The Geometry of Numerical Reasoning: Language Models Compare Numeric Properties in Linear Subspaces

Ahmed Oumar El-Shangiti[1] **Tatsuya Hiraoka**[1] **Hilal AlQuabeh**[1]
Benjamin Heinzerling[3,2] **Kentaro Inui**[1,2,3]
[1] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
[2]Tohoku University
[3]RIKEN
ahmed.oumar@mbzuai.ac.ae

Abstract

This paper investigates whether large language models (LLMs) utilize numerical attributes encoded in a low-dimensional subspace of the embedding space when answering questions involving numeric comparisons, e.g., *Was Cristiano born before Messi?*. We first identified, using partial least squares regression, these subspaces, which effectively encode the numerical attributes associated with the entities in comparison prompts. Further, we demonstrate causality, by intervening in these subspaces to manipulate hidden states, thereby altering the LLM's comparison outcomes. Experiments conducted on three different LLMs showed that our results hold across different numerical attributes, indicating that LLMs utilize the linearly encoded information for numerical reasoning.

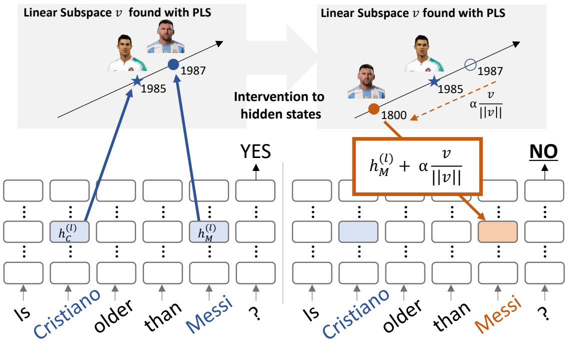

Linear Subspace v found with PLS
1987
1985
Intervention to hidden states
Linear Subspace v found with PLS
1987
1985
1800
$\alpha \frac{v}{||v||}$
YES
$h_C^{(l)}$
$h_M^{(l)}$
$h_M^{(l)} + \alpha \frac{v}{||v||}$
NO
Is
Cristiano
older
than
Messi
?
Is
Cristiano
older
than
Messi
?

Figure 1: Summary of our approach. We extract contextualized numeric attribute activations and then train k-components PLS model on the activations to predict their values and then use the first component of the PLS model to do an intervention at the last token of the second entity in the logical comparison.

1 Introduction

Language models (LMs) store large amounts of world knowledge in their parameters (Petroni et al., 2019; Jiang et al., 2020; Roberts et al., 2020; Heinzerling and Inui, 2021; Kassner et al., 2021). While prior work has evaluated parametric knowledge mainly via behavioral benchmarks, more recent work has analyzed how knowledge is represented in activation space, for example, localizing relational knowledge to specific layers and token representations (Meng et al., 2022; Geva et al., 2023; Merullo et al., 2024) or identifying subspaces that encode numeric properties such as an entity's birth year (Heinzerling and Inui, 2024). However, analysis of LM-internal knowledge representation has been limited to simple factual recall, e.g., for queries like "When was Cristiano born?" (Answer: 1985) or "When was Messi born?" (Answer: 1987). If and how the mechanisms responsible for simple factual recall also participate in more complex queries, e.g., "Is Cristiano older than Messi?", is not understood so far. A possible mechanism by which an LLM answers this query is a multi-step process consisting of first recalling the respective birth years of the two entities, comparing the two years, and then selecting a corresponding answer.

Herein, we focus on LLM's ability of arithmetic operations (Dehaene, 2011). The LLM's ability to handle numbers has been discussed after the advent of pre-trained language models (Spithourakis and Riedel, 2018; Wallace et al., 2019). With modern LLMs such as the LLaMA family (Touvron et al., 2023), Heinzerling and Inui (2024) shows that LLMs map numerical attributes such as *(Cristiano, born-in, 1985)* and *(Messi, born-in, 1987)* to low-dimensional (Linear) subspaces and prove that those subspaces are used during knowledge extraction. However, it is not clear whether the LLMs use those subspaces to solve logical reasoning such as the relation *(Cristiano, born-before, Messi)*.

In this study, we tackle the research question: **do LLMs leverage the linear subspace of entity-numerical attributes when solving numerical reasoning tasks?** We investigate whether the linear subspace is indeed used in the logical reasoning tasks. We first show the LLMs' capability to solve

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 550–561
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Experiment	Question	Response
Extraction	Birth year of Albert Einstein?	1879
	What is Isaac Newton's year of death?	1727
	Latitude of Cairo?	30.04° N
Reasoning	Einstein born before Newton?	No
	Einstein died before Newton?	No
	Is Cairo's latitude higher than Jerusalem's?	Yes

Table 1: Samples from Extracting Information and Comparisons Experiments

the numerical reasoning tasks from the viewpoint of behavioral observation: testing the performance of the reasoning task with in-context learning (§3). We then examine the representations of LLMs (§4). We identify the linear subspace corresponding to the numerical attributes with partial least-squares (PLS (Wold et al., 2001)) and intervene in the representation to test whether the model utilizes the linearly represented information (see Figure 1).

The experimental results on the three numerical properties (the birth/death year of a person and the latitude of location) and on three LLMs (LLama3 8B (Dubey et al., 2024), Mistral 7B (Jiang et al., 2023), and Qwen2.5 7B (Team, 2024) all instruction based models) demonstrate that LLMs leverage the numerical information represented in the linear subspace for the reasoning tasks.

2 Outline of Experiments

This section outlines our methodology to investigate the process of LLMs to solve the numerical reasoning.

2.1 Model and Dataset

In this work, we focus on the three numerical properties: the birth years of person entities, the death years of person entities, and the latitudes of location entities. Table 1 exemplifies the questions and expected responses for both tasks. For the knowledge extraction task, we create the question-answer pairs by extracting 5,000 entities alongside their numerical attributes from Wikidata (Vrandečić and Krötzsch, 2014). After filtering out entities that the LLM does not know (§3.1), we created the 5,000 questions about numerical reasoning that include two entities each. For all experiments, we used Llama3-8B-instruction following model (Dubey et al., 2024) as the LLM and later validate our finding on two additional models (see § 4.3).

2.2 Design of Experiments

We conducted the experiments in two phases to investigate the LLM's ability to utilize the linear subspace for numerical reasoning.

Data Pre-processing (§3): We began by evaluating the LLM's ability to handle both knowledge extraction and numerical reasoning tasks by inputting questions and evaluating its response. To focus the subsequent experiments on entities for which the LLM has reliable numerical knowledge, we filtered out any entities that the LLM could not answer correctly during this initial behavioral experiment.

Internal Representation Experiments (§4): In the second phase, we examined the inner workings of the LLM when solving the knowledge extraction (§4.1) and the numerical reasoning (§B.1). Here, we focus on analyzing the hidden state of each entity representation at a particular layer for knowledge extraction. For the case of numerical reasoning, we investigated the activations of the last token's representation. We denote the hidden state of the i-th input at the l-th layer as $h_i^{(l)}$. To investigate whether knowledge of numerical attributes is stored in low-dimensional subspaces, we applied PLS (Wold et al., 2001) for each representation (Heinzerling and Inui, 2024). Partial Least Squares (PLS) offers an alternative to Principal Component Analysis (PCA) for dimensionality reduction, especially when predicting one set of variables from another. PLS seeks to maximize the covariance between the input matrix $\mathbf{X}$ and the response matrix $\mathbf{Y}$ by projecting both onto a latent space. Through PLS, we identified components that represent the linear structure of each numerical attribute, allowing us to analyze how the LLM might utilize these subspaces for reasoning. To further test this, we intervened in the hidden state $h_i^{(l)}$ by incorporating the 1st PLS component v, as follows:

$$h_i^{(l)} \leftarrow h_i^{(l)} + \alpha \frac{v}{||v||}, \quad (1)$$

where α is a hyperparameter derived from the first PLS component, and $||v||$ is the Euclidian norm (L2-norm) of the vector v. Intuitively, this intervention edits the numerical attribute captured by the LLM. For instance, if the numerical information *(Cristiano, born-in, 1985)* is shifted to *(Cristiano, born-in, 2020)*, an LLM that genuinely relies on a linear subspace for reasoning would adjust its interpretation accordingly, reflecting the change in its responses (Figure 1).

3 Data Pre-processing

The purpose of this experiment is to assess whether the LLM possesses knowledge of the numerical attributes of the entities prepared for this study, and to evaluate its capability to perform numerical reasoning tasks. Additionally, by conducting behavioral experiments focused on information extractions, we aim to filter out entities for which LLM lacks sufficient knowledge, therefore creating a refined dataset to be used in the subsequent numerical reasoning tasks. For both tasks, extraction and reasoning, we prepared ten distinct prompts. The prompts that demonstrated the best performance in preliminary tests were selected for further investigation of the internal representations (§4). Appendix 4 lists the complete list of prompts in the experiments.

3.1 Knowledge Extraction

To assess the LLM's knowledge extraction of entity numerical attributes, we conducted a zero-shot question-answering task, in which we asked direct questions about numerical attributes for various entities. The results summarized in the top half of Table 2, demonstrate that the LLM correctly answered at least 67% of the prepared questions with the best-performing prompt for each task.

3.2 Numerical Reasoning

For the numerical reasoning task, we created 5,000 question samples using a pair of unique entities, selected after filtering out those that the LLM could not answer correctly in §3.1. Each question was designed to prompt the model to perform numerical reasoning, with binary (Yes/No) answers indicating correctness. The results, shown in the bottom half of Table 2, reveal varying levels of accuracy across different prompts. The LLM achieved around 75% for birth/death year prediction, but only 56% for latitude-related questions, suggesting differences in task difficulty.

4 Internal Representation Experiments

This experiment aims to train a PLS model to identify low-dimensional linear subspaces within the activation space, which could potentially be efficient in predicting numerical attributes for various entities. We then demonstrated the causal relationship within these subspaces by implementing targeted interventions which shows that indeed there is a causal effect between the identified linear subspaces and the logical comparison answers by the model. We validate our hypothesis by running three models on three numerical attributes.

	Prompts									
Task	1	2	3	4	5	6	7	8	9	10
BP	66.0	70.0	67.4	66.2	72.3	67.6	66.9	66.6	68.2	71.3
DP	63.4	65.5	61.5	61.5	67.0	65.0	63.3	60.1	61.7	66.1
LP	47.6	72.0	69.0	70.0	69.0	68.5	61.5	69.0	69.0	66.6
BC	57.0	56.6	75.6	67.0	62.5	50.0	74.5	57.0	71.7	62.1
DC	53.5	50.3	74.8	58.7	50.5	50.2	50.3	61.8	50.1	56.6
LC	53.0	56.0	50.0	37.8	55.0	51.2	55.0	50.0	50.0	50.2

Table 2: Experiments 1 and 2's Results for three tasks, and 10 different prompts for each. The accuracy of exact matching is reported, except for the Latitude task, where we relaxed the predicted and ground truth to be rounded to the integer part. **BP**: Birth Prediction, **DP**: Death Prediction, **LP**: Latitude Prediction, **BC**: Birth Comparison, **DC**: Death Comparison, **LC**: Latitude Comparison

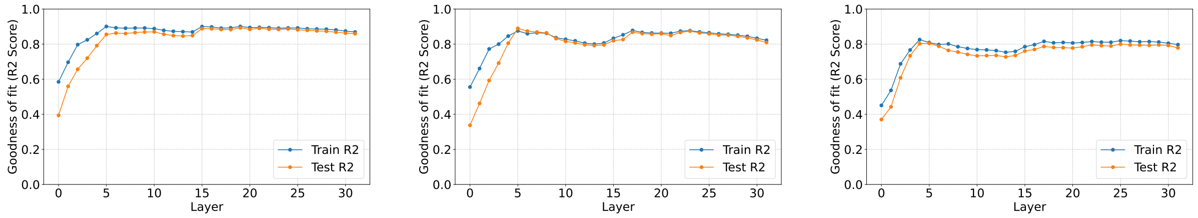
We also fitted another PLS model to evaluate Yes/No comparison reasoning related to these numerical attributes (see appendix B.1).

4.1 Prediction of numerical attributes with PLS

The training procedure consists of the following steps: (1) we first filter out the entities that the model predicted their comparison incorrectly (Section 3.2). (2) We feed a context vector that contains the comparison prompt (e.g., *Was Cristiano born prior to Messi?)* (3) We extract the hidden states of the last token of each entity from the LLM's hidden states at a particular layer. (4) These hidden states are then used to train a PLS model with a 5 component to predict the corresponding numerical attribute of each entity based on their corresponding model representation (activations). Figure 2 depicts the results achieved by $five$ components PLS model, measured by the coefficient of determination R^2. The goodness of fit exceeds 0.8 for all measured properties, indicating that the information encoded in these attributes can be extracted with low-dimensional (linear) subspaces.

4.2 Intervention using PLS Components Vector

While the previous experiments with the PLS model establish correlation, they do not demonstrate causality. For this purpose, we perform interventions at a particular token within a designated model layer, chosen based on the correlation strength identified in predicting numerical attributes from each task (Section 4.1). We fix the first entity and intervene at the last token of the

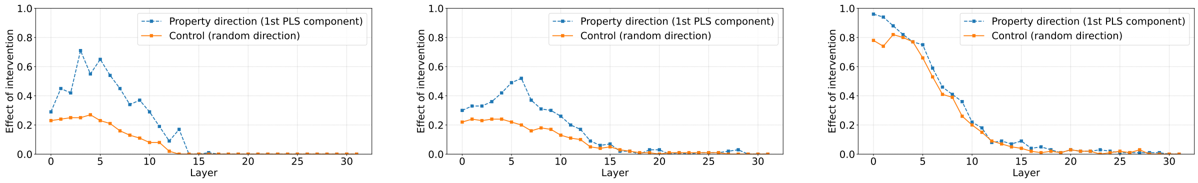


(a) Birth Attribute

(b) Death Attribute

(c) Latitude Attribute

Figure 2: The R^2 score of predicting entity's numerical attributes, using a 5-Component PLS model.



(a) Birth Attribute

(b) Death Attribute

(c) Latitude Attribute

Figure 3: The effect of the intervention—specifically, the ratio of flipped answers after performing intervention—was analyzed within the identified model subspace of each layer and compared to the effects observed in a randomly selected direction sampled from a normal distribution.

second entity. This token's hidden state is then updated by a scaled version of the first component direction from the PLS model to the original hidden state $h_i^{(l)}$ as illustrated in equation (1).

In Figure 3 we compare the effect of our intervention per layer against a random vector from the normal distribution. It is measured by the Effect of Intervention metric (EI) (equation 2), f and f' are the clean and patched models.

$$\mathrm{EI} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[f(x_i) \neq f'(x_i)\right] \qquad (2)$$

The results clearly demonstrate the superiority of our intervention method, particularly evident in Subfigures a and b. In subfigure c, related to the Latitude numeric attribute, the gap between our method and the baseline narrows, suggesting that the direction may not be significant for this attribute. This could reflect the mode's nearly random response in the behavior experiment (Section 3.2). Additionally, the intervention's effect is notable only in the first $\approx$ 50% of the model layers, after which it diminishes to zero, aligned with prior research on inference time theory. We also tested the generalization of our approach on unseen samples, as shown in appendix, Figure 9 and additional models (see § 4.3).

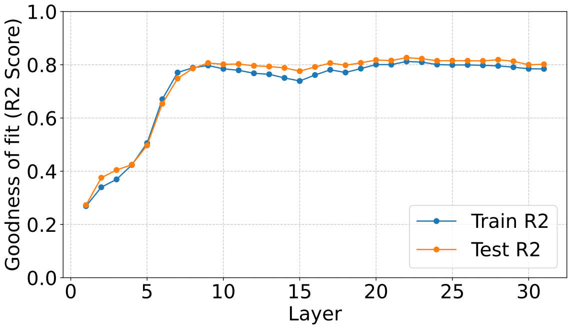


Figure 4: R^2 score of predicting entity's birth years attributes, using a 5-Component PLS model trained on Mistral 7B Instruct activations.

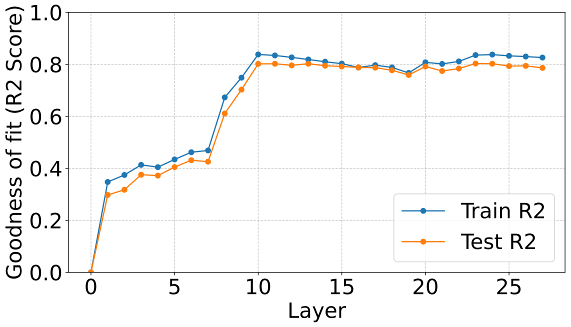


Figure 5: R^2 score of predicting entity's birth years attributes, using a 5-Component PLS model trained on Qwen2.5 7B Instruct activations.

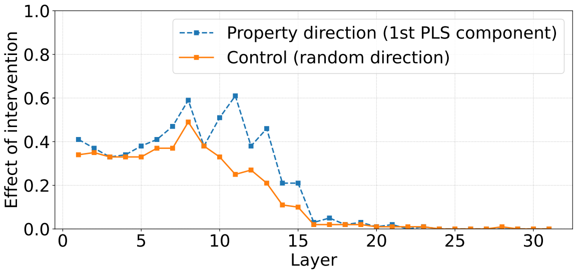


Figure 6: The effect of the intervention(i.e. the ratio of the flipped answers) in the identified subspace in each layer of the Mistral 7B Instruct model, compared to a random direction from a normal distribution.

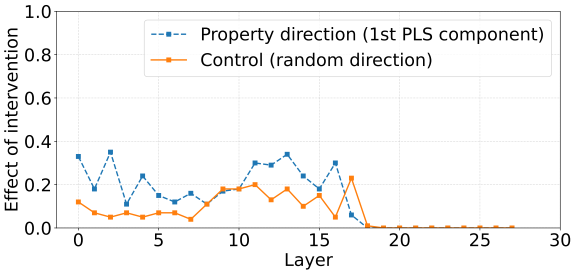


Figure 7: The effect of the intervention(i.e., the ratio of the flipped answer) in the identified subspace in each layer of the Qwen2.5 7B instruct model, compared to a random direction from a normal distribution.

4.3 Experiments on Additional Models

To further validate our hypothesis generalization, we run the same experiments on two additional language models for the *birth* property. Those additional models are Mistral-7B-intruct (Jiang et al., 2023) and Qwen2.5-7B-Instruct (Team, 2024).

PLS models trained on models' activation have crossed an R^2 score of 0.8 suggesting that the information encoded in those models' activations can be extracted using low-dimensional (linear) subspaces (see Figure 4 and Figure 5).

The Effect of Intervention (EI) results shown in Figures 6 and 7 of the Mistral 7B Instruct and Qwen-2.5 7B Instruct models, respectively, demonstrate the same behavior seen in the previous experiments. For the EI of Mistral, we can see that the peak was around the 11th layer and then continued to decrease until it finally disappeared around the 16 layer (Figure 6). When compared to other models, Qwen2.5 has shown two clear differences. First, we can observe two peaks for the EI with almost the same value of the EI, early around the $third$ layer and later one around layer 12, while other models have shown only one peak. Second, unlike other models, Qwen2.5 7B kept bouncing around almost the same EI values and suddenly become None at around layer 16 (Figure 7). One reason that might explain the difference between Qwen2.5 7B and other models, is that Qwen2.5 7B uses only 28 layers, while other models in the experiments are formed of 32 layers.

Task	Model	Prompts									
		1	2	3	4	5	6	7	8	9	10
BP	Mistral 7B	72.65	72.63	74.68	75.36	73.64	75.44	73.86	74.81	72.90	73.56
	Qwen2.5 7B	40.82	34.68	33.95	34.59	33.95	36.72	36.96	32.61	39.07	34.32
BC	Mistral 7B	53.60	64.84	64.02	53.10	61.88	57.66	53.00	67.06	64.68	50.00
	Qwen2.5 7B	29.20	58.10	38.88	26.22	49.76	40.54	6.20	3.84	9.16	6.98

Table 3: Exact Matching Accuracy of Mistral 7B and Qwen2.5 7B Models on Birth Date Numerical property extraction and Comparison Tasks Across Prompt Variations. All models are instruction-based models. **BP:** Birth Prediction and **BC:** Birth Comparison tasks are evaluated.

5 Conclusion

In this research, we empirically demonstrate that the model answers numerical reasoning questions, such as "Was Cristiano born before Messi?" using a two-step process. First, it extracts numerical attributes for each entity from a linear subspace. The second step involves utilizing these linear directions to answer the logical question. Specifically, subspaces are identified through PLS regression, where directions in low-dimensional subspaces of the activation space encode numerical property information. We illustrate this approach using three numerical attributes: Birth, Death, and Latitude across three LLMs. The reasoning step is validated using causal interventions along the direction of the first component of the PLS model, where these interventions successfully alter the model's answers.

6 Ethical Statement

Our work adheres to the ACL Code of Ethics and maintains a high standard of ethical research practice. We ensure that our methodology, data usage, and model development follow responsible AI principles, and that there are no ethical violations in our study. Our research does not involve the use of sensitive or private data, nor does it contribute to any potential harm or bias propagation. We remain committed to transparency, fairness, and the responsible application of large language models in line with ACL's ethical guidelines.

7 Limitations

This work has several limitations we plan to address in future work:

- Error Analysis: While the experimental results demonstrate the model's ability to map numerical properties to low-dimensional subspaces and use them for reasoning tasks, we have not conducted a thorough error analysis to understand the model's types of mistakes. Identifying patterns in erroneous outputs could guide improvements in both model design and training.
- Limited Scope of Numerical Attributes: Our experiments are restricted to three types of numerical attributes: birth year, death year, and geographic latitude. It remains unclear whether our findings extend to a broader range of numerical properties, such as financial data, time intervals, or other continuous variables. We plan to investigate this in future work.
- Intervention Hyperparameter Sensitivity: The success of the intervention experiments relies heavily on the choice of the scaling factor α applied during the intervention. We have not explored the full sensitivity of the model's performance to this hyperparameter, which could introduce biases or instability in real-world applications.

References

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics*. Oxford University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 Montgomery, 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. *Preprint*, arXiv:2407.21783.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric properties in language models. *Preprint*, arXiv:2403.10381.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

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. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. A mechanism for solving relational tasks in transformer language models.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? *arXiv preprint arXiv:2405.02421*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. In *Pre-print*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57:78–85.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

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.

Svante Wold, Michael Sjostrom, and Lennart Eriksson. 2001. Pls-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024. Interpreting and improving large language models in arithmetic calculation. *arXiv preprint arXiv:2409.01659*.

A Background

Generative-Transformer Language Models. Transformer models, particularly in generative contexts, have revolutionized natural language processing tasks due to their self-attention mechanisms. These models map an input sequence $x_1, x_2, \ldots, x_n$ to a corresponding sequence $y_1, y_2, \ldots, y_m$ using multi-layer perceptron, and multi-head self-attention layers, which compute attention scores based on the query-key-value system. Mathematically, for a given layer l, the attention output A_l is computed as:

$$A_l = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V are the query, key, and value matrices, and d_k is the dimension of the keys. By stacking multiple layers of these attention mechanisms and multi layer percptron, transformers efficiently capture long-range dependencies in text. The autoregressive nature of generative transformers allows them to generate coherent text sequences by predicting the next token based on previous tokens.

Representation Analysis of Transformer Language Models. Representation analysis of transformers has revealed important insights into how these models store and manipulate information across layers. Research has shown that transformer language models develop complex, hierarchical representations that can be understood by analyzing the attention patterns and hidden states at different layers (Niu et al., 2024). For example, studies have found that early layers capture syntactic structures, while deeper layers capture more semantic information (Hernandez et al., 2023). Recent work also uses probing techniques to analyze how specific linguistic features are represented, contributing to a growing understanding of model interpretability (Vulić et al., 2020).

Intervention and Activation Patching. One technique that has gained attention in the analysis of neural models, including transformers, is **activation patching**. This involves replacing activations in a specific layer with those from another input in order to study the effect of those activations on the final output. By intervening at different points within the model, researchers can better understand how information is processed and transformed throughout the network. This method has been useful in dissecting how specific neurons or attention heads contribute to a model's behavior, allowing for targeted interventions that shed light on model interpretability.

Linear Hypothesis in Representation. The **linear hypothesis** posits that the representations formed by transformer models are linearly separable. This means that complex patterns, such as syntactic and semantic categories, can be distinguished by applying a linear transformation to the learned embeddings (Park et al., 2023). The key idea here is that the hidden representations of different tasks or features align in such a way that linear classifiers can achieve good performance with minimal processing, a phenomenon observed across a range of neural architectures. Connecting this with the previous analysis, it appears that transformers structure their internal space in a way that is amenable to linear separation of features, thus facilitating tasks such as classification and regression.

Partial Least Squares (PLS). Partial Least Squares (PLS) offers an alternative to Principal Component Analysis (PCA) for dimensionality reduction, especially when predicting one set of variables from another. PLS seeks to maximize the covariance between the input matrix $\mathbf{X}$ and the response matrix $\mathbf{Y}$ by projecting both onto a latent space. The key idea is to find latent variables $\mathbf{T} = \mathbf{XW}$ and $\mathbf{U} = \mathbf{YC}$ that best capture this covariance.

The predictive relationship between $\mathbf{X}$ and $\mathbf{Y}$ is then modeled as:

$$\hat{\mathbf{Y}} = \mathbf{XWP}^T, \tag{4}$$

where $\hat{\mathbf{Y}}$ is the predicted output matrix, $\mathbf{P}$ are the loadings, and the quality of this prediction can be assessed using the coefficient of determination R^2. The R^2 value measures how well the model explains the variance in $\mathbf{Y}$, where higher values indicate a better fit between predicted and actual outputs.

PLS is preferred over regression when predictors (or columns of $\mathbf{X}$) are not independent or when the number of predictors exceeds the number of observations, making it suitable for high-dimensional data. For transformers, applying PLS helps uncover how input embeddings influence predictions by focusing on the shared variance between input features and outputs (Heinzerling and Inui, 2024).

B Related Work

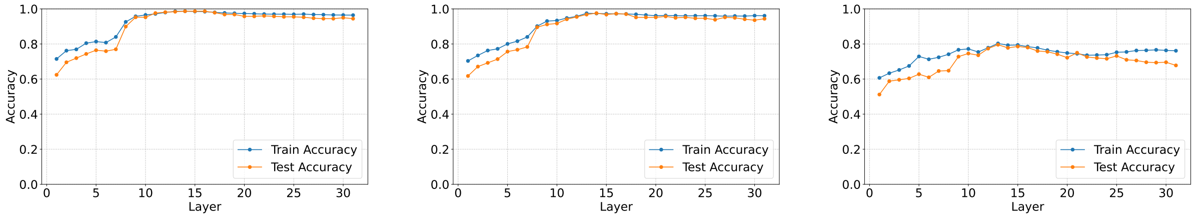
After the appearing of pre-trained language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT (Radford and Narasimhan, 2018), researchers have had interests in the numerical capability of language models. (Spithourakis and Riedel, 2018) evaluates the pre-trained language models from viewpoints of the output capability of numerical tokens, the behavioural side of the numeracy. (Wallace et al., 2019) focused on the numerical knowledge stored in the embeddings, which is the internal side of the numeracy. Zhang et al. (2024) investigated the internal working of the recent large language models when processing arithmetic calculation.

Knowledge of entities such as named entity has also been payed attention to by many researchers. Considering the pre-trained language models as a knowledge base (Petroni et al., 2019; Jiang et al., 2020), behavioral (Shin et al., 2020) and internal (Meng et al., 2022; Dai et al., 2022) analysis have been studied.

With much larger scale of language models such as GPT3 (Brown, 2020) and LLaMA (Touvron et al., 2023) and the technique of in-context learning, the capability of reasoning acquired by the language models has started to be discussed. (Merullo et al., 2024) examined the internal working of language models when solving the reasoning task of the entity-entity relation such as *(Paris, capital-of, France)*. Heinzerling and Inui (2024) provides a deeper observation of the reasoning of the entity-numeric relation such as *(Dijkstra, born-in, 1930)*. They reveal that the entity-numeric relations are stored in the language models' representation as keeping their monotonic structure. Following this work, we further dive into the numerical reasoning that requires the extraction of the entity-numeric knowledge and the comparison of the two numerical information such as *(Bellman, born-before, Dijkstra)*.

B.1 Logical Comparison with PLS

In this experiment, we feed the entire context vector containing a comparison into the model and extract the last hidden state of the last token for each comparison sample. We train a PLS model on these activations to predict the comparison results (i.e. Yes or No). We aim to make sure that the Yes/No task is predictable from model activations using a low-dimensional (linear) subspace. Figure 8 illustrates the accuracy of the 5-components PLS model in predicting the comparison results giving the model activations. The model shows near-perfect performance of the Birth and Death tasks, while less robust on the Latitude task. This outcome is consistent with findings from the Behavioral experiments in Section 3.2.

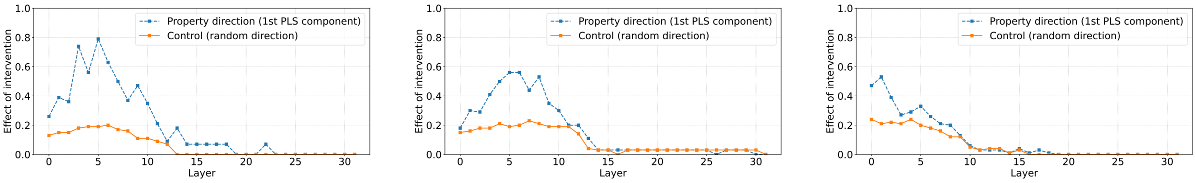


(a) Birth Attribute

(b) Death Attribute

(c) Latitude Attribute

Figure 8: The accuracy of predicting Yes/No in a comparison task of numerical attributes, using a 5-Component PLS model.



(a) Birth intervention

(b) Death intervention

(c) Latitude

Figure 9: Intervention graphs for out-of-distribution data samples on birth, death, and latitude tasks.

Birth	**Death**	**Latitude**
Did {entity_x} come into the world earlier than {entity_y}? Answer with Yes or No.	Did {entity_x} die before {entity_y}? Answer with Yes or No.	Is {entity_x} located at a higher latitude than {entity_y}? Answer Yes or No.
Is {entity_x}'s birthdate before {entity_y}'s? Respond with Yes or No.	Did {entity_x} pass away earlier than {entity_y}? Respond with Yes or No.	Is {entity_x} farther north than {entity_y}? Answer Yes or No.
Was {entity_x} born prior to {entity_y}? Output only Yes or No.	Was {entity_x}'s death prior to {entity_y}? Provide only Yes or No.	Does {entity_x} have a higher latitude value than {entity_y}? Answer Yes or No.
Did {entity_x} enter life before {entity_y}? Answer with Yes or No.	Did {entity_x} pass on before {entity_y}? Answer Yes or No.	Comparing latitudes, is {entity_x} north of {entity_y}? Answer Yes or No.
Was {entity_x}'s birth earlier than {entity_y}'s? Output only Yes or No.	Did {entity_x} die first compared to {entity_y}? Respond only with Yes or No.	In terms of latitude, is {entity_x} above {entity_y}? Answer Yes or No.
Was {entity_x} born first compared to {entity_y}? Respond with Yes or No.	Was {entity_x}'s death earlier than {entity_y}'s? Answer with Yes or No.	Is the latitude of {entity_x} greater than the latitude of {entity_y}? Answer Yes or No.
Is {entity_x} older than {entity_y}? Reply only with True or False.	Did {entity_x} precede {entity_y} in death? Reply only with True or False.	Geographically, is {entity_x} at a more northern latitude than {entity_y}? Answer Yes or No.
Did {entity_x} precede {entity_y} in birth? Respond only with True or False.	Did {entity_x} pass before {entity_y}? Respond only with True or False.	Does {entity_x} have a more northerly latitude compared to {entity_y}? Answer Yes or No.
Did {entity_x} arrive before {entity_y}? Answer only with True or False.	Did {entity_x} die earlier than {entity_y}? Answer only with Yes or No.	Is {entity_x} positioned at a latitude north of {entity_y}? Answer Yes or No.
Is {entity_x} senior to {entity_y}? Reply only with Correct or Incorrect.	Did {entity_x} pass away first compared to {entity_y}? Reply with Correct or Incorrect.	Considering only latitude, is {entity_x} more northward than {entity_y}? Answer Yes or No.

Table 4: Comprehensive list of prompts for our three tasks: for Birth, Death, and Latitude