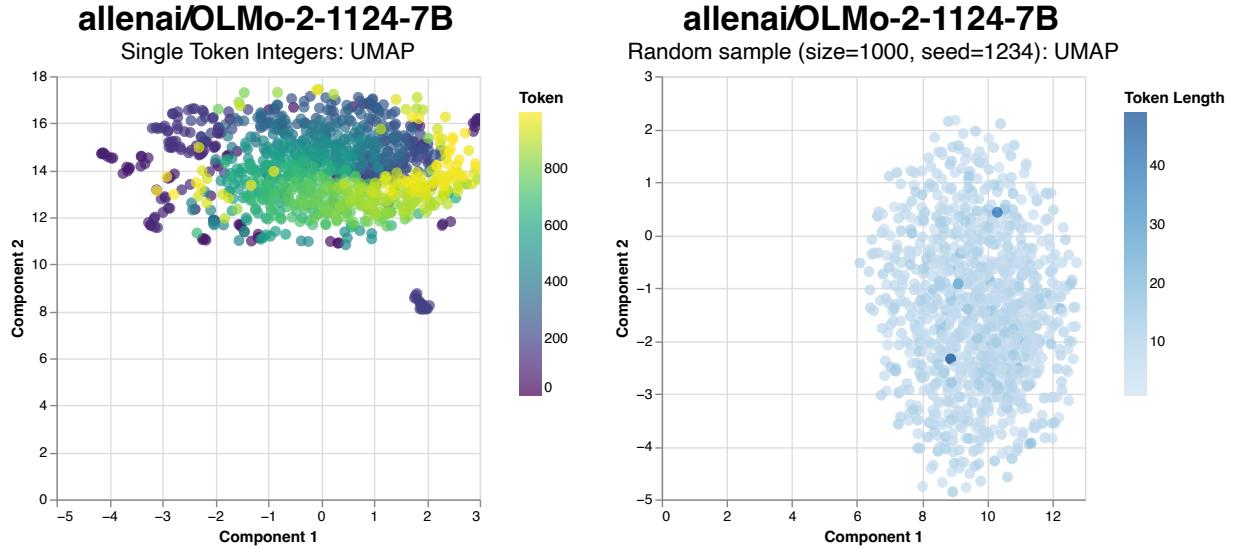


Figure 10: UMAP visualization with Euclidean distance

### Correlation with mathematical properties

dimension	property	encoding	correlation	p_value
514	log	direct	-0.67287	8.4465e-133
3085	even	direct	-0.60990	6.3840e-103
3085	numbers	direct	-0.60990	6.3840e-103
3085	log	direct	-0.60653	1.6467e-101
514	fibonacci	gauss	0.37430	1.3043e-34
2538	fibonacci	gauss	0.35112	2.1919e-30
514	primes	gauss	0.26358	2.3511e-17
695	primes	fourier_cos_t10	-0.22069	1.6980e-12
2538	fibonacci	fourier_cos_t5	-0.18336	5.2002e-09

Table 3: Feature-sequence correlations in OLMo-2.

A lot of features (tbl. 3, fig. 11) correlate very significantly with magnitude, giving confirmation of the semantic connection between the embedding value and the number represented. We find:

- Dimension 514 has a high correlation (0.37) with Fibonacci numbers using Gauss one-hot encoding, and with prime numbers with Gauss one-hot encoding (0.26).
- There are other dimensions are also correlated with Fibonacci and Primes, to a lesser extent, and Fourier encoding seems to show weaker ties than one-hot Gaussian encoding.

Although while these can be useful clues about the relation between the features and the numerical properties they're correlated with, they do not explain by themselves how and why the features are tied to those properties. The correlation with Fibonacci numbers in particular is high enough to not resemble a pure coincidence, especially in such a high-dimensional space. We can speculatively make some hypotheses:

- Features in the embedding might work as hierarchical detectors, with some of them being broad-scope, general detectors of numbers of interest (dimension 514), and others being more specific to certain properties (2538 to Fibonacci numbers).

- The correlation with the features is tied with the geometry of the embedding space. The Fibonacci numbers have ties to spiral structures (in particular golden spirals), which may have a relation to the self-similar structures observed.

Further study would be needed to untangle the relationship between mathematical sequences and features, although having a dimension (514) with such a strong correlation with Fibonacci numbers and such a low p-value doesn't seem dismissible over random chance.

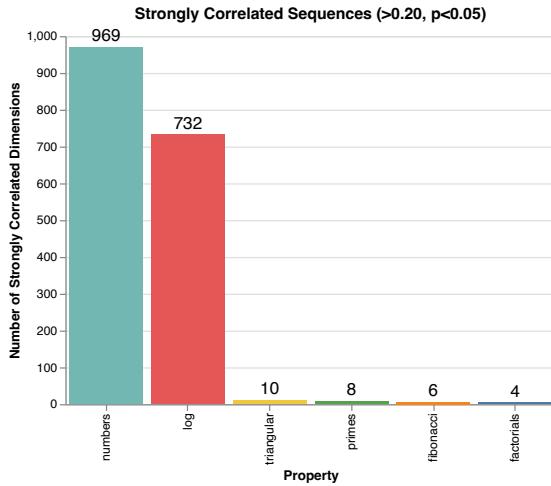


Figure 11: Mathematical sequences against the number of associated strongly correlated embedding dimensions.

## Llama-3.2-1B-Instruct

### Linear analysis

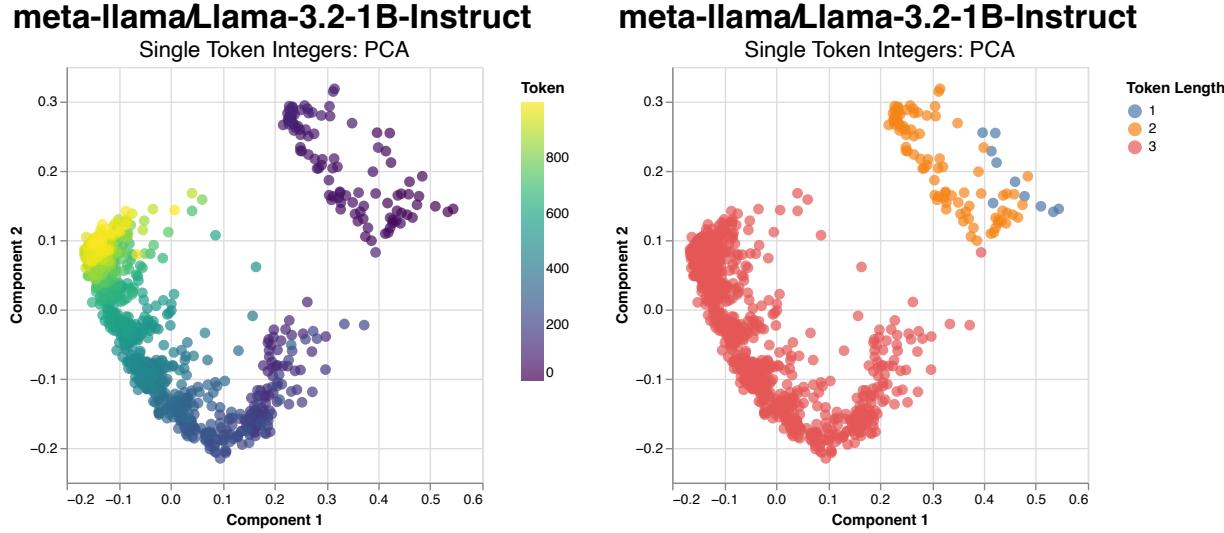


Figure 12: PCA visualization of Llama embeddings.

The LLaMa PCA plot (fig. 12) shows a similar picture to the OLMo one: a curve with smooth, gradual transitions between numeric quantities. The self-similar, recursive structure based on digit count is immediately visible in the PCA plot, as there is a much more striking division and a separation between numbers of different digit size. It is remarkable how the replication of the same structures for different digit counts stays present in different models.

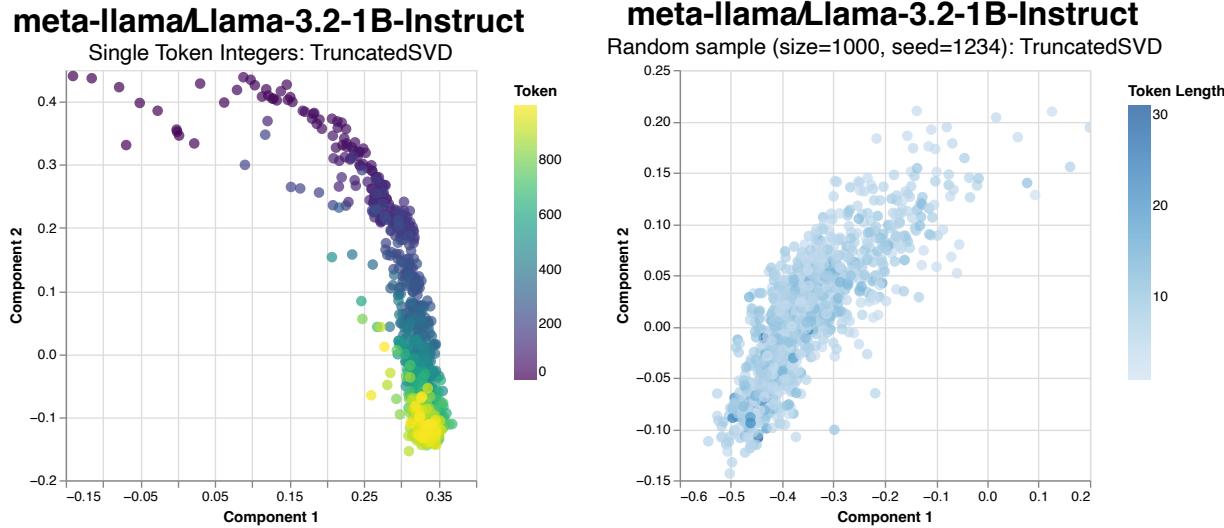


Figure 13: SVD visualization of Llama embeddings compared to random sample.

The SVD plot shows a linear arrangement - numbers form an almost straight diagonal line from small to large values. However, this apparently linear picture changes drastically once we also take into account the third component, showing a much more tridimensional picture, looking like the curves observed in PCA after a diagonal rotation. The 3D projection (fig. 15) really shines here, showing how the same structure is repeating at different distances and angles of rotation.

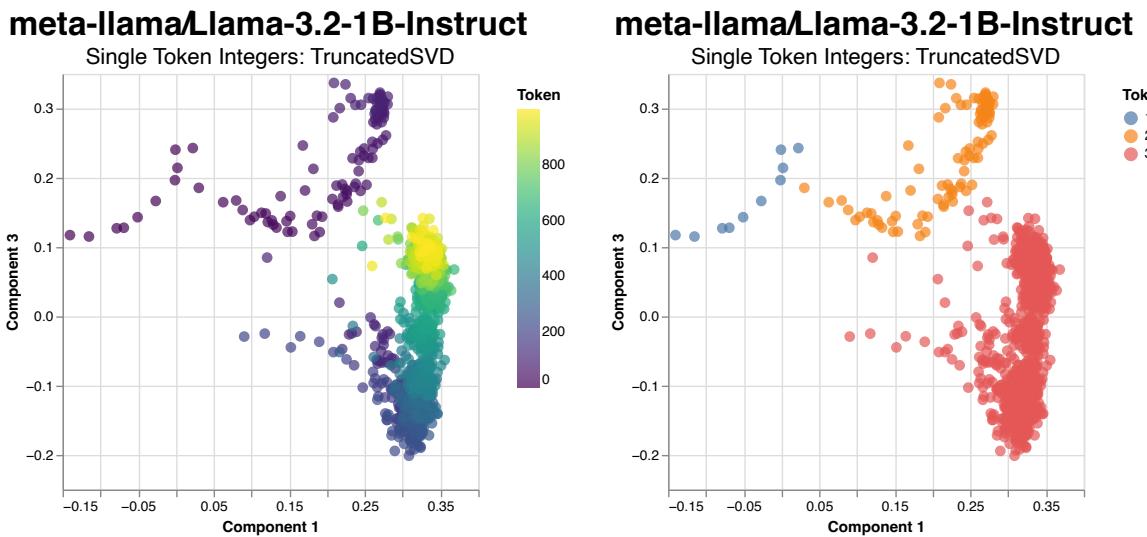


Figure 14: Llama SVD visualizations of first and third component.

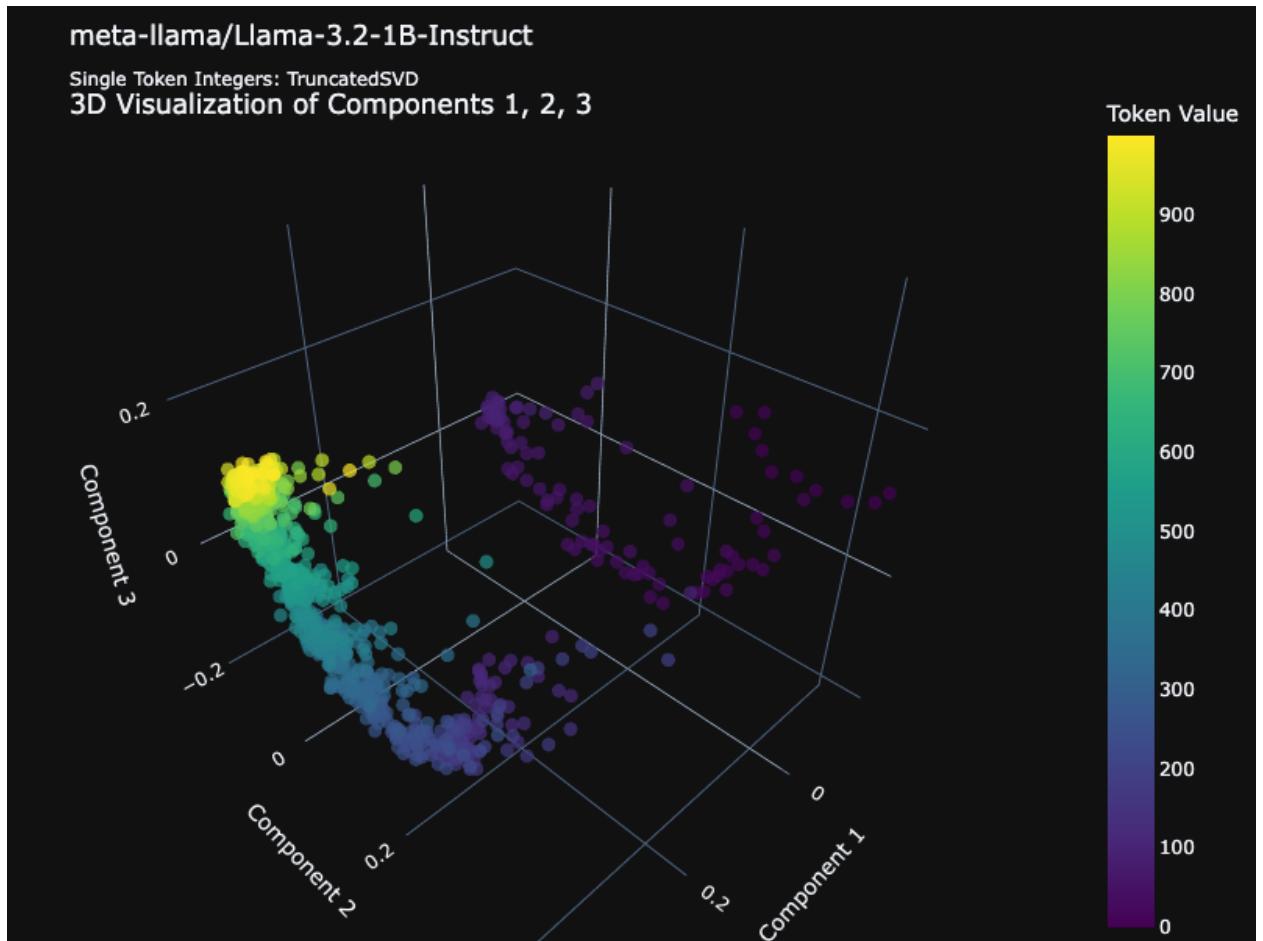


Figure 15: 3D projection of LLaMa 3.2 embeddings after SVD

## Explained variance

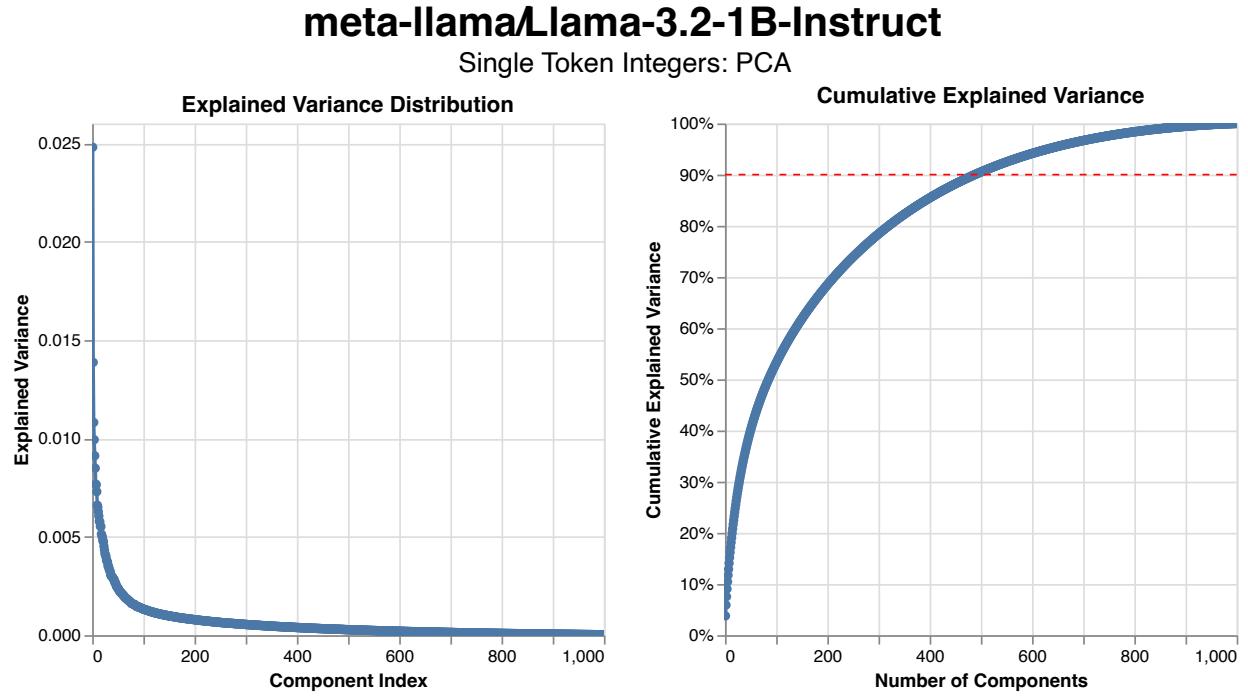


Figure 16: Llama PCA explained variance.

The explained variance plot reveals slightly higher information concentration than OLMo-2. Llama-3.2 reaches 90% explained variance with approximately 500 components compared to OLMo-2's 500 components. This suggests more efficient numerical encoding, despite the smaller model size. A possible reason is the bigger training dataset of LLaMa 3 (Grattafiori et al., 2024), having 15 trillion ingested during training against OLMo's 5 trillion (OLMo et al., 2025) could have lead to different convergence patterns.

## Non-Linear analysis

These nonlinear projections reveal dramatically different organizational patterns from both the linear methods and from OLMo-2's structures.

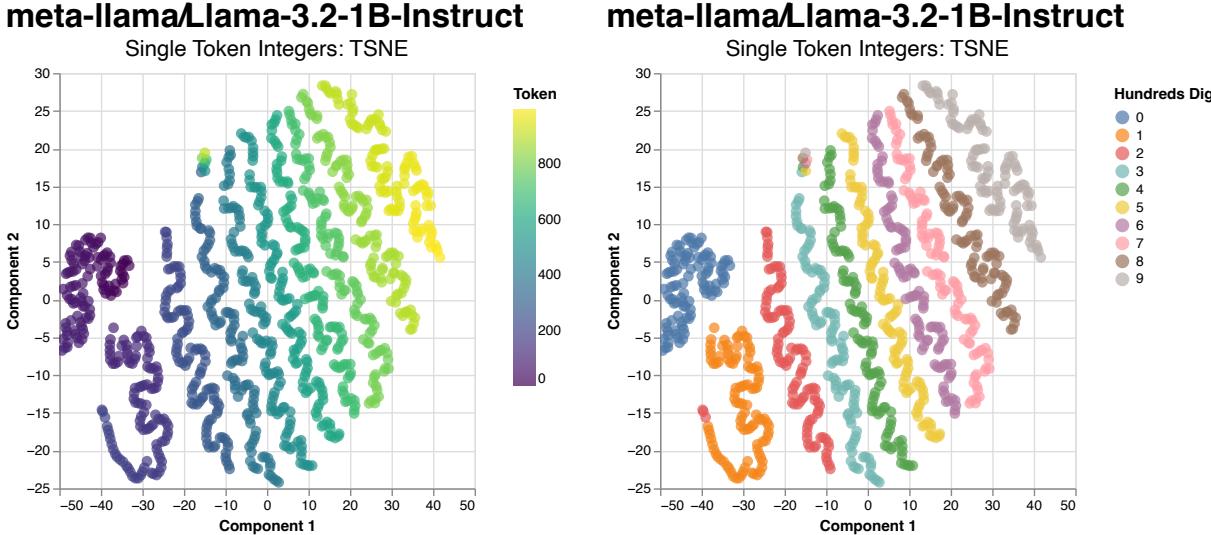


Figure 17: 2D t-SNE structure in Llama

The t-SNE visualization (fig. 17) is very unusual, and show continuous, winding structures that might look like they had been uncoiled or unwound from a higher-dimensional spiral arrangement. The mathematical progression follows these winding paths smoothly, and switching the coloring to highlight the hundreds digit reveals that each filament clusters neatly for its hundreds' group. We can also see another interesting phenomenon: some numbers divisible by 100 cluster between the teal and green filaments.

By applying the procedure in 3D, the helical structures become more visible. Interactive visualizations are helpful in this context, as they give a sense of depth, but still the helices are clearly visible. With this visualization we get singular helices clustered by the hundreds' digit, also due to setting a low perplexity value to let the structures emerge locally.

The UMAP visualizations (figs. 19, 20) resemble OLMo's ones. It's also interesting to see that changing the distance function to Euclidean doesn't have particular effects, unlike the previous OLMo visualization. This can be an indication that the numeric data is more centered towards the mean. Even though the 2D visualization is interesting, the 3D one (fig. 21) complements the picture we started seeing with t-SNE, by giving the idea that beyond having dimensions that represent the spiral structure at a cluster-local level, there might be a more global geometric phenomenon going on, making the whole structure approximate a half-sphere.

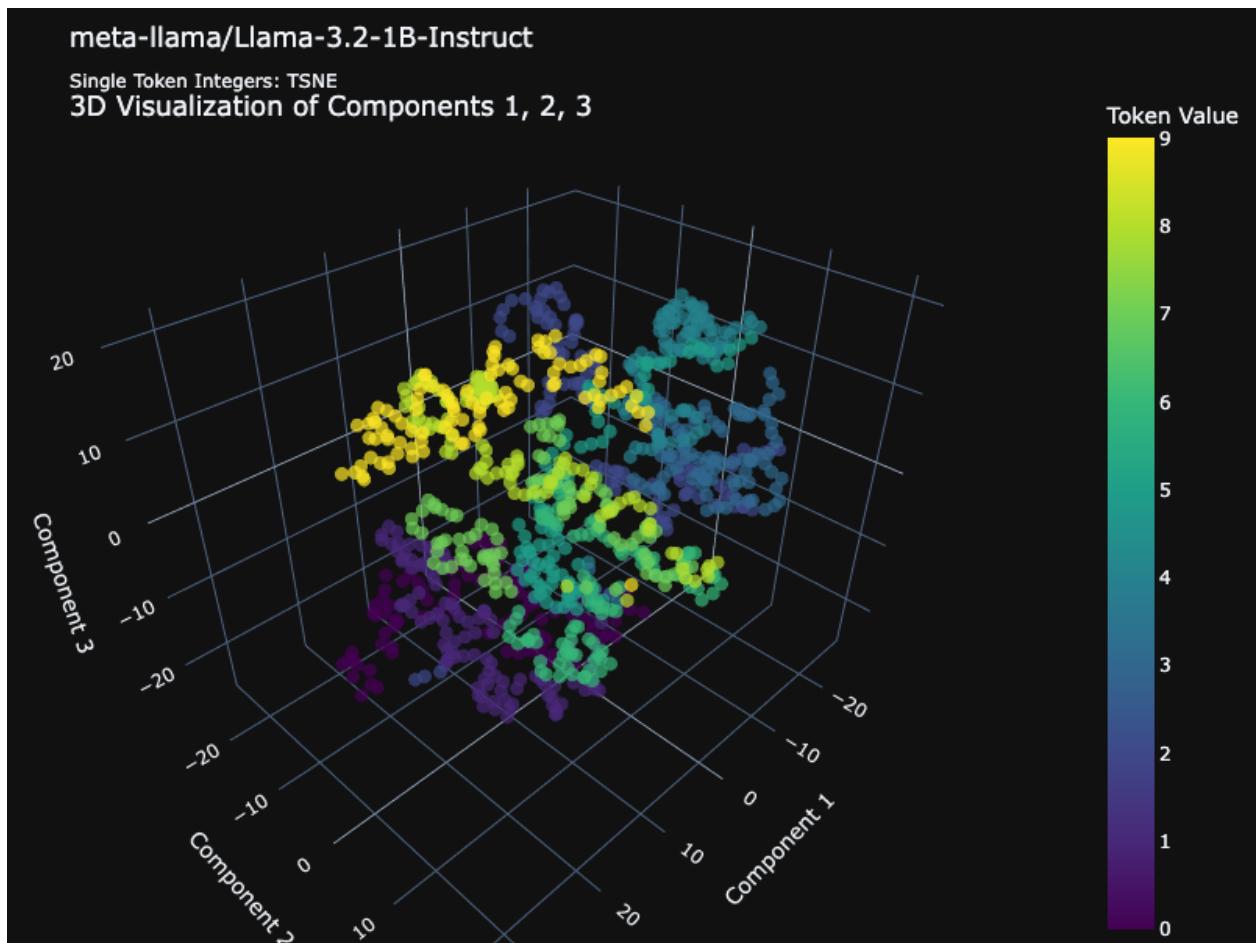


Figure 18: 3D t-SNE structure in Llama

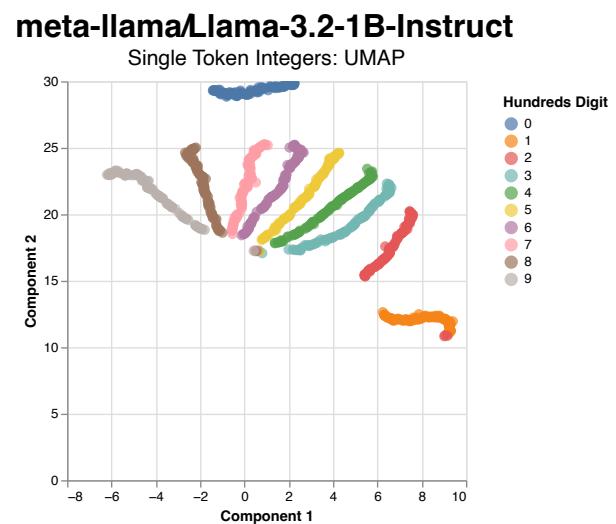
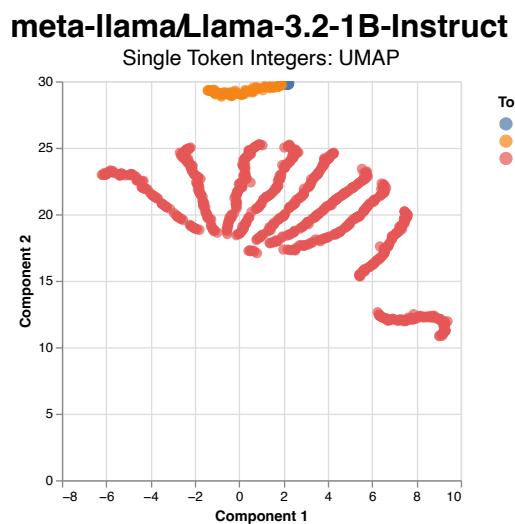
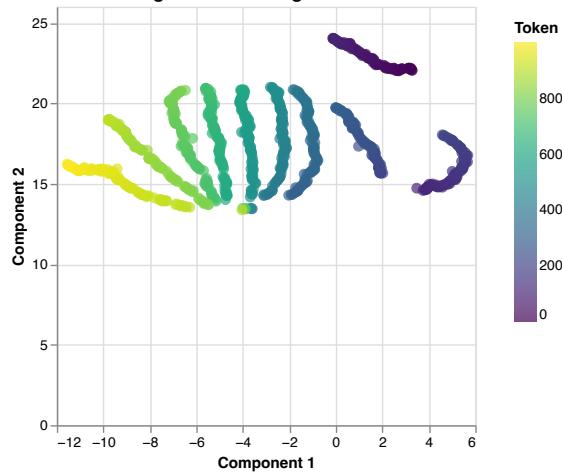


Figure 19: UMAP in Llama with cosine similarity.

### meta-llama/Llama-3.2-1B-Instruct

Single Token Integers: UMAP



### meta-llama/Llama-3.2-1B-Instruct

Random sample (size=1000, seed=1234): UMAP

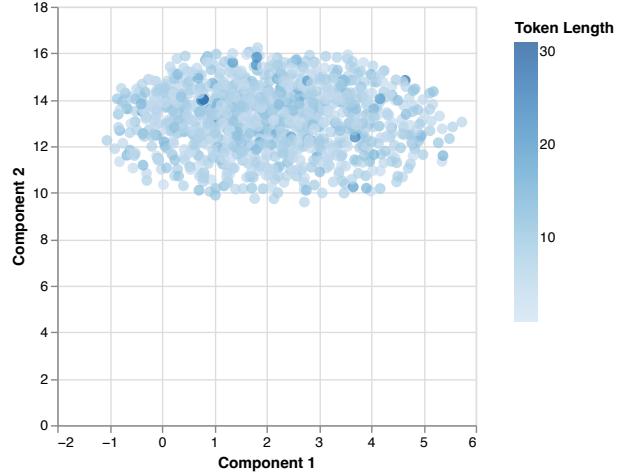


Figure 20: UMAP in Llama with Euclidean distance.

### meta-llama/Llama-3.2-1B-Instruct

Single Token Integers: UMAP

3D Visualization of Components 1, 2, 3

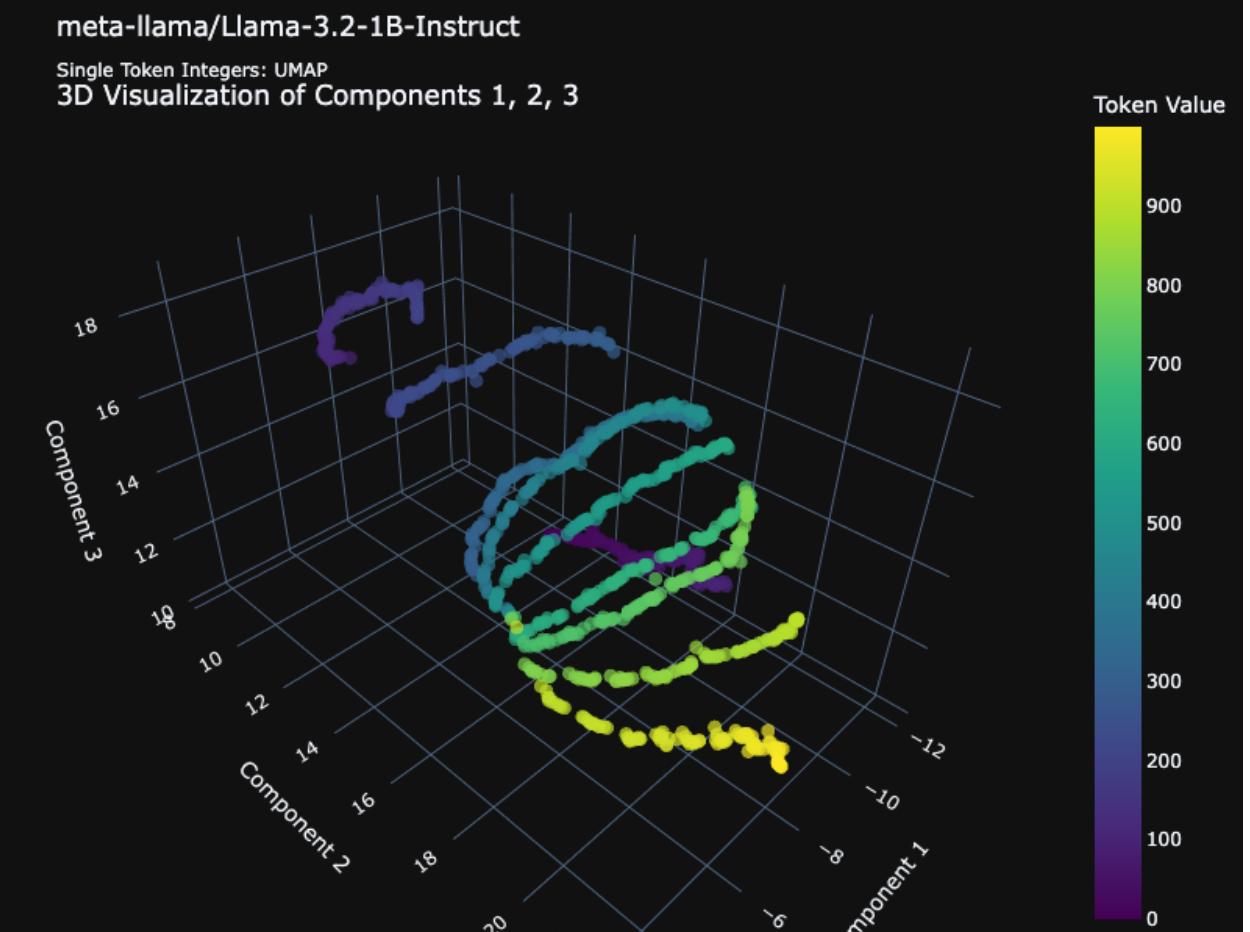


Figure 21: 3D cosine similarity UMAP in Llama.

## Correlation with mathematical properties

dimension	property	encoding	correlation	p_value
417	log	direct	-0.70445	9.1543e-151
1929	log	direct	0.69842	3.7225e-147
1511	log	direct	-0.69462	6.3413e-145
1601	log	direct	0.68892	1.2059e-141
1511	numbers	direct	-0.66539	7.2717e-129
1929	fibonacci	gauss	-0.53600	1.8846e-75
1601	fibonacci	gauss	-0.48667	1.3712e-60
417	fibonacci	gauss	0.46017	1.4848e-53
1447	fibonacci	gauss	-0.40412	1.4218e-40
1929	triangular	gauss	-0.43179	1.1163e-46
1601	triangular	gauss	-0.42298	1.1313e-44
881	primes	gauss	-0.29401	2.1688e-21
665	primes	gauss	0.28418	4.9509e-20

Table 4: Correlations between Llama embedding dimensions and mathematical sequences.

The picture being shown through the correlations is a very surprising one, as it reinforces clearly what was observed previously through the OLMo model. The most correlated properties are magnitude-related, with the logarithm and the plain numbers on top. The maximum absolute correlation with Fibonacci numbers has soared to around 0.5, making a very strong case for these dimensions to either be feature detectors for certain sequences, or to incidentally have connections with the sequences due to intrinsic geometric properties. The case for the second hypothesis is also reinforced by how strikingly we were able to show a very interesting geometric landscape for this model, and how this seemed to coincide with examining a model that was trained on a much larger scale.

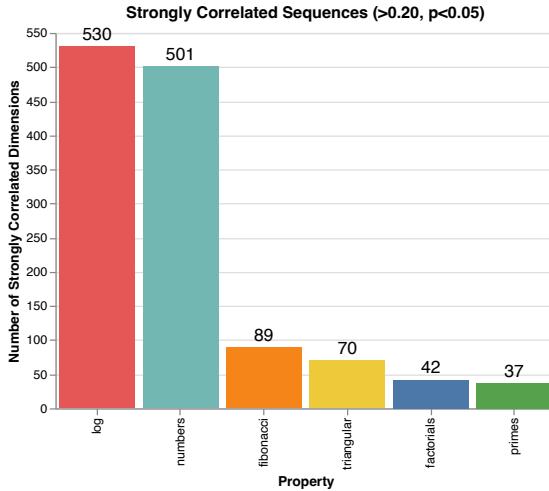


Figure 22: Dimensions strongly correlated with properties in Llama 3.2

Taking a look at how many dimensions are correlated quantitatively (fig. 22), most of them are still related to magnitude, but the number of them that are related to properties by more than .20

# Conclusions

With visualizations and the data, we were able to paint a picture of interesting phenomena coming through the embeddings layers of two distinct LLMs, OLMo and Llama. The most important difference between these two models that we were able to discuss is the scale, and we have shown that with bigger scale come more visible patterns, in both correlations with mathematical sequences and geometric structure.

## Key Findings

### Structured Numerical Embeddings

Our analysis revealed that numerical tokens are not randomly distributed in embedding space but follow organized patterns:

**Geometric Organization:** Principal component analysis showed that numerical embeddings lie on low-dimensional manifolds. OLMo-2 exhibited U-shaped curves with recursive patterns across digit ranges, while Llama-3.2 displayed more linear arrangements. Both models required only ~500 components to capture 90% of variance, indicating substantial dimensionality reduction from the full 4096-dimensional space.

**Mathematical Property Correlations:** Multiple embedding dimensions showed \*significant correlations with mathematical properties:

- Magnitude and digit count exhibited the strongest correlations ( $r > 0.67$ ) - Primality was encoded across multiple dimensions - Fibonacci number proximity showed notable correlations, though potentially influenced by magnitude effects
- Binary properties like evenness were systematically represented

The consistent organization of numerical embeddings according to mathematical properties suggests that neural language models might spontaneously develop structured representations during training. This organization goes beyond simple ordering, incorporating complex mathematical relationships like primality and sequence membership.

## Limitations

The analysis was limited to two models and a specific set of mathematical properties. Further investigation with controlled experiments would be needed to establish causal relationships.

Whether these findings extend to larger models, different architectures, or other mathematical domains remains to be determined. The observed structures may reflect training data properties as much as emergent organizational principles.

## Future Directions

This work suggests several research directions: investigating mathematical representations across model scales and architectures, exploring computed embedding approaches that leverage discovered geometric

structures, and examining whether similar organizational principles apply to other domains where specialized representations might be beneficial.

The findings contribute to understanding how neural language models organize mathematical information and suggest that structured representations may emerge naturally in systems trained on numerical data. While modest in scope, these results provide a foundation for further investigation into the relationship between representational structure and mathematical reasoning capabilities in artificial systems.

# Bibliography

- Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbing, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J. S., Jain, C., Weber, A. A., Jurkschat, L., Abdelwahab, H., John, C., Suarez, P. O., Ostendorff, M., Weinbach, S., Sifa, R., ... Flores-Herr, N. (2024, March 17). *Tokenizer Choice For LLM Training: Negligible or Crucial?* <https://doi.org/10.48550/arXiv.2310.08754>
- Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanusse, F., McCabe, M., Ohana, R., Parker, L., Blancard, B. R.-S., Tesileanu, T., Cho, K., & Ho, S. (2023, October 4). *xVal: A Continuous Number Encoding for Large Language Models.* <https://doi.org/10.48550/arXiv.2310.02989>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024, November 23). *The Llama 3 Herd of Models.* <https://doi.org/10.48550/arXiv.2407.21783>
- Huh, M., Cheung, B., Wang, T., & Isola, P. (2024, May 13). *The Platonic Representation Hypothesis.* <https://doi.org/10.48550/arXiv.2405.07987>
- Kantamneni, S., & Tegmark, M. (2025, February 2). *Language Models Use Trigonometry to Do Addition.* <https://doi.org/10.48550/arXiv.2502.00873>
- Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605. <http://jmlr.org/papers/v9/vandermaaten08a.html>
- McInnes, L., Healy, J., & Melville, J. (2020, September 18). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* <https://doi.org/10.48550/arXiv.1802.03426>
- McLeish, S., Bansal, A., Stein, A., Jain, N., Kirchenbauer, J., Bartoldson, B. R., Kailkhura, B., Bhatele, A., Geiping, J., Schwarzschild, A., & Goldstein, T. (2024, December 23). *Transformers Can Do Arithmetic with the Right Embeddings.* <https://doi.org/10.48550/arXiv.2405.17399>
- Millidge, B. (2023, February 4). *Integer tokenization is insane.* <http://www.beren.io/2023-02-04-Integer-tokenization-is-insane/>
- Millidge, B. (2024, July 7). *Right to Left (R2L) Integer Tokenization.* <http://www.beren.io/2024-07-07-Right-to-Left-Integer-Tokenization/>
- Murray, A. L. (2010). Can the existence of highly accessible concrete representations explain savant skills? Some insights from synesthesia. *Medical Hypotheses*, 74(6), 1006–1012. <https://doi.org/10.1016/j.mehy.2010.01.014>
- OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., ... Hajishirzi, H. (2025, January 15). *2 OLMo 2 Furious.* <https://doi.org/10.48550/arXiv.2501.00656>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.* <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>
- Singh, A. K., & Strouse, D. J. (2024, February 22). *Tokenization counts: The impact of tokenization on arithmetic in frontier LLMs.* <https://doi.org/10.48550/arXiv.2402.14903>
- Skalise, J. (2023). *Some Arguments Against Strong Scaling.* <https://www.lesswrong.com/posts/DvCLEkr9pXLnWikB8/some-arguments-against-strong-scaling>
- Sutton, R. (2019, March 13). *The Bitter Lesson.* <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 1). *Attention Is All You Need.* <http://arxiv.org/abs/1706.03762>

