Language Models Encode Numbers Using Digit Representations in Base 10

Amit Arnold Levy[*]
University of Oxford
amit.levy@keble.ox.ac.uk

Mor Geva
Tel Aviv University
morgeva@tauex.tau.ac.il

Abstract

Large language models (LLMs) frequently make errors when handling even simple numerical problems, such as comparing two small numbers. A natural hypothesis is that these errors stem from how LLMs represent numbers, and specifically, whether their representations of numbers capture their numeric values. We tackle this question from the observation that LLM errors on numerical tasks are often distributed across *the digits* of the answer rather than normally around *its numeric value*. Through a series of probing experiments and causal interventions, we show that LLMs internally represent numbers with individual circular representations per-digit in base 10. This digit-wise representation, as opposed to a value representation, sheds light on the error patterns of models on tasks involving numerical reasoning and could serve as a basis for future studies on analyzing numerical mechanisms in LLMs.

1 Introduction

Despite their high performance on various challenging tasks (Bubeck et al., 2023; Bommasani et al., 2021; Trinh et al., 2024), large language models (LLMs) often struggle with simple numerical problems, such as adding or comparing the magnitude of two small numbers. While previous works commonly attribute such failures to different limitations in the representations of LLMs (e.g., McLeish et al., 2024; Nogueira et al., 2021), *how* LLMs represent numbers is still an outstanding question.

Recently, Zhu et al. (2024) used linear probes to predict the number encoded in a hidden representation, showing high correlation with the expected value. However, the probes exhibited low accuracy, suggesting that a linear representation alone is not sufficient to explain how LLMs can often perform exact numerical operations, such as addition and multiplication. Maltoni and Ferrara (2024) have
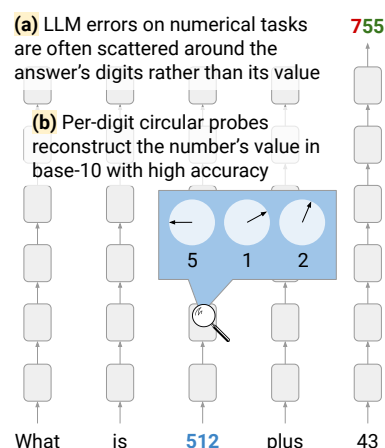

Figure 1: An illustration of our key findings, suggesting that LLMs represent numbers on a per-digit base-10 basis: (a) on simple numerical tasks, LLMs often make errors that are close to the answer in 'digit space' rather than in value space, (b) though probing the exact number is hard, digit values can be decoded accurately.

suggested that LLMs may do arithmetic in "value space", but then we would expect to see a normally-distributed error pattern, which we will see is not the case in widely-used models.

We approach the above question by observing that when models make numerical errors, the errors are often distant from the correct answer in value space but close in 'digit space'. For example, consider the simple addition problem "$132 + 238 + 324 + 139 = $" where the correct answer is 833. LLMs are more likely to generate errors with high string-similarity to the correct answer, such as "633" or "823", than natural errors like "831" or "834", which are close in value, as if the model's internal algorithm misreads one of the digits in the input. We show this rigorously in §2.

We argue that such scattered error distributions are unlikely to occur in models that directly manipulate numbers in a value space. For example, in multi-operand addition, if the model represents

---

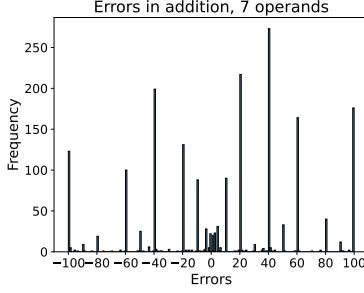[*] Work done at Tel Aviv University.

Figure 2: Error distribution in 7 operand addition.

each number in a value space and then translates the result back to tokens after addition, we would expect a normal error distribution around the correct answer. This distribution would arise from noise in the addition operation and the representations themselves. However, the observed scattered errors (in §2) suggest that the model may represent numbers in a fragmented manner, for example based on their individual digits.

To test this hypothesis, we first train probes to recover the number value and digit values from hidden representations of numbers. Our experiments with Llama 3 8B (Dubey et al., 2024) and Mistral 7B (Jiang et al., 2023) show that, while probes fail to recover the exact number value directly (which agrees with Zhu et al., 2024), the hidden representations of a number contain an orthogonal circular representation for each digit in base 10 (as illustrated in Figure 1). This observation holds across both models, which use different tokenization schemes for numbers. Moreover, causally intervening on these circular digit representations (i.e., performing +5 mod10) often modifies the value of the number accurately.

To conclude, our work proposes that the scattered errors LLMs demonstrate on arithmetic problems stem from a fragmented digit-wise representation of numbers. We show that this hypothesis holds in practice; it is possible to accurately recover and modify the digit values from number representations in base 10, but not the number values. Our findings provide a basis for understanding mathematical operations in LLMs and mitigating numerical errors. We release our code at https://github.com/amitlevy/base10.

2 Model Errors on Numerical Tasks are Scattered Across Digits

We analyze the distribution of errors by Llama 3 8B on two simple numerical tasks with numbers within the range 0 to 999, which the model rep-

Digit	Correct		Incorrect	
Units	4,232	94%	259	6%
Tens	4,351	97%	140	3%
Hundreds	4,054	92%	356	8%

Table 1: Accuracy of Llama 3 8B in comparing the magnitude of two numbers differing by one digit.

resents as individual tokens. We find that errors are distributed in a digit-wise manner, where an incorrect prediction is close to the correct answer in string edit distance but not in value space.

Task 1: Multi-operand addition We generated 5,000 queries of addition of $N = 7$ operands, which sum into a number between 0 and 1000, and calculated the errors of the model on these queries. Figure 2 displays the error distribution, showing that most errors are exact multiples of 10 and 100. Further, when considering the error distributions for any number of operands between 4 and 8, we observe that about 80% of the errors are in a single output digit, which is often not the units digit. A similar error analysis of GPT-4o (OpenAI et al., 2024) on 20-operand addition tasks showed similar trends (§A.1).

Task 2: Comparison of two numbers We consider all pairs of numbers between 0 and 999, which differ from each other in only a single digit—the units, tens, or hundreds place. Given a pair of numbers, the model needs to indicate which number is larger, e.g. *"between 121 or 171, the larger number is:"*. Table 1 shows that errors are distributed approximately equally between the digits. This indicates that the model's likelihood of making a mistake is not significantly affected by the numerical closeness of the numbers, as would be expected if numbers were represented in value space.

The evident base-10 digit-related error trends in both tasks lead to the hypothesis that LLMs may represent numbers in base 10 as opposed to in a linear value space, which we test in the next section.

3 LLMs Represent Numbers Digit-Wise in Base 10

We test our hypothesis and show that LLMs represent numbers on a per-digit base 10 basis.

3.1 Experimental setting

Probing We train digit-wise probes that estimate the value of a number from its hidden representa-

Basis	2	3	4	5	6	7	8	9	10	11	12	13	14	1000	2000
Llama 3 8B	0.16	0.06	0.16	<u>0.67</u>	0.05	0.08	0.06	0.07	**0.91**	0.08	0.06	0.06	0.06	0.00	0.00
Mistral 7B	0.13	0.02	0.13	<u>0.72</u>	0.02	0.05	0.05	0.08	**0.92**	0.12	0.04	0.06	0.05	0.01	0.00

Table 2: Accuracy in predicting all digits of digit-wise circular probes in various bases, averaged over layers ≥ 3.

tion by predicting the numeric values of its digits. Let M be a pre-trained transformer-based language model (Vaswani et al., 2017) with L layers and a hidden dimension d, and denote by $\mathbf{h}_j^\ell$ the hidden representation of the j-th input token at layer ℓ. In the following, we omit the position index and use $\mathbf{h}^\ell$, as in our experiments we always consider the last position of the input (i.e., the last numeric token). For a digit i, a base b, and a layer $\ell \in [L]$, we train a circular probe (Engels et al., 2024) that given the hidden representation $\mathbf{h}^\ell$ of a number x, predicts the numeric value of its i-th digit in base b:

$$\mathbf{P}_{i,b}^\ell = \arg\min_{\mathbf{P}' \in \mathbb{R}^{2 \times d}} \sum_{\langle \mathbf{h}^\ell, x_i \rangle \in \mathcal{D}^\ell} \left\| \mathbf{P}'\mathbf{h}^\ell - \texttt{circle}_b(x_i) \right\|_2^2$$

(1)

$\mathcal{D}^\ell$ is a training set consisting of pairs $\langle \mathbf{h}^\ell, x_i \rangle$ of the ℓ-th layer hidden representation and the i-th digit of a number x, and

$$\texttt{circle}_b(t) = [\cos(2\pi t/b), \sin(2\pi t/b)] \quad (2)$$

maps a digit in base b to a point on the unit circle.

Using the set of probes for some layer ℓ, we define a function that reconstructs the value of a number x from its representation $\mathbf{h}^\ell$. For every digit i, define a function $\texttt{digit}_{i,b}^\ell : \mathbb{R}^d \to [b]$ that predicts the value of that digit by applying $\texttt{digit}_{i,b}^\ell := \frac{b}{2\pi} \cdot \texttt{atan2}(\mathbf{P}_{i,b}^\ell \mathbf{h}^\ell)$.[1] Concatenating the outputs of the functions for all the digits of x provides an estimation of its value in base b. For example, the value of a 3-digit number would be reconstructed in base b from its ℓ-layer representation by concatenating $[\texttt{digit}_{3,b}^\ell, \texttt{digit}_{2,b}^\ell, \texttt{digit}_{1,b}^\ell]$.

In addition to the circular probes, we trained linear probes, which have been used recently to extract various features from LLM representations (Belinkov, 2022; Park et al., 2023; Gurnee and Tegmark, 2024). While the linear probes showed similar trends to the circular probes, we observed they are less effective in predicting numerical values from LLM representations. This observation agrees with recent findings that some features in LLMs have non-linear representations (Engels

et al., 2024) as well as with the circular patterns observed in PCA plots (see §A.2). Therefore, in our experiments we focus on circular probes.

Data For each positive number $x \in [2000]$ we feed "$\langle x \rangle$" (the value of x as a string) as input to the model and extracted the hidden representations from every layer $\ell \in [L]$. In cases when x is tokenized into multiple tokens, we take the representation at the last position (we assume that M is an auto-regressive model). For each basis b, we randomly split the numbers into train and validation sets with 1800 and 200 numbers, respectively.

Models We analyze two popular auto-regressive decoder-only LLMs: Llama 3 8B (Dubey et al., 2024) and Mistral 7B (Jiang et al., 2023). Llama's tokenizer contains individual tokens for all numbers between 0 and 999 inclusive, which is the common choice for modern LLMs (e.g., GPT-4 Singh and Strouse (2024) and Claude Sonnet 3.5). Mistral 7B was picked for having a different tokenization from Llama, specifically a single token per digit, which can be expected to impose a stronger bias towards digit-wise representations of numbers.

3.2 Probing recovers digit values in base 10 but not the whole number value

Table 2 shows the average probe accuracy over layers ≥ 3 in predicting all the digits of the number correctly (maximum accuracy results show similar trends; see §A.3). We do not consider the early layers as multi-token numbers require multiple layers to contextualize (see Figure 7 in §A.3).

The highest accuracy of 0.91 for Llama 3 8B and 0.92 for Mistral 7B is achieved when reconstructing the numbers in base 10. Moreover, for all other bases, accuracy is substantially lower, typically not exceeding 0.2, serving as a natural baseline for the base 10 results. Specifically, classifying the number directly (base 2000) succeeds in only $< 1\%$ of the cases, which further shows that the direct circular representation of the value in the hidden space is not accurate enough for arithmetic, similarly to the linear representation mentioned earlier. Interestingly, base 5 also has relatively high

---

[1] atan2 computes the two argument arctangent, which we convert from a signed to an unsigned angle between 0 and 2π.

accuracy, though significantly below base 10.

Overall, these results show that while reconstructing the number value directly generally fails, reconstructing digit-by-digit in base 10 succeeds with high accuracy. Importantly, while such a representation has advantages (see discussion in §5), it is surprising considering that LLMs typically have individual tokens for multi-digit numbers, which is not naturally base 10. We show evidence that the probes extend to representations of word form numbers, without being trained on them, in §A.4.

3.3 Modifying a digit representation modifies the whole number value accordingly

Our experiments suggest that models may represent numbers in a per-digit base 10 basis rather than store the number value directly. Here we conduct a causal experiment to test if this digit-wise representation is used by Llama 3 8B during inference.

Experiment Since the digit representations are circular in base 10, if we flip a number's hidden representation along the two directions of the probe (Eq. 2), we would expect the modified representation to encode the same number but with one digit flipped, i.e. the digit corresponding to the probe will now take a value of $v + 5 \pmod{10}$ where v was the original digit value before the intervention. For example (Figure 3), flipping the tens digit in the representation of 375 is expected to produce a representation of 325. For more details see §B.

To test this intervention, we consider the model's inference pass on a query "$\langle x \rangle + 0 = $" with some number x, for which the model initially generates x as the output. Then, we intervene on the representation of x at layer ℓ, apply the procedure described above to change one of x's digits, and continue the model's run to obtain a new output x'. Let x_i and x'_i be the i-th digits of x and x', we then check whether $x'_i = x_i + 5 \pmod{10}$ and for all $j \neq i$ that $x'_j = x_j$. We further define the prediction to be "close" to the intended result if it is closer to the intended result than an off by 1 error in the intervention digit. We conduct this experiment using all natural numbers 0 through 999. For each number, we perform the intervention once for every digit at layer 3, where the probes extract the number with high accuracy and before the information would propagate to the last position from which the prediction is obtained.

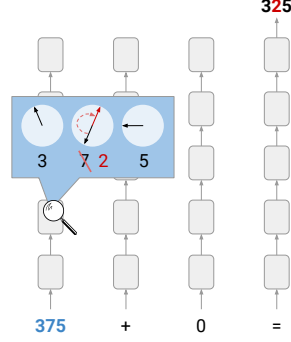**Results** For the hundreds digit, the exact intended result was achieved 15% of the time, while 47% of

Figure 3: An illustration of our intervention on number representations via circular per-digit probes in base 10.

the results were 'close' to the intended number, e.g., the digit was changed but with an error of 1 from the intended outcome. These numbers were respectively 10% and 50% for the tens digit, and 15% and 50% for the units digit. As a baseline, using a linear intervention following Zhu et al. (2024), but with the appropriate change to the normalization such that a specific number is targeted instead of a general direction, the exact result is achieved in less than 1% of the cases. A random baseline would be replacing the numeric token with another random numeric token in the range of the intervention, leading to a random baseline accuracy of 0.1%.

We conclude that there is a causal significance to the digit-wise circular representation, but there might be secondary representations or that some information might transfer before layer 3.

4 Related Work

Representation of numbers in LLMs There has been some investigation into how LLMs may represent numeric magnitude, involving linear probes of hidden representations (Zhu et al., 2024; Heinzerling and Inui, 2024) and embeddings (Wallace et al., 2019). In Gould et al. (2024) the authors looked into modular features of the first layer's hidden representations, and observed that modulus 10 seems of particular importance, but did not look beyond the units digit. To the best of our knowledge, no prior work has succeeded in training probes that extract the value of a held-out number from an LLM representation with the precision necessary to explain LLMs' successes on arithmetic.

Mechanistic interpretability of arithmetic tasks There has been much interest in looking into how LLMs may perform arithmetic tasks. Recent work has largely focused on either performing in depth

analysis of the algorithms learned by toy models (Maltoni and Ferrara, 2024; Nanda et al., 2023; Quirke and Barez, 2024; Yehudai et al., 2024) or analyzing information flow in trained open source LLMs (Stolfo et al., 2023; Chen et al., 2024). Most recently, Zhou et al. (2024) demonstrated that LLMs utilize Fourier features for arithmetic operations, with distinct roles for low- and high-frequency components. Our work complements these efforts and provides a basis for future work in this avenue, by analyzing the representations of numbers in modern LLMs.

Failures of LLMs on arithmetic tasks Razeghi et al. (2022) have looked into the performance of GPT-J 6B on arithmetic tasks, showing it is correlated with the frequency of the terms in the training dataset, which potentially suggests that LLMs may not be reasoning at all. While explaining LLM errors with number frequencies is valuable and may be more plausible in terms of the performance seen in older models, Llama 3 8B can perform 7 operand, 2-digit addition (10^{14} possible problems) with about 50% accuracy, which is far beyond the number of problems that could possibly be in the training data.

5   Conclusion and Discussion

While previous research has demonstrated that linear probes struggle to accurately extract numerical values from hidden representations — which are necessary for performing exact arithmetic operations like addition and multiplication — our findings indicate that circular digit-wise probes can effectively achieve this in two models with different tokenization. We have further demonstrated that editing these representations can alter the encoded number and consequently the model generation. These nonlinear representations align with Engels et al. (2024), who showed circular representations for the days of the week and months of the year.

Why would models construct digit-wise base-10 representations? Digit-wise representations may be more robust to noise in computations. If the number 120 is represented in value space, and has 1% of relative noise introduced as a result of an operation, it may now be represented as 121 instead, leading to a mistake in the model's generation. Conversely, if 120 is represented in 'digit space', an error of 1% is not enough to change any of the digits independently. That is, the model can

self-correct the number after the operation. Regarding the specific usage of base 10, one can presume it is because of the bias in the model's training data. That is, the model often has uses for the digits of a number, which biases the model toward learning to represent numbers in base 10, and as a result using that representation during operations.

Limitations

Our experiments show that the digit-wise circular representations exist and can be extracted, and that they are more significant causally than previously described representations of magnitude and are sufficient for arithmetic. However, we do not show conclusively that the representation is the only representation of numeracy in the hidden representations of LLMs. That is, there may be a superposition of multiple redundant representations. Finally, our focus was exclusively on the natural numbers - which are only a subset of the numeric values that exist. Nevertheless, the natural numbers are the most prevalent and a natural starting point, and it could be expected that the digit-wise base 10 representation extends also to fractions, which we leave for future work to explore.

Acknowledgements

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak,

Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Junhao Chen, Shengding Hu, Zhiyuan Liu, and Maosong Sun. 2024. States hidden in hidden states: Llms emerge discrete state representations implicitly. *arXiv preprint arXiv:2407.11421*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Mont-

gomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. 2024. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Davide Maltoni and Matteo Ferrara. 2024. Arithmetic with language models: From memorization to computation. *Neural Networks*, 179:106550.

Sean Michael McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, and Tom Goldstein. 2024. Transformers can do arithmetic with the right embeddings. In *ICML 2024 Workshop on LLMs and Cognition*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of transformers with simple arithmetic tasks. *arXiv preprint arXiv:2102.13019*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec

Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geome-

try of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Philip Quirke and Fazl Barez. 2024. Understanding addition in transformers. In *The Twelfth International Conference on Learning Representations*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaditya K Singh and DJ Strouse. 2024. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Gilad Yehudai, Haim Kaplan, Asma Ghandeharioun, Mor Geva, and Amir Globerson. 2024. When can transformers count to n? *arXiv preprint arXiv:2407.15160*.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. 2024. Pre-trained large language models use fourier features to compute addition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Fangwei Zhu, Damai Dai, and Zhifang Sui. 2024. Language models know the value of numbers. *arXiv preprint arXiv:2401.03735*.
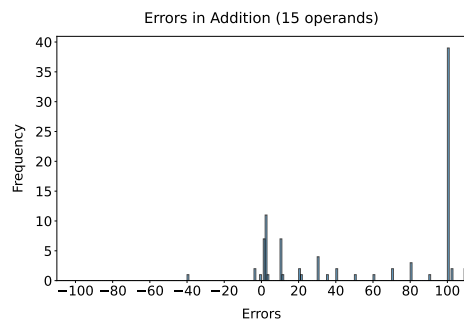
Figure 4: Error distribution in 15 operand addition for GPT-4o.

A Additional Results

A.1 Error patterns of GPT-4o

We conducted an additional error analysis using GPT-4o (OpenAI et al., 2024) on 15-operand addition tasks. Increasing the number of operands was necessary due to the model's high accuracy on simpler addition problems. The results were consistent with the trends observed in Figure 2, showing the majority of errors are at multiples of 10, as seen in Figure 4. This indicates that the fragmented error distribution identified in smaller models persists in larger models.

Increasing the number of digits instead of the number of operands leads to errors in multiples of 100 and 1,000 as well, showing that the error distribution stays indicative of a fragmented representation also for other digits.

A.2 PCA of hidden representations

We visualized the hidden states for natural number tokens 0 to 999 in layer 2 of Llama 3 8B, projected onto their top two principal components. In Figure 5 we can see that there are two half circles, one contained at the edge of the other. One is a half circle of all the numbers, and the next is of all numbers 0-99.

An interesting observation is that within each half-circle, the numbers increase in a clockwise direction, indicating that the model may represent digits circularly. In the circle for the numbers 0-99, the numbers increase clockwise, and again when you look at the half-circle that contains the rest of the numbers. This indicates that at least the hundreds digits and tens digits are represented circularly.

In Figure 6 we can see that the circular pattern in the tens digit also extends to all numbers 0 to 999, when the dominance of the hundreds digit is

removed through averaging out all numbers into 10 groups by their tens digit.
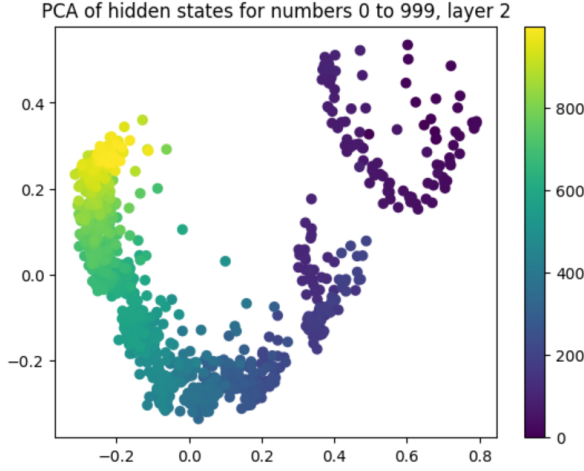

Figure 5: Visualization of the hidden states for natural number tokens (0 to 999) in layer 2 of Llama 3 8B, projected onto their top two principal components.
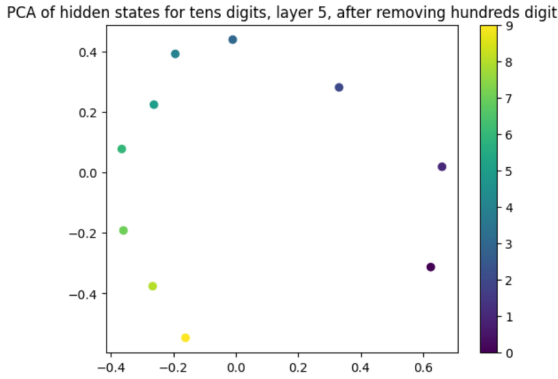

Figure 6: Visualization of the averaged hidden states for natural number tokens (0 to 999), grouped by their tens digit (0—9), in layer 5 of Llama 3 8B, projected onto their top two principal components. For example, numbers like 101 and 406, both having a tens digit of 0, are grouped together.

A.3 Accuracy of circular probes

In the main results we showed the accuracy of the circular digit-wise probes, averaged over layers ≥ 3. Here we will show this choice is justified as can be seen in Figure 7. While there is significant variations between layers, the accuracy is especially low before the contextualization that happens in the first 3 layers.

Another interesting question is which layer's set of digit-wise circular-probes have the highest accuracy in predicting the number, and how accurate is it. The corresponding results can be seen in Ta-

ble 3. It can be observed that in Mistral 7B, in the best layer, the circular probes achieve perfect accuracy on the validation set. That is, the number can always be recreated perfectly.
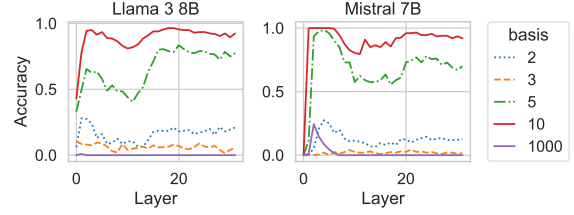

Figure 7: Accuracy of the circular probes in different bases across layers in Llama (left) and Mistral (right).

A.4 Probing representations of numbers in word form

Since numbers can also be represented in word form, i.e. twenty-two for the number 22, we further tested if our digit-wise circular probes extend to these representations, without being explicitly trained on them. Concretely, we evaluated the probes' accuracy for Llama 3 8B on the numbers 'zero' through 'fifty' in word form.

We observe that the accuracy varies depending on the layer used, with a peak of 68.6 accuracy when using the representations at layer 14. This is an encouraging sign that the circular probes generalize beyond the specific setting they were trained on, which further supports our causal results.

B Causal Intervention Details

We provide additional details on the interventions performed in §3.3. In practice, since the two directions of the circular probe are approximately orthogonal, we project the hidden representation onto each direction, subtract these components to remove the original digit representation, and then add the components back with their directions reversed to modify the digit. We also scaled the projection by a fixed constant ($a = 19$), assuming that if the model has multiple representations for numbers, scaling the representation will make the model place more weight upon it. The exact constant was chosen through binary search, in order to select the largest scaling factor such that the model still predicts a number, as it was observed that with a very high scaling factor the model starts predicting non-numeric tokens.

Basis	2	3	4	5	6	7	8	9	10	11	12	13	14	1000	2000
Llama 3 8B	0.24	0.10	0.25	<u>0.84</u>	0.08	0.10	0.10	0.11	**0.96**	0.12	0.09	0.10	0.11	0.00	0.02
Mistral 7B	0.28	0.04	0.22	<u>0.98</u>	0.04	0.08	0.12	0.23	**1.00**	0.29	0.08	0.18	0.10	0.14	0.03

Table 3: Accuracy of the digit-wise circular probes for different bases in predicting all digits correctly, taking the layer with the highest accuracy.