UNIVERSITÀ
DI TORINO

Università degli Studi di Torino

Corso di Laurea Magistrale in Informatica

Concrete Numeric Representations in LLM Embeddings

Tesi di Laurea

Relatore/Relatrice
Prof. Di Caro Luigi

Correlatore/Correlatrice
Dr. Torrielli Federico

Candidato/a
Gentiletti Emanuele
900831

Anno Accademico 2024/2025

Contents

Introduction

This work started with a simple premise: why are LLMs bad at math?

This is not really a hard question to answer. Most of the LLMs to date are not built with that purpose in mind, and can rely on tool calling to give good answers to quantitative and numerical questions.

There is a tremendous investment in computing resources that is directed towards arithmetic operations that make up the inner workings of LLMs, computations that the LLMs themselves aren't capable of leveraging to answer arithmetic questions. It feels like witnessing a fundamental disconnection, where the LLM is segregated from the capabilities that make its own functioning possible.

Savant syndrome is a very rare disorder. It manifests primarily in people with autism spectrum disorders (Murray, 2010) or after traumatic episodes. The people affected by it possess extraordinary qualities in certain areas, like arts, music or mathematics, while usually showing significant impairment in others. One of the possible areas in which savants may show exceptional aptitude is calculation: calendrical savants are able to instantly know the day of the week of dates far in the future. These skills are unlikely to be the product of algorithmic calculation (Cowan & Frith, 2009), so alternative hypotheses emerged.

What I propose here is that the Savant condition can be seen as a parallel to the bridging of this capabilities gap in LLMs. In particular, what is taken in consideration here is the use of concrete representations as described in (Murray, 2010), where abstract numerical concepts are transformed into "highly accessible concrete representations" that can be directly manipulated rather than computed through algorithmic steps. This reification process - the conversion of abstract concepts into concrete entities - appears to provide savants with immediate access to numerical relationships that would otherwise require complex calculations.

This is not meant necessarily to give a comprehensive explanation of the phenomenon on an empirical basis, as that would be hard to establish from the basis of current knowledge about both savant cognition and neural network representations. Rather, it serves as a conceptual framework for exploring whether similar representational advantages can be induced in artificial systems.

This idea is explored in two ways:

- by a literature review, that is meant to clarify what can function as concrete representations in this context
- by an exploration of numerical embeddings, that is meant to show whether the learned representation of current language models already tends to conform to certain geometrical objects or structures. We show that there is remarkable structure and patterns in the learned representation of current LLMs.

The Transformer architecture and vector representations

The inductive bias of Tokenization

Modern LLMs are built on the Transformer architecture (Vaswani et al., 2023), which operates by converting input text into sequences of discrete tokens that are then mapped to high-dimensional vector representations. This initial tokenization 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 the numerical data to arithmetical tasks.

The most used algorithm for tokenization is currently Byte-Pair Encoding, 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 $'1234' \rightarrow [1, 2, 3, 4]$), 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, 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 a lesser arithmetic performance, as the grouping doesn't match the positional system's <way of calculating?>. 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). 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 clustering scheme is corrected.

There can be different hypotheses on why this might be, 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.
- 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)

Still, we are left with the fact that the learned representations work better for a tokenization strategy different from the one the model was trained for. 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) lead to embeddings that maintain that bias even when learned in a L2R fashion. This can be a

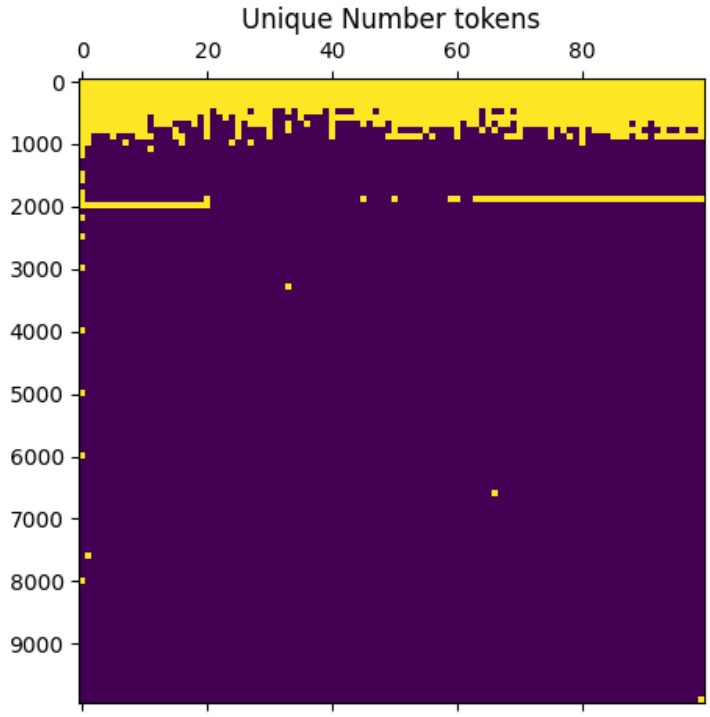


Figure 1: GPT-2 number tokenization. Each row represents 100 numbers, yellow squares mean that the number is represented by a single token, purple ones by multiple (Millidge, 2023)

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.

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

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-3.5/4	L2R chunks of 3 digits
Claude 3/4	R2L chunks of 3 digits

Reification as computed embeddings - xVal

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. Numerical values represented by a single embedding vector associated with the [NUM] special token.

This fits very well with the idea of reification: the embedding is no longer just a representation, but it contains and has properties of the object it represents.

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. This work has been inspired by the xVal paper, with one of its initial goals being to find good representations for computed numerical embeddings.

The search for better suited representation

A case study of a Savant patient, DT (Murray, 2010), has been reported of having a mathematical landscape with the following characteristics:

- Has sequence-space synesthesia with a “mathematical landscape” containing numbers 0-9999
- Each number has specific colors, textures, sizes, and sometimes movements or sounds
- Prime numbers have special 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 visual structure for DT

Sequence-space synesthesia consists in the visualization of certain sequences in physical space.

Embeddings Analysis

The analytic part of this work consists in the search for structure in LLM numerical embeddings.

As stated previously, recent open source models mostly employ an L2R tokenization scheme. There are no large scale open source models using R2L tokenization as of the time of writing, but the improvement in performance observed when using R2L tokenization in L2R-trained models could be a hint that L2R embedding representations still have similar properties to the R2L ones.

We're looking for clues of mathematical properties being encoded in the embeddings. As the results show, we find strong correlations between embedding dimensions and mathematical properties, the strongest one being magnitude (data) and a very interesting one being the distance between the number and their closest Fibonacci number.

Methodology

For each model, we extracted the embeddings corresponding to integer numbers representable by a single token. The embeddings were then analyzed using dimensionality reduction techniques (PCA, SVD, t-SNE and UMAP) to garner statistics and produce visualizations that could potentially highlight structures in the data. As a final stage, the embeddings were directly tested for correlations with mathematical properties of the number:

- magnitude
- relation to digit count
- primality
- evenness
- being a perfect square
- being a Fibonacci number

For binary properties, the correlation was measured with vectors with value 1 for the indexes corresponding to numbers where the property is true and 0 elsewhere. To account for the continuous nature of the embeddings, smoother functions that might relate to the same properties were also taken into account:

- "squareness": an approximate measure of closeness of the number to a perfect square

$$\text{squareness}(n) = \begin{cases} 0 & \text{if } n \leq 0 \\ 1 - 2 \cdot \min\left(\sqrt{n} - \lfloor\sqrt{n}\rfloor,\ \lceil\sqrt{n}\rceil - \sqrt{n}\right) & \text{if } n > 0 \end{cases}$$

- prime proximity: distance between the number and the nearest perfect prime, measured in integers between the two.

- Fibonacci proximity: distance between the number and the nearest Fibonacci number, measured in integers between the two.

OLMo-2-1124-7B

Linear analysis

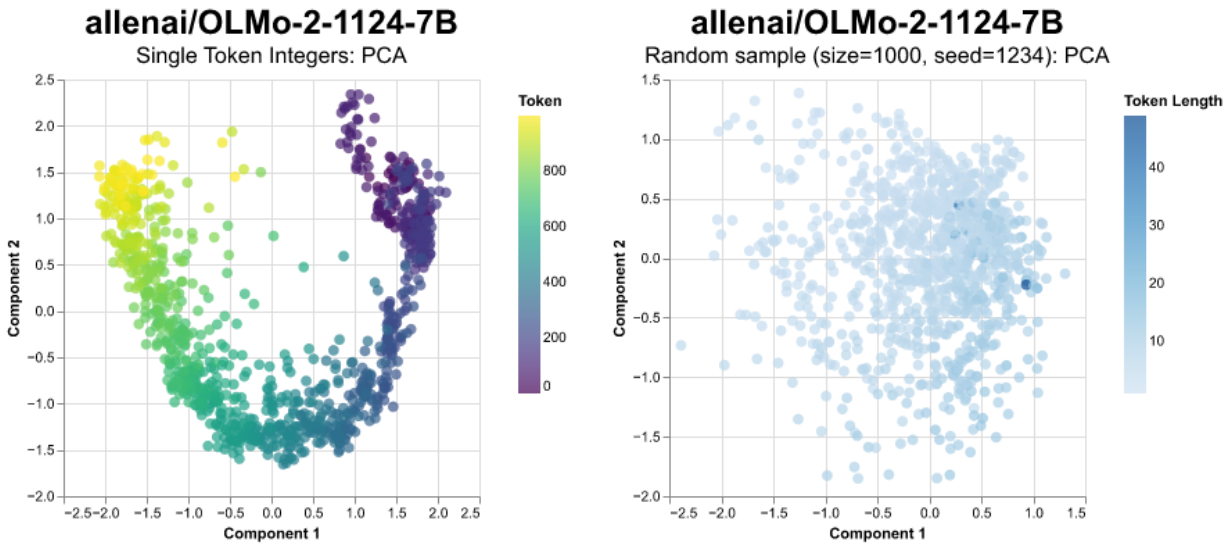


Figure 2: Principal components 1 and 2 of the OLMo model. Random embeddings sample for comparison.

Projecting the numerical token embeddings onto the first two principal components reveals a U-shaped curve. This structure constitutes a one-dimensional manifold embedded within the two-dimensional principal component space.

The manifold structure is particularly significant because it demonstrates that numerical tokens do not occupy the embedding space randomly. Instead, 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.

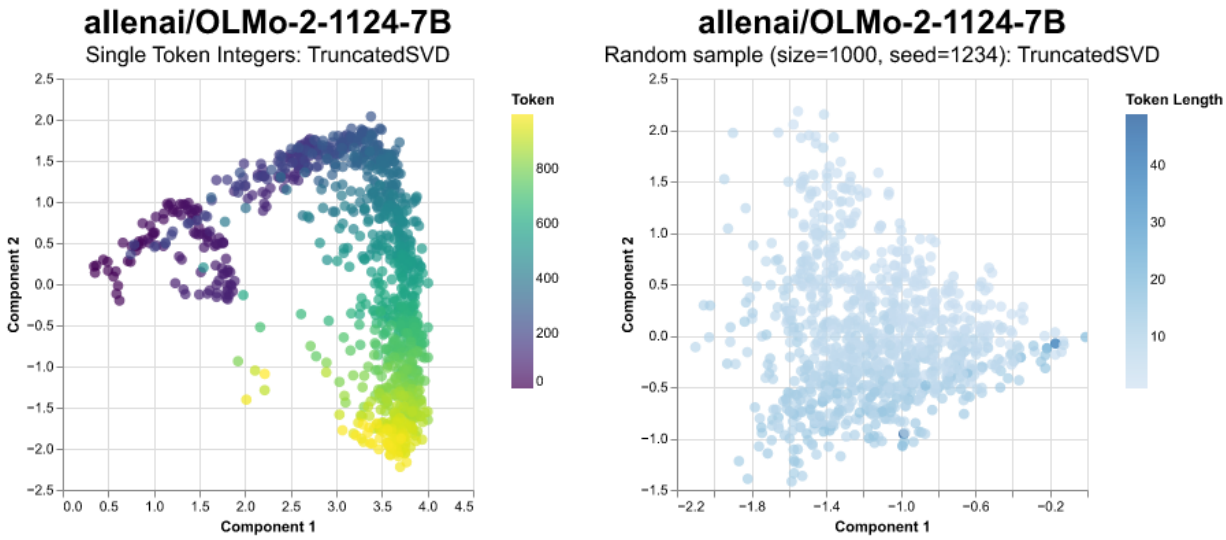


Figure 3: SVD for the two main components of the OLMo model, with random embeddings sample for comparison

The SVD visualization, lacking the data centering done in the PCA, shows a much more consistent geometric structure, suggesting that the encoding of information might be done in absolute distances rather than just with relative positioning between data points. There is also a very notable recursive, fractal structure, repeating itself for numbers with one, two and three digits.

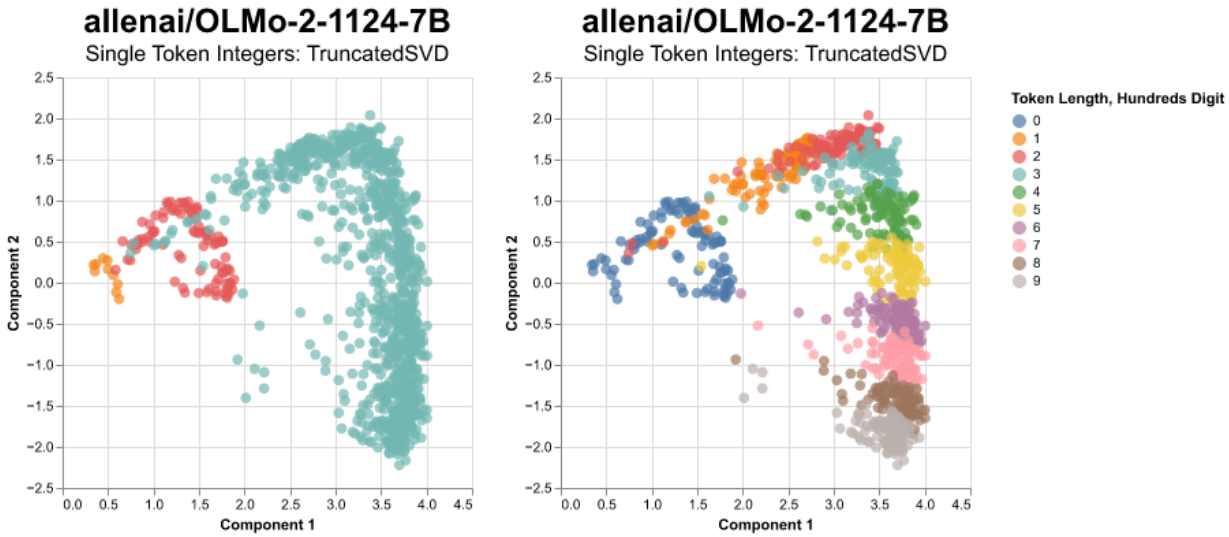
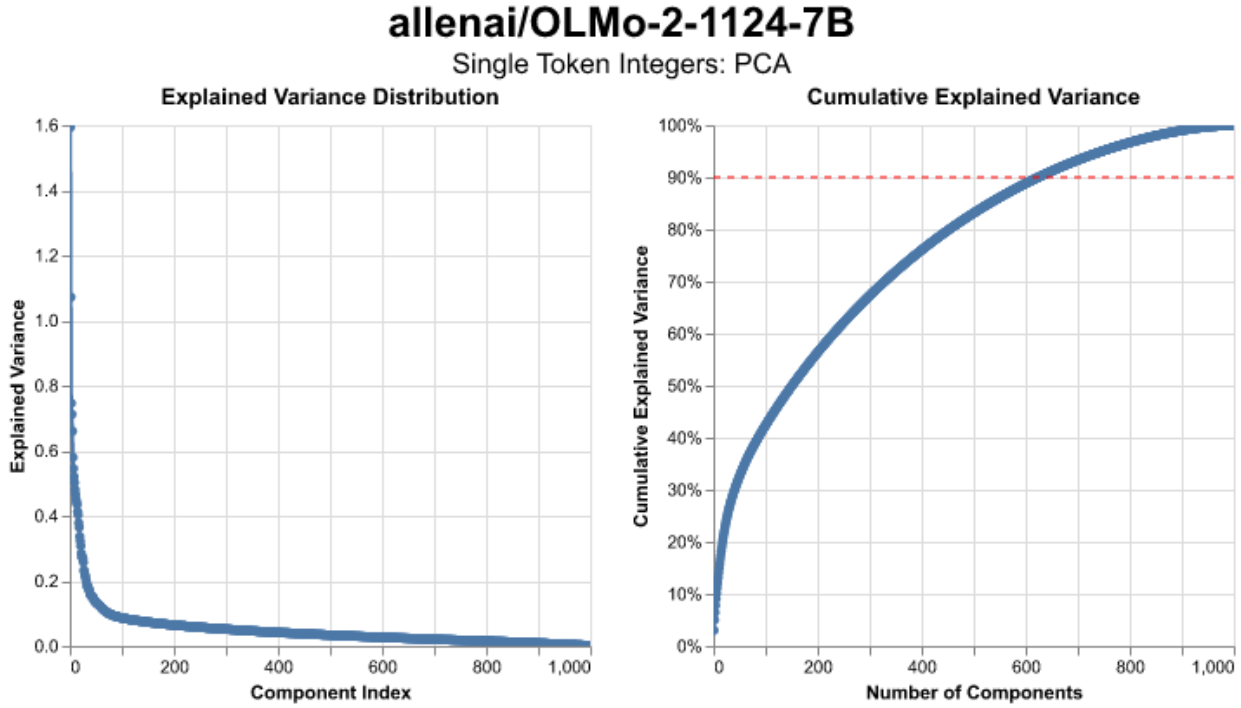


Figure 4: SVD coloring done by digit length and hundreds digit, highlighting the clustering properties of the embeddings.

Explained variance



The explained variance by component plot (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 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.

Non-linear analysis

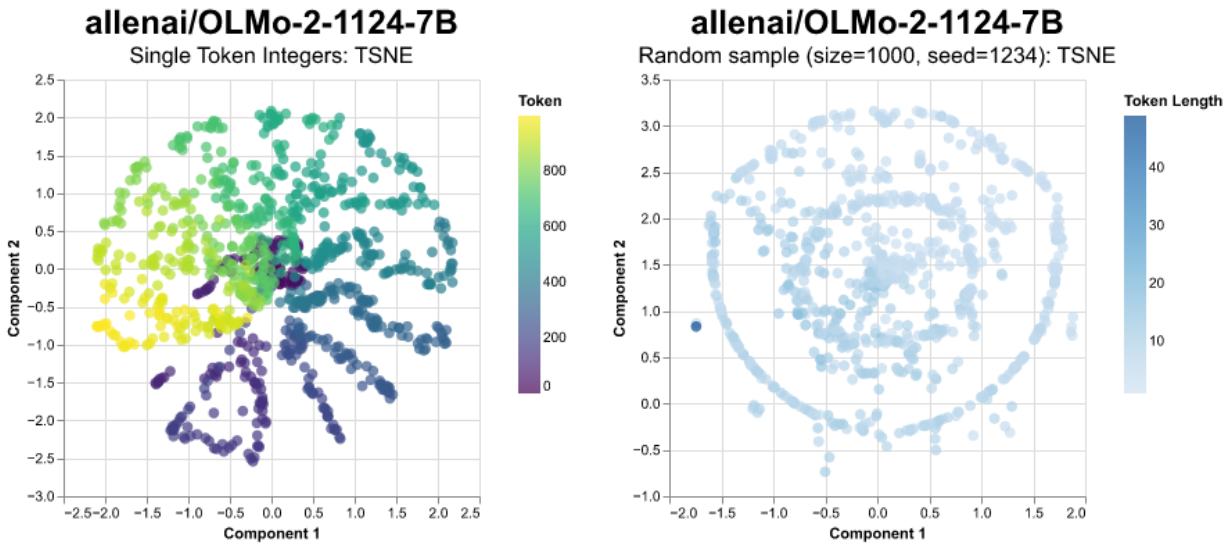


Figure 5: t-SNE visualization for OLMo embeddings.

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

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

The t-SNE visualization 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 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.

UMAP has been run using both Euclidean and cosine distances, 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.

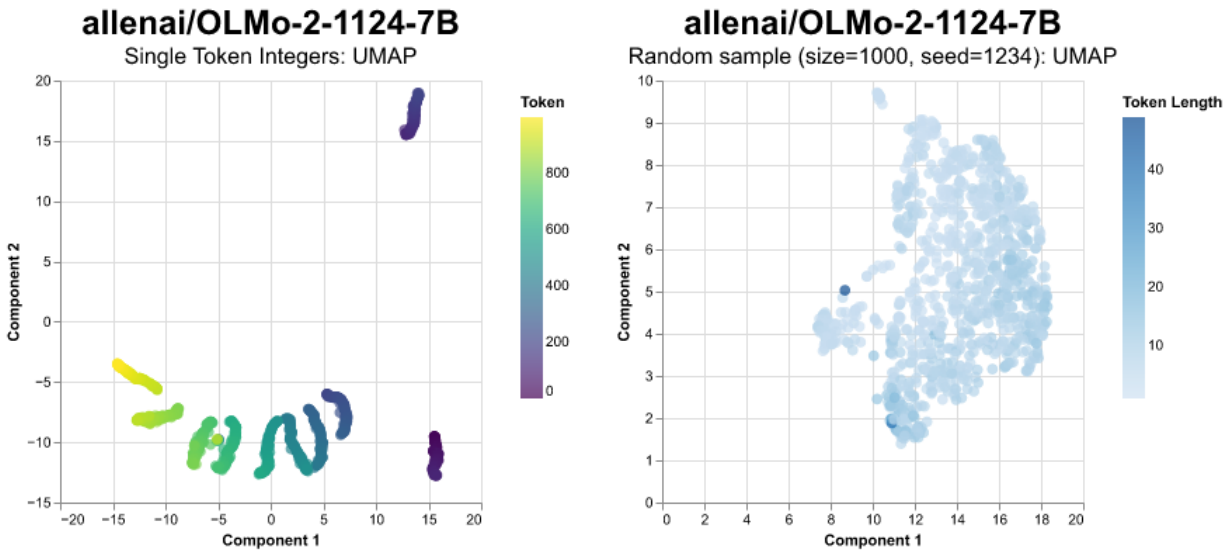


Figure 6: UMAP visualization with cosine distance

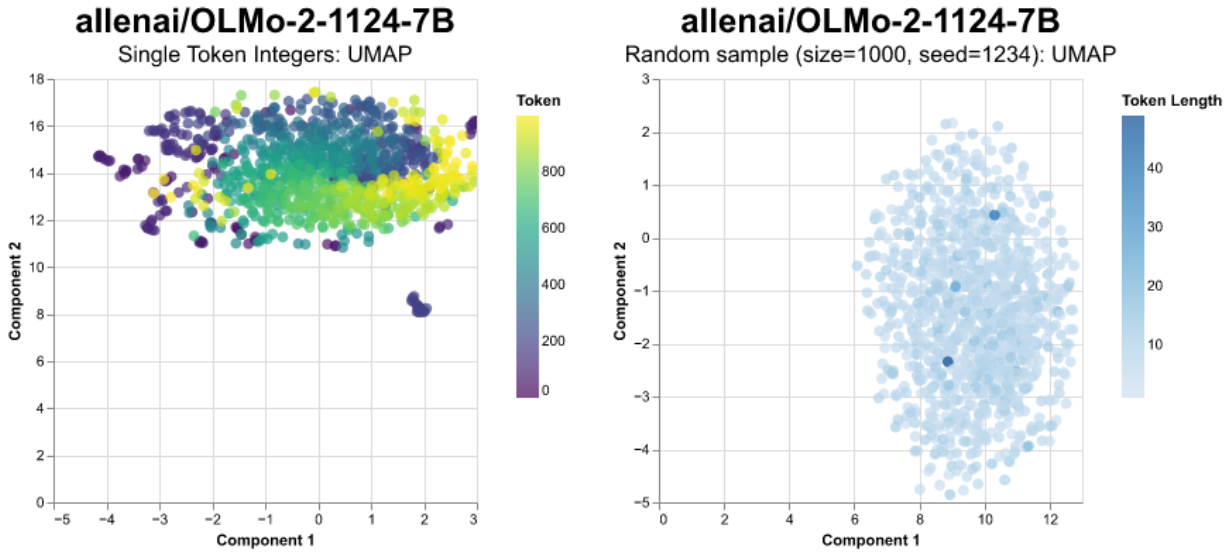


Figure 7: UMAP visualization with Euclidean distance

Llama-3.2-1B-Instruct

Linear analysis

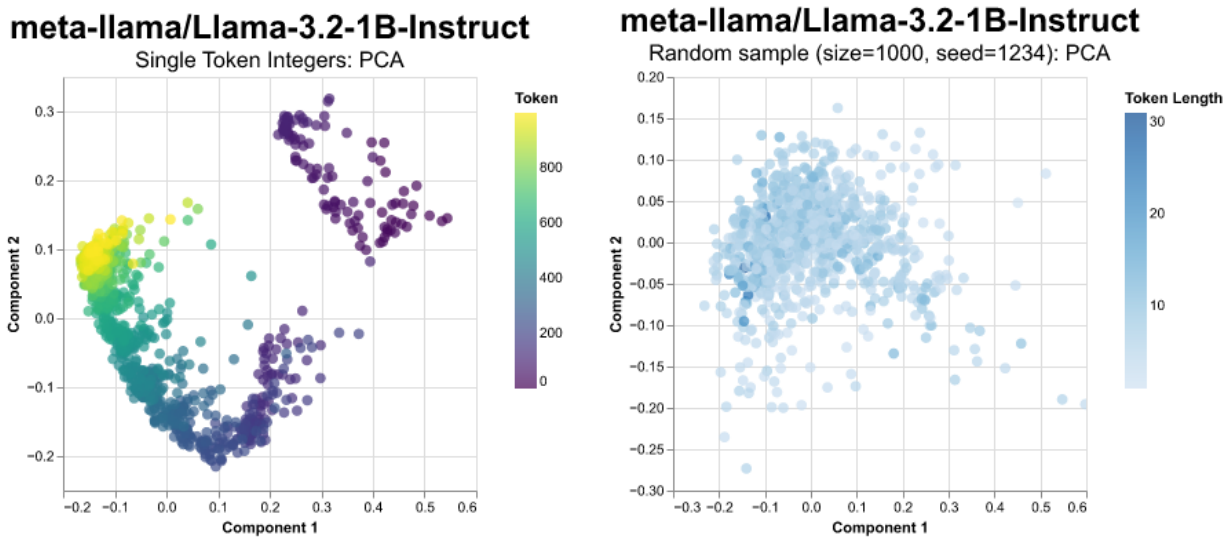


Figure 8: PCA visualization of Llama embeddings.

The PCA projection shows a continuous, arc-shaped curved manifold, with smoother transitions between numbers and a distinct separation with numbers close to 0. As with what was seen with OLMo, it looks like the PCA centering might end up destroying geometric relationships that are better preserved in the SVD visualizations.

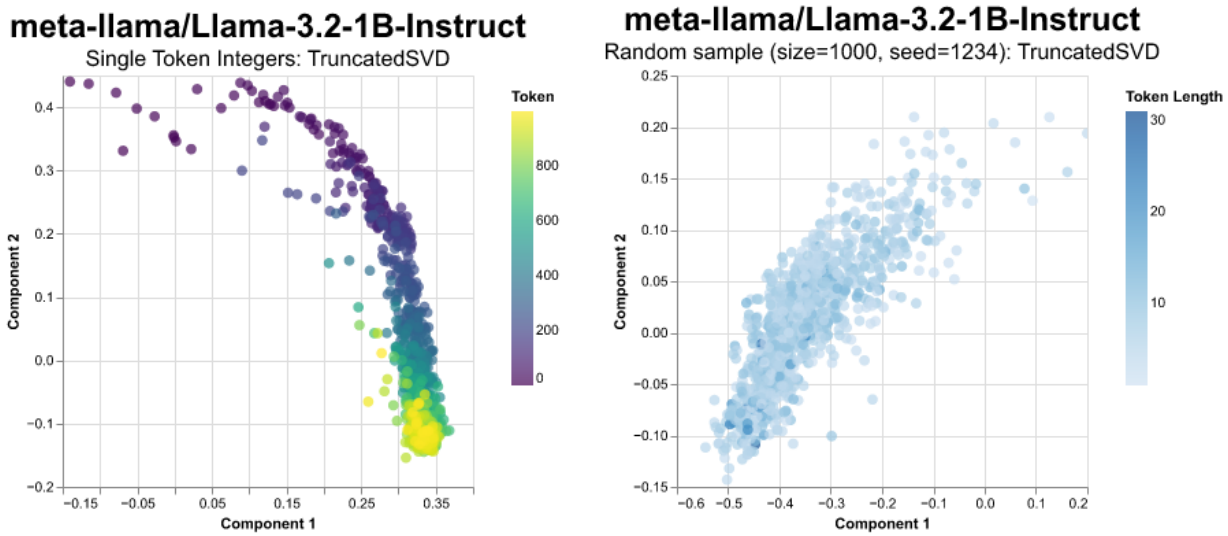


Figure 9: SVD visualization of Llama embeddings

The SVD plot shows a remarkably linear arrangement - numbers form an almost straight diagonal line from small (yellow) to large (purple) values. This linear structure is much more pronounced than OLMo-2's curved SVD patterns, and it is a unique shape rather than a recursive, recurring pattern.

The digit-based coloring reveals clear but subtle clustering by mathematical properties. Unlike OLMo-2's

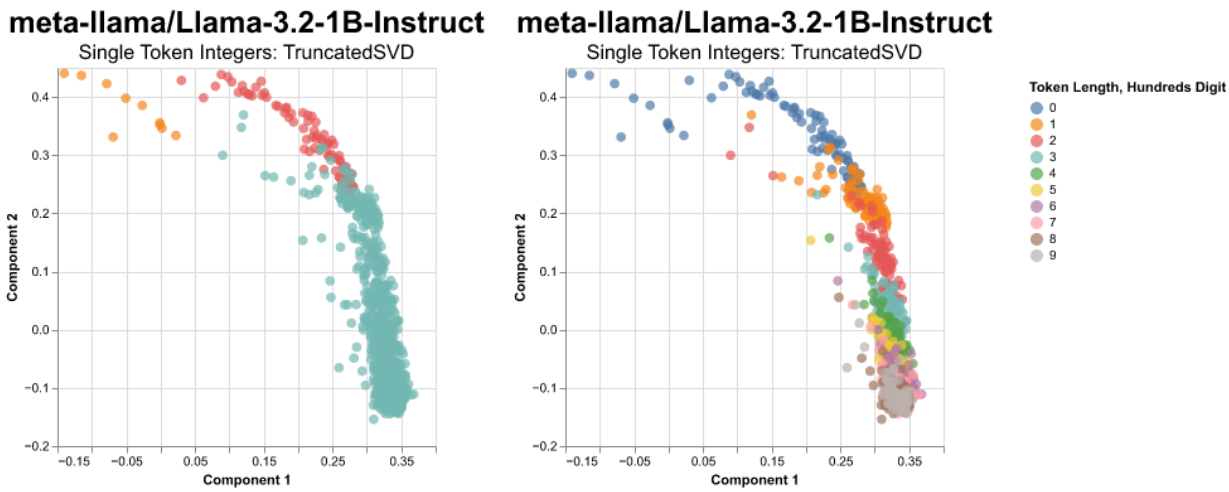


Figure 10: Llama SVD visualization by digit

distinct spatial territories, Llama-3.2 shows gradual transitions along the linear arrangement while maintaining digit-based organization patterns.

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 in the smaller model.

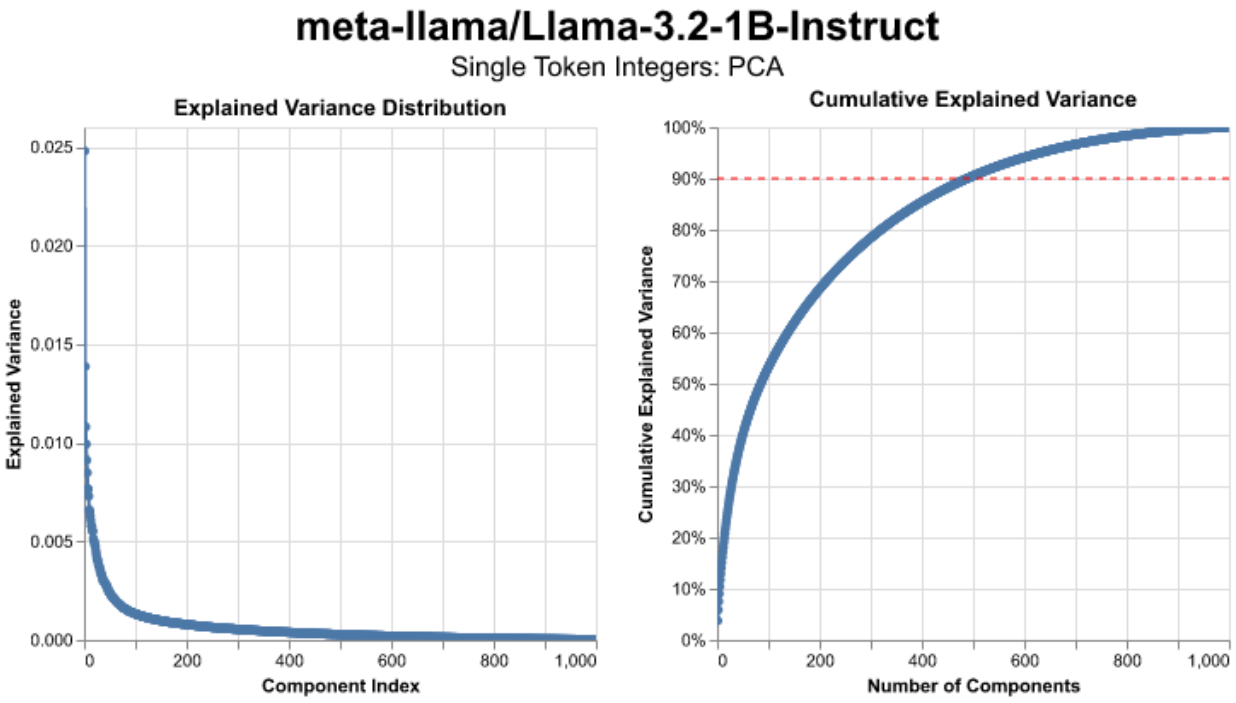


Figure 11: Llama PCA explained variance.

Mathematical Property Detection in OLMo-2 Embeddings: Analysis Report

2.3 Analysis Techniques

- **Correlation analysis**: Pearson correlation between embedding dimensions and mathematical properties
- **Dimensionality reduction**: UMAP, t-SNE, PCA, and TruncatedSVD for visualization
- **Statistical significance testing**: P-values computed for all correlations

3. Results

3.1 Dimensionality Reduction Analysis

Visualization through multiple dimensionality reduction techniques revealed consistent organizational patterns:

- **Token length clustering**: Numbers grouped primarily by digit count (1-digit, 2-digit, 3-digit)
- **Hierarchical structure**: Within length clusters, sub-structures emerged based on individual digit properties
- **Smooth transitions**: Continuous neighborhoods suggest interpolation capabilities between similar numbers

The consistency across different reduction methods (UMAP, t-SNE, PCA, TruncatedSVD) indicates robust underlying structure rather than technique-specific artifacts.

3.2 Correlation Analysis Results

3.2.1 Magnitude and Scale Detection

The strongest correlations observed were with magnitude-related properties:

Dimension	Property	Correlation	P-Value	Interpretation
514	magnitude	-0.673	8.4×10^{-133}	Strong magnitude detector
3085	magnitude	-0.607	1.6×10^{-101}	Secondary magnitude detector
2538	magnitude	-0.567	3.0×10^{-86}	Tertiary magnitude detector

The negative correlations indicate these dimensions activate more strongly for smaller numbers, which aligns with frequency distributions in natural language (smaller numbers appear more often in text).

3.2.2 Structural Property Detection

Digit count showed the next strongest correlations:

Dimension	Property	Correlation	P-Value
514	digit_count	-0.485	5.3×10^{-60}
1012	digit_count	-0.463	3.3×10^{-54}
2538	digit_count	-0.454	4.1×10^{-52}

3.2.3 Fibonacci Sequence Detection

Notably, specific dimensions showed significant correlation with Fibonacci proximity:

Dimension	Property	Correlation	P-Value
3085	fibonacci_proximity	0.410	8.6×10^{-42}
1318	fibonacci_proximity	0.383	2.2×10^{-36}

This finding is particularly interesting as Fibonacci sequences were not explicitly present in the model's training objective.

3.2.4 Prime Number Detection

Earlier analysis with binary prime detection revealed weaker but statistically significant correlations:

Dimension	Property	Correlation	P-Value
3278	is_prime	0.200	1.6×10^{-10}
1266	is_prime	-0.200	1.6×10^{-10}

Multiple dimensions showed both positive and negative correlations with primality, suggesting ensemble-based detection mechanisms.

3.3 Multi-Modal Mathematical Dimensions

Several dimensions showed significant correlations with multiple mathematical properties:

Dimension 3085: - Magnitude: r = -0.607 - Fibonacci proximity: r = 0.410 - Digit count: r = -0.452

This suggests some embedding dimensions function as multi-modal mathematical property detectors rather than single-property specialists.

3.4 Visualization Analysis

The Fibonacci detector analysis (Dimension 3085) revealed:

- **Linear relationship**: Clear positive correlation between Fibonacci proximity and dimension activation
- **Bimodal distribution**: Distinct separation between high and low Fibonacci proximity groups
- **Nonlinear response curve**: Gradual activation changes rather than binary on/off responses
- **Actual Fibonacci number clustering**: True Fibonacci numbers (0,1,1,2,3,5,8,13…) clustered at high activation values

4. Discussion

4.1 Evidence for Specialized Mathematical Representations

The analysis provides evidence that OLMo-2 has developed specialized mathematical property detectors within its embedding space. Key observations supporting this conclusion:

1. **Statistical significance**: P-values as low as 10^{-133} indicate these correlations are not due to chance
2. **Multiple independent detectors**: Different dimensions specialize in different mathematical properties
3. **Hierarchical organization**: From basic properties (magnitude) to complex patterns (Fibonacci sequences)
4. **Consistent visualization patterns**: Multiple dimensionality reduction techniques show similar organizational structures

4.2 Comparison to Mathematical Cognition Research

These findings show interesting parallels to research on mathematical cognition:

- **Magnitude representation**: The strong magnitude correlations align with research on the "number sense" and logarithmic number representation in human cognition
- **Specialized modules**: The existence of property-specific dimensions resembles theories of domain-specific cognitive modules
- **Hierarchical processing**: The organization from basic (magnitude) to complex (Fibonacci) properties mirrors developmental theories of mathematical understanding

However, we note these are structural similarities rather than direct evidence of identical mechanisms.

4.3 Implications for Model Enhancement

The identification of specialized mathematical dimensions suggests potential enhancement strategies:

1. **Targeted amplification**: Specific dimensions could be enhanced during fine-tuning for mathematical tasks
2. **Ensemble utilization**: Multiple property detectors could be combined for more robust mathematical reasoning
3. **Transfer learning**: These representations might transfer to related mathematical domains

4.4 Limitations and Future Work

Several limitations should be noted:

- **Single-token limitation**: Analysis focused only on numbers representable as single tokens
- **Correlation vs. causation**: While correlations are strong, causal relationships remain to be established
- **Model-specific findings**: Results are specific to OLMo-2 and may not generalize to other architectures
- **Limited mathematical domain coverage**: Analysis focused on basic properties rather than advanced mathematical concepts

Future work should investigate: - Multi-token number representations - Cross-model validation of findings - Performance correlation between embedding structure and mathematical task performance - Enhancement experiments using identified specialized dimensions

5. Conclusion

This analysis demonstrates that OLMo-2-1124-7B has developed sophisticated mathematical property detection capabilities within its embedding space. The emergence of specialized dimensions for magnitude,

digit count, Fibonacci sequences, and prime numbers occurs without explicit mathematical training objectives, suggesting these representations arise naturally from language modeling.

The hierarchical organization observed—from basic magnitude detection to complex sequence recognition—provides insights into how large language models internally represent numerical concepts. While these findings show structural similarities to theories of mathematical cognition, further research is needed to establish the functional significance of these representations for mathematical reasoning performance.

The identification of specific mathematical property detectors opens new avenues for model interpretation and enhancement, potentially leading to more effective approaches for improving mathematical capabilities in large language models.

References

[Note: This would include relevant citations to savant syndrome research, mathematical cognition studies, and embedding analysis literature in the actual thesis]

Paragraphs yet to contextualize - not a real section

In (Mottron et al., 2006), the hypothesis is also that the capabilities of the savant might come from privileged access to lower-level perceptual processing systems that have been functionally re-dedicated to symbolic material processing. This suggests that mathematical savants may bypass high-level algorithmic reasoning entirely, instead leveraging perceptual mechanisms that can directly recognize patterns in numerical relationships - much like how we might instantly recognize a face without consciously processing its individual features.

Given the strong association between sequence-space synesthesia and Savant syndrome, the additional sensorial information wouldn't just be an aberration, but possibly the encoding of structural and mathematical information on a hijacked channel. The encoding, giving the subjects the extraordinary abilities they are gifted with,

Bibliography

Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, 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

Cowan, R., & Frith, C. (2009). Do calendrical savants use calculation to answer date questions? A functional magnetic resonance imaging study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1522), 1417–1424. https://doi.org/10.1098/rstb.2008.0323

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/

Mottron, L., Lemmens, K., Gagnon, L., & Seron, X. (2006). Non-algorithmic access to calendar information in a calendar calculator with autism. *Journal of Autism and Developmental Disorders*, *36*(2), 239–247. https://doi.org/10.1007/s10803-005-0059-9

Murray, A. L. (2010). Can the existence of highly accessible concrete representations explain savant skills? Some insights from synaesthesia. *Medical Hypotheses*, *74*(6), 1006–1012. https://doi.org/10.1016/j.mehy.2010.01.014

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

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