hello Community today we are investigating how to override new knowledge and we have some brand new research how to do this in a professional way so let's start with this you know it is easy open your eyes shipes your dare propietary llm with this content here and this is the shape of the Moon and you say yep and then I have a question if I want to override this knowledge because now my moon should have a shape like a shoe box you know and now I want to calculate the gravitational field of the high tide in I didn't do England so you see now this is not a change of a single point of information it will change it a complete interconnected logical reasoning connectome in the llm how we do this in a professional way welcome to this video and I would say this is a simple example I know but imagine meta would would ship their llm with this content this is the shape of the moon or if you want to make it a little bit more spicy for my friends in the United States of America you ask me what are the achievements of the last three presidents and maybe you do not agree with what you get back from the model from the llm by Elon or maybe you don't agree with the llm from Mark so you see sometimes you might find yourself in a position you want to change some particular content because you don't agree and you don't want to have the system in your system and if I say hey I believe in science and I think this is the shape of the moon now you understand how to overwrite knowledge in an llm and we have some beautiful ideas now I know if I go out on the street and I ask 99% of the people that I meet the standard answer I will receive is Rag and you're right just look at the video just three months ago we had a deep dive here into the anthropic implementation of a system with contextual Retriever with prompt cashing and everything great but you know what time moved on now at the very last day of December 2024 here we have a brand new publication and it is now about multi-agent filtering of retrieval augmented generation because you know what they write here just days ago the existing rack system frequently struggle with the quality of the retrieval documents as irrelevant or noisy documents degrade the performance of the system increase the computational overhead and undermine the response reliability with rag so they say hey rag is so we have to improve it and had a new idea and of course they go 2025 with multi-agent filtering of retrieval augmented generation and if you read my post here on my channel here just 12 hours ago I I give you more details about it and I said hey this is actually quite a good idea so you have in addition to the re indexer and and ranker and whatever you have now you add another three specialized highly specialized EI system specialized agent to improve direct performance and you might say hey what a timely coincidence that just two days ago you presented us a new form of agents that are so easy to code look this is it this is the code agent this is the complete code and then just yesterday you might say we showed us here to build multi AI agents system that are now absolutely perfect now we understand how to build this better rack system but you know it's not about the or the retriever here it is also about the augmentation if you still want to go with frag because think about it the augmentation phase it augments your original query with the extracted external data so the shape of your prompt is now simply the query and more relevant information and knowledge and you know this is what we call in context learning but now think about this if I now want to change here a particular date a particular fact here in the predefined knowledge of the llm from Elon or whatever and I want to change the date of birth of Mard in this particular llm I can easily say Hey you know I find out this is an incorrect date please insert the

correct date now normally the llm will not just modify this because you see if I simplify a tensor structure here the tensor weight structure here with a simple Network we just change here one single point of effect but you know what if we now say hey I want to change the achievement of mod into llm I now change here a complete sub Network as complete sub grow graph and this subgraph it is not perfectly integrated in the rest of the graph of the history of the time of moard of the achievement of Mozart where Mozart was living the people Mozart Matt cooperated if this does not fit I have a system that will hallucinate as hell so therefore the challenge of today is how to find the perfect mathematical methodology if I want to change the content of a predefined llm how I do this that I still have the complete reasoning pattern of the old llm and everything was trained for I don't know how many millions of dollars but I don't just want to change here a particular port a particular sub Network that is highly interconnected if you think this is the reasoning connect home how I change subconnect homes with other data so you see we are also in a rag system but you know we want more so if you switch here between this and this you see we have to care about the network topology how we do this in a operational methodology now you know I have here multiple videos and in context learning normal fine tuning and supervised e Plus or scaling H reasoning with multicol research for ICL for small particular llms or how we do a visual in context learning with the new methodology by Stanford but today today we have some new research we can do it even better in context learning so what we want we want to have to take here a standard l from Elon or whatever and we say okay I'm happy with this I just don't agree that the predefined knowledge I want to change here minimal Parts here of the network so I want to override this knowledge with my belief what I want to promote here with my llm and I don't want to start a new pre-training phase I don't want to start fine-tuning I don't want to invest any money I just want to do it here if it's a black box show me the methodology to do it here new money black box approach this is what we're going to talk about now you remember when we had to look at this paper here the geometry the geometry of categorical and hierarchical Concepts in llm and you remember they talked about immediately formalized the representation of categorical concept as polytopes in the representation space and we used them the formalization to provide a relationship between the hierarchical structure of the concept of the llm and their geometry of their represent ation in their tza networks this already was amazing but now we build on this inside or do you remember this study here October 2024 not all language model features are linear from MIT and we talked here about we believe that uncovering the true nature of models representation is necessary for discovering the underlying algorithm to use those representation so there are something about this representation that we have to talk and we have to have a deep dive into representation to understand those representation remember this particular paper that we look at Harvard University MIT and Cornell University we spent quite some intense discussion on this for the world model remember building generative models that meaningfully capture the underlying logic of the domains they model would be really immensely valuable and MIT tells us our results suggest new ways to assess how close a given model is to the goal we will build on this inside also so if you want to prepare for this video maybe have a look at the paper in detail but the main paper we're going to examine today is here December 29 in context learning of representation Harvard University Harvard University and Harvard University and

University of Michigan on and why it catched my attention yeah we talking about physics intelligent and eii and this is my topic so let's just Define the terms before we start here in context learning ability of llm to learn and adapt new task or new patterns by simply seeing a few examples within your prompt without requiring any update to the llm parameters there's no fine-tuning now representation in llms and code knowledge about the world in the form of internal numerical representations tensor structure Vector structure associated with each token you remember the tokenization and these representations are typically learn during the pre-training phase and their relationship reflect the semantic con con actions remember the content that we want to map into a mathematical space a vector space and you might say wait this is not a real thing this is so simplified hey we can do better we deserve better yes so let's go here scientific the process begins with a tokenization we have our tokenizer breaking down the input ta into smaller units we call the tokens great then tokenizer creates your lookup table using them EMB badings as a reference and since the Transformers don't inherently process here the sequences in order we have our additional our positional encodings here added to the Token imp paddings and then the magic if you want of the Transformer begins no token embeddings withal encodings and then they passed to a series of our Transformer layers and you remember from other videos in the layer we have our self attention the key mechanism have a Transformer and our feed forward Network where we have the nonlinear activation function to really learn the stuff but then then comes the most important point the residual stream in the architecture of a transformer so the Transformer layer output and not just the output from the self attention and the feed forward Network we are adding this to the in initial embeddings to form here the residual stream at each layer for each layer and this maybe you don't see it often in the visualization this is an integral component of the success of the Transformer and it is crucial to understand in context learning so the term representation that we find here in this literature it focuses here on the activation within the residual stream of very specific Transformer layer that we will pick out of the architecture and analyze their content some further information if you would like to know you remember that the vectors in the residual stream of our architecture are contextual ized so this means their values are not solely determined by the initial embedding at the start but they are of course influenced here with the complete context of the entire uh token length context window that the llm processes and the specific layers the activations are extracted from so this means each layer you remember this from the classical architecture of a transformer computes different aspects of the input of the query and that the different layers capture different cont textual information also called features if we build on this understanding you understand immediately that now the residual connections are crucial since in there's no fine-tuning there's no pre-training nothing the tens of Weights are untouched we just have in context learning so our learning mechanism has to be different and our learning mechanism is based on the residual connection so adding the previous layers output to the new transformation ensures that the model doesn't forget information about original input and this helps prevent here and we had the problem of the vanishing gradient this is also called the gradient problem here in very deep newal networks and this was the solution years ago so each layer only learns to change a portion of this particular representation each layer is responsible for adding a small correction or if you want a refinement to

the Token representation in a particular mathematical space building on what was computed already in the previous layers the presentations are therefore not directly modified but added which is why it is important to track where the information is at any given layer if you want to optimize in context learning so this means that the information accumulates throughout here the residual stream remember classical idea shallow layers perform the basic operation and the deeper layers May capture here a more complex contextual semantic relationship and in this way the information learned in the earlier layers is not lost as the token is computed in the deeper layers so this residual connection is essential for complex reasoning just to make it clear this is not a static encoding I got a lot of of comments here on those videos no we are here in a dynamic reorganization of the internal learning process of the llm so the radial shifts in this representation within the residual stream across multiple layers and this is now what reveals the development of the in context learning representations and by measuring that representation I will show you this how we do this at each layer of the residual stream this now this new research now shows us here the gradual overriding of the pre-trained semantics by Inc context information in the de layers and you say hey what here it happens this is where the magic happens the gradual overriding of the pre-trained semantics of your llm by this new in context information that we provide with our prompt and this is now modifying here especially for a complex reasoning task in the deeper layers so we achieved it we found exactly the place where it happens now we have to do an experiment by feeding our llm with a particular chosen sequence of token corresponding to graph traversals I've explained it in a minute the self attention mechanism causes now a reorganization of the token embeddings in the residual stream and this this particular reorganization is now detectable is now visible with a tool and this tool is simple here at PCA demonstrating here the emergence of the graph topology this is heavy stuff and reading this the first time I did not understood what this means so remember the self attention mechanism allows the LM to adjust the residual stream representation of each token based on the relation it infers in the context window but now we know where the learning in the ey is happening within context learning yeah for my viewers who want your visual guide the real interesting question is how is the learning happening in this Transformer within context learning because if we understand exactly how the Learning Happens the inner workings we can now optimize the in context learning so in short short summary short break the representations are not just the token EMB badings a lot of people make this mistake but the evolving and contextualized activation vectors within the llms residual stream of each layer and we will examine each each layer one by one I'll show you this in five minutes so the self attention mechanism is now the main driver for the dynamic reshaping of these representations of in context learning allowing here our llms to learn new semantic rowes in context so not the single point of the date of moort but it's complete Mozart sub Network and this might also explain why the transform architect Ure is so good at in context learning let's come back to the paper and you you remember there was something about physics yes let's focus now on physics so let's do an experiment so very easy here we have figure one this is the first figure so we just start with this paper we have here a grid and this grid we fill in with words just words that we know that the llm has a context has been learned has been trained on and in this grd we now put a specific code a specific way how we go from Apple bird milk sand sun

plain Opera and it doesn't make sense and this is exactly what we want we want to have new data where we can be sure that the LM has not been trained on Apple bird milk s plane Opera but we wanted the LM learns this new information this new knowledge so you do this then you generate a sequence of token following a random walk on the grid and this is the input the contents of a llama 3.18 B model and then you see when is our in context learning happening now look at this and you might say what is this now you understand immediately those are representation those are representation of our Transformer architecture with different context length 200 400 1,400 and you might say hey wait a minute this grit topology that we see here what a coincidence if I provide enough information enough context length 1,400 you know this could look if if you a little bit Yeah know it like a grid no look also the grid lines this could be a grid Now understand what this is this is not a visualization what is really an object living inside the llm and I get the comment on other videos this is here a very particular projection of the representation from the recal stream activation in the layer 26 of our transform architecture and now the aors tell us as the number of in context examplars increase there's now a formation of a representation mirroring the grid structure underlying the data generating process and you say this is not possible that's what I said so this data generation process it was based on a topology of a grid structure where we defined with a simple random walk here a particular sequence and you want to tell me that I can find inside the llm if I do here an analysis of the maximum variance of particular components with PCA I can find a similar pattern like the grid pattern in uh if you want visualization I don't think so however the paper says hey it introduces here now a graph tracing task just showed you this is it as a controlled the way to explore here this in context representation shift and as you see very short context we have not enough data this could be any geometrical figure but this looks like a grid structure with grid lines now a matrix so if we think about node edges and context the node of the graph are represented by words that the llm has likely seen before during the pre-training Apple bird and so the edges are the connections between the notes defined by the graph structure a square grid a ring hexagonal grid the context is now the input to the LM llm consist of sequences of token representing random walks along the graph effectively specifying the graph structure without explicit labels or verbal description of this structure so this is now an abstraction of learning happening it is not just that single datas are learned but kind of a graph structure is learned and you remember my example with mozard this is now kind of a subgraph topology that we see is now forming inside the llm and this is what we call learning now we have a tool we have here something very particular our PCA to reduce the dimensionality and visualize here the token representation in 2D but you remember the xaxis and the Y AIS are very specifically chosen for yeah the variation in the data so only if we have here the most significant axis of the varation in the data revealing here it will reveal here the underlying structure of the representation of the original data and if you think this is just happening in layer 26 have a look at all the layers 0 2 4 6 to 30 you see that slowly slowly slowly this grid lines are emerging layer 30 is already could be identified as a grid structure you say okay is this just llama what what about we go from a 1 billion to an 8 billion or we go to Google's Jam too with a 2 billion and a 9 billion you see this seems to happen in the Transformer models in general if we have some deep layers interesting let's do another experiment now now we position

them not in a grid but we position them in a ring and then we generate some data this is what should be learned and now we have pairs of neighbors apple and banana this is now a group and tuplet orange and onion orange and onion those seven and six are now a group and you get the idea this is now the data we want to learn in context learning let's have a look at this if we have a very short context length and a larger context length if we look now at the specific layers in the later layers here even if we go here to layer 26 if you have an extreme short context length even 100 does not really bring out here the ring but if you just go to a context length of 400 look here layer 16 already shows you kind of a rink and layer 26 even have some more information here so those are the alteration of the representation in accordance with the context specified semantics the ring structure in our data if you want the hidden pattern in our data and this are now representations that emerge within our transform architecture now let's look at this here the Llama 1B the Llama 3.1 8B the ring structure for a JMA 2 and a Gemma 29b so interesting also other Ms seem to develop this and now the Au to say as we show increasing the amount of context for 100 to 400 leads to a sudden reorganization of representation in accordance with the graph's connectivity and the Au go on and say this suggests that the llms can manipulate their internal representation in order to reflect concept semantics specified entirely in context in context learning in line with theories of inferential semantics from cognitive science this is amazing simply amazing the Deep newal Network organizes in a form that there is if you want a sudden reorganization of a better optimal representation in accordance with the graph connectivity it would make sense no it is something like an energy optimization if you think physics this is absolutely amazing for in context learning and you know we say okay let's do this with a hexagonal structure and yes if you look so again this is the grid structure this is the ring structure and if you look at the different Ms you see yeah it could you could identify also a hexagonal structure a little bit of fantasy you need a little bit of fantasy sometimes but it yeah there is a hexagonal structure so he like damn this is really working and you would say but wait this is just insides from computer science you know and you might say okay but now we really want to understand this phenomena we really want to understand how the learning is happening because we want to optimize our in context learning so we need now more powerful tools to explore here the inner workings of the Transformer in reference to in context learning because now we are in 2025 so you might say come on we now bring in the big guy physics complete mathematical physics but this will be the topic of the next video because I think we already had some excitement today in this video so if you want to change your llm content if you want to override specific knowledge like the form of the Moon to your preferred form to your preferred predefined content you don't want to accept what is given to you by meta or Microsoft in their predefined AI system but you want to make sure that it is according to your facts that you have the beliefs that you have the content that you want to generate then maybe you have to modify the knowledge that is encoded in the standard blackbox llms and you can do this with in context learning and this we're going to explore in mathematical details in the next video I hope you enjoyed it I hope you had a little bit of fun in this video and it would be great if you subscribe because then we will enjoy the next video