FOUNDATION OF DIGITAL BUSINESS

Surojit Mookherjee

Vinod Gupta School of Management

Indian Institute of Technology Kharagpur

Week 06

Lecture 29

Lecture 29 : Gen AI -Rethinking the Enterprise Agenda

Good morning. So, in this module on generative AI—the next productive frontier—I will be talking today about rethinking the enterprise agenda, and then I will also discuss some elements of enterprise AI strategy. So, we are basically talking about how to introduce generative AI to industrial enterprises or commercial establishments. So, how do we start the journey?

So, hence, we have to start by thinking about how we rethink our entire enterprise agenda and what the components of that would be. So, we can break it up like this: as you can see, there are three steps to the future. Foundation models will fundamentally change the nature of digital interfaces enabling richer textual, voice, image and video interactions as multimodal inputs are now an accepted form of any data, literally. So, be it video, image, free text, or whatever, it is data for us that we can utilize to exploit the foundation models.

Knowledge bases will allow the development of intelligent autonomous agents that will unlock massive productivity gains. So, we are now talking about agentification. So, just an output, a prediction, a classification, or whatever is not good enough for us—we want to take it further. So, that is our journey of automation. Can that tool, can these models, can these AI models do the job for us?

So, let them predict something and also execute it. We will talk more about this in the future slides. GenI will be embedded into all applications enabling a move towards a new computing architecture. So, you will find literally GenI will be a component of any enterprise tool for example. So, you must be using or you know about heard about say

CRM tool, customer relationship management tool or an enterprise resource planning tool.

These are there from several years, decades. But now they are modernizing, they are advancing, they are augmenting those products with embedded gen AI architecture. So, that is adding lot of value to these products which we will gradually see. So, to begin with what is really meant by a foundation model. It is a large scale AI model trained on vast quantity of broad unlabeled that can be adapted

data fine tune to a wide range of downstream tasks. So, what people have done, you must have heard about all those chat GPT's and things like that. So, they have trained them on whatever data was available when it started into chat GPT 3.0, I think the date was whatever was available, but they were trained as of November 2021 or something like that.

So, when they are released in 2022 or 2023. So, that was the cut off line till of November 2021, whatever most of the things available on the internet has been used to train the ChatGPT 3.0 version. So, the term was popularized by Stanford Institute for Human Centered Artificial Intelligence. If I say that these models serve as a base or foundation upon which many different applications can be built. So, literally speaking, you can think of it like a library, but all knowledge has been

captured into one tool, and now you can use that tool for whatever purpose you want or where your requirement is. They represent a paradigm shift where, instead of building many specialized models from scratch, developers can leverage these powerful pre-trained models. So, that is why they are called pre-trained models, and adapt them, saving significant time, resources, data costs, money, etc. To give you an example or an analogy, literally. So, you can think of a student who has read and learned from a massive library covering countless subjects.

They have a broad understanding of the world, language, and various concepts. So, you hire that person. He has captured all the knowledge, spent many years going through whatever literature, etc., and he knows most of the subjects. Now if you want the student to become a specialized in say law or you want to become a specialized them some specific legal textbooks and case studies to study,

and this is what in this jargon is called fine-tuning. When you talk about fine tuning the foundation model, this is what is meant that I take a topic or an area and now I give some extra knowledge to this student or the person—or, actually speaking, to the foundation model. So, they can now leverage their broad foundational knowledge to quickly excel in this specialized legal domain.

Similarly, it can be for true for any profession you take this student now can be since he or she knows in general most topics what human beings know and now you can make him a specialized lawyer, doctor, engineer or some extra specific domain knowledge, which in computer terminology we call fine-tuning of the So, the characteristics of foundation models are: they will be large-scale and trained on broad data. So, they are pre-trained on diverse and extensive datasets, which can include anything on the multimodal inputs like text, image, code,

codes, videos, audio files—you name it, anything can be used. It is self-supervised learning. So, these are not labeled data. It is not possible to label so much of vast things and train a computer. It is self-supervised learning.

In this approach the model learns to predict parts of the input data itself without requiring very explicit human provided labels for every piece of data. For example, a language model might learn by predicting the next word in a sentence or filling in the missing words. the transformer model where you give an input and then you expect the computer to predict the next word and so on. So, it goes on building a sentence, then a paragraph, then whole set of text scripts. Adaptability and transfer learning once pre-trained, the foundation models can be adapted or fine tuned to perform specific tasks with relatively small amounts of time.

task specific level data which I told you about just in the previous slide. Generative capabilities found particularly LLMs large language models and image generation models can create new content, text images code etcetera that is similar to the data they were trained on. So, you can ask a LLM model to write a story or a script, film script or drama script. create an image, create videos. So, we have gone to that level and the video qualities are improving so much day by day that you can literally do a complete movie using these LLM models.

So, give them the prompt and give them the basic ideas that model will create the entire script, will create also the video and almost like a human like video this level it has come today. Emergent properties due to the scale and vastness of the training data. They can exhibit emergent properties in the sense they can tell you some extra information which they were actually not trained over. So, this is what we are now getting into the we are saying that computers are becoming more and more intelligent because as an almost like reaching that human behavior just I can answer and add something new out of whatever my cognitive skills are.

Some examples of foundation models, you must have the GPT series, GPT-3, GPT-4 by OpenAI, they are known famous for the strong text generation and understanding capabilities. BERT was developed by Google, bi-directional encoder representation from transformers. The language model primary used for understanding the meaning and the context of the text so that they can go on predicting new word to complete the sentence, paragraphs, etc. DALI and DALI2 by OpenAI, Stable Diffusion by Stability AI and Imagen by Google these models that can generate images from text descriptions. And recently we are hearing about a product called Sora also from OpenAI is a text to video model.

And the model generates short video clips based on user prompts and also extend existing video So, they are coming up with very real like even Google Gemini has also come up with the news that their video qualities are also very real like. The AI agents about the automation part which was referring to. The democratization of creativity will change the nature of a relationship with the work. Now, if you imagine these tools enabled at scale in an enterprise, relationship manager gets a well structured talk track for a client based on analysis of client data and recent interaction.

So, you want to do a client interaction, client delivery, you want to give a talk for a business deal and you get your tool to prepare your lecture notes from based on whatever client inputs you give client data and certain recent interactions from emails, exchanges etcetera you can share feed to the tool and the tool will come up with a deliverable in the sense what your speech should be. The artists creating their own movies as generative production tools are proliferating. So, this will have tremendous impact good or bad.

So, we will not discuss that part. But, this is what is the now the technical capabilities have reached. Researchers and drug companies are using protein sequencing data to create new drugs. Today all the proteins in the world have been sequenced, the structures have been discovered the help of open AI and that is why in 2024 one of the two Nobel laureates in chemistry was an AI scientist.

And same for physics, in physics if you find out in 2024 one of the Nobel Prize Godfrey Hinton who is known as the Godfather of AI was given a Nobel Prize in physics, but he is not a physicist he is an AI scientist. So, that is you can imagine what is the impact AI or generative AI or deep neural network are playing a role in today's everyday life, business and also big time in scientific discoveries. So, they will be used to discover drugs, vaccines, new materials, nano materials, even new completely new molecular formula etcetera. So, they will unleash innovation at an unprecedented scale.

So, it is actually we are unable to imagine also to what scale it can go to. So, that is why people are very optimistic about these tools at the same time there is a other side of it will become too powerful for us to really manage and control. Because there are all kinds of people in the world and once we have such a powerful tool it can be put to certain wrong uses as well. So, that always fear factor remains and we should be very concerned about this negative sides of the tool. So, what we need to do is practice what we call responsible AI, which I will cover again in my future class.

Now, back to the topic: rethinking the enterprise agenda. Organizations will benefit from putting in place a JNI strategy that integrates with their overall approach to customer engagement, digital operations, and technical architecture. So, the first thing you should think about is where to apply is think about the customer engagement area, that is where you can get immediate benefit, improve customer satisfaction, and directly benefit your business. The technology is nascent and fast-evolving.

Keeping abreast of the changes is critical. Proving success through agile experiments and then scaling with full functional automation is critical to realizing benefits. Now, this is the place where we are stuck today. So, all sorts of pilots are being tried out, but scaling has yet to happen to the desired extent that is required. Many things are there; we will gradually discuss some of these, and you will come to know why scaling up is still not keeping the right pace.

Key decisions relate to the choice of LLMs and cloud vendors, integration with the digital platform and the enterprise data ecosystem, and the right security architecture. So, today we are using many enterprise applications, like ERP—for example, SAP ERP—for managing the business. So, that is creating a lot of transactions, as all your transactions are done through those tools. So, all your transaction data, master data—all of these reside in your enterprise data system, along with your CRM software, etc.

So, all of the softwares which you use are repositories of all the entire companies data. So, that is one major source for your LLM application. So, integrate these data systems with the other digital platforms whatever you are using like cloud or service or other your enterprise your improvised data centers and with your cloud vendors if you are using cloud and then with the LLM of your choice.

the open source LLM whichever you have selected of your choice and of course, keep in mind the security architecture. So, that is one of the major thing which you must keep in mind because whenever you are going to cloud, when you are going to use these models, you will be exposed to lot of security vulnerabilities. JNI also speeds up software coding by converting natural language instructions to complex code. So, we are talking about no code, zero code regime. Today literally you can talk to the computer and the computer will write out the software code.

They are still in of course, in a nascent stage, but these simple programs can be written. So, you can imagine quickly imagine that all of us are literally becoming software programmers, because I can talk I can tell the computer what I want and then the computer gives the code for me. I may not be skilled in programming, so I may not be able to find out whether the code is correct or not, but I can always use another programmer to test it out and tell me whether this code is good enough or not, but at least the code writing time has got drastically reduced from weeks or months to minutes or hours. So, we are talking about that sort of improvement in productivity. So, the elements of enterprise AI strategies will be.

Aligning AI properties with priorities with the overall business strategy. So, you have to align it with your business strategy. It cannot be a standalone project or venture because then you will not get the desired business benefit which you want, and you are spending a lot of money on your AI experiments. The CXO group must agree on a clear AI policy and deployment goals. So, the company leadership team should be very clear about what they want from this AI application.

Identify the AI business sponsors and transformation leads, and define KPIs. The first thing should be to identify who will lead each of these projects. So, they are the sponsors, which is very important; otherwise, the project will not succeed. Create an investment plan, get board approvals, and allocate funds. Work with business leads to create an AI use case roadmap.

Again, a very important step: we now have to democratize the process. That is, you have to involve all your mid- to senior-level executives to work and find out what the AI use cases you want to select for your initial investment will be. Identify the owners and policies for responsible AI, AI governance, and ethics. So, this will come up every now and then. So, whenever you start an AI venture, the first thing you should keep in mind is about the responsibility, because you will be handling data, and with data comes the fear of bias.

So, whatever the input is, the output will be the same. Biased data will give you a biased output. And now you will open up to public to use the product whatever you have developed because you want to do it for your business and then the public will be able to find out or detect whatever bias is there in the tool and then that can bring in lot of problems for the organization. So, think of responsible AI think of a governance team think of a steering committee or whatever to oversee what is being used for the project. So, that is the part of the responsible AI.

Build by your partner approach could be you can build means you to do it in house through in house talents, you can buy an off the shelf product or you can jointly work with the outsourced partners. So, these are the three modes of delivering projects each has its own advantage disadvantage which it is not part of this syllabus. So, we will take it offline. To continue on this elements of enterprise AI strategy, given the intense scrutiny from the market and board in depth planning and successful initiation of the journey is very So, before you start, before you jump into an AI project, do a market survey, have a intense discussion at the board level and then only you begin.

Organizations can take a dual approach during the initial pilots. One is a bottom-up approach that empowers gen AI champions for the grassroots innovations. So, you can encourage all your employees to start experimenting with the gen AI. Because it is so commonly easily available that you do not even have to spend money because it is available free. So, initial trials can be done in a small scale using this free open source tools to the limited extent they allow.

And the other approach is a top down which is driving it from the top. So, you utilize something called a create something called a center of excellence to design, build and deploy the priority high value and complex use cases and also centrally define the technology stack, governance, talent and risk processes. So, you do it from the top down the CEO decides, creates an CEO, brings in some senior people for the governance and to decide which projects to undertake. decides which technology to use buy or use etcetera and then involves up to a team to do the development and then use it at a company wide level. So, both can be approached top down or bottoms up each has its own merits and demerits.

Effective communications. or the value created must be disseminated across the enterprise to gain buy-in or build support and drive the organizational engagement for broader adoption of AI initiative. So, you must have the communication plan because this is an intense part of change management which will come because of you are bringing in a completely new technology to your enterprise level and which will involve all your So, you must communicate about the project across your board across the company organization. Some enterprises are equipping a significant portion of the workforce with JNI tools to enhance day to day productivity at scale.

These tools encompass an enterprise version of ChatGPT. So, I will talk about in another session about how governments are encouraging their employees to participate in such experiments, because how do you get the buy-in from your employees. So, one of them approach is to give them a taste. of the advantages which these tools bring in their productive work life. So, if you are using GenAI obviously, life is becoming much simpler because it can summarize things for you, it can write things for you, it can draft a mail for you.

So, when people start seeing the advantage they get kind of habituated to these products or the advantage and then that is how you get the acceptance or the buy in from your So, that is the point is that you need to encourage your employees to experiment or play with Gen AI tools for their day to day work. Enterprise need to evaluate and design the right architecture across different technology components, because of all these technology things you need to find out which is the right technology. The first thing is the choice of models for text image, there are plethora of tools available you will get confused. So, here you might need some expert advice from consultants etcetera or maybe your CIOs

organization can do find out talk to people attended conferences etcetera to find out which of these products are really being very useful and used by. your peer group, your competitors etcetera. So, broadly speaking GenAI we classify in two types closed source and open source. Closed source models are the large models enabling them to handle diverse complex queries. They do not require infrastructure that are available immediately for enterprises because they are on their clouds like OpenAI's GPT-4 or GPT-3.5.

or Google's Palm 2, Anthropix cloud, etc. So, there is a cost attached for access and use. Somehow they allow you to use it free, but if you want to use it for commercially obviously you have to pay, but you do not have to have bother about any infrastructure component for this using these tools. The open source models need to be managed by individual enterprises.

They are more controlled and typically smaller in size like Metas, Lama 2, Falcon or Mistral. The less versatile and tend to hallucinate provide incorrect responses more than the larger closed source models. So, both again have advantage disadvantage which you can find out, but this is not the purview of this study because otherwise it will become too long. So, I will not be talking about that, but you can always read about these and find out from the internet. The choice of models is closely intermedia cloud platform providers.

So, OpenAI for instance is available on Azure enterprise and OpenAI's own enterprise cloud while other closed source LLMs are exclusively available on other enterprise cloud providers. So, there are many cloud providers who are hosting such closed source LLMs. To improve accuracy of GNI agents, enterprises need to make models contextual to their prior proprietary data and various tools and techniques are available for enterprises. So, in this lecture session I will be mentioning lot of names, tool etcetera, but I will not be discussing of any of these tools because otherwise it will be too long and too complicated, but you can the names are for reference for you.

So, you can if you are interested you can always read about these from in the internet. The fourth critical step involves around the seamless integration of models with existing enterprise applications which was just now talked about that for your enterprise applications they will need to be integrated to this tool such that your proprietary data your company data can be used to create that small language models. And you can use internal query, so you get the responses relevant to and in context to your company data, organizations data. So, there will be the responses very specific to the organization not generic from across the world.

The effectiveness of AI is inherently tied to the quality of the data it receives. The modern data platform needs added capabilities around data governance and data security. So, we have been talking about this governance and security for obvious reasons, but I am sure by now you have realized what are those reasons. The open source elements are best leveraged for use cases at scale where the token volumes are high. A typical use case of high token volume is customer service quality analysis which require giving all customer call transcripts etcetera.

So, this is again a bit becoming bit more technical. So, we have to again decide on the work load to decide which module. So, we have to because they are very power hungry.

So, we have to so that is one of the problem with using AI models etcetera all these chat GPT general gen AI models is that they large language models they consume lot of compute power. So, the more complex the more input load you give through all these tokens etcetera the more calculation it will have to do and so more time and more power.

Closed source LLMs which are the proprietary ones the smaller ones are ideal for getting started pilot and rapid experimentations. So, these are better suited for deployment of use cases that required versatility and ability to handle complex conversations like financial planning fraud investigation with low to medium token volumes. So, in summary the open source LLMs like CHAD, GPD they provide transparency, customization and freedom while the closed source LLMs will prioritize performance and commercial interests. So, that is how these products are because since they are being so powerful and so commonly used they are being classified in grouped in these two ways for commercial exploitation by the companies.

The difference between open source and closed source further. is accessibility and adaptability. Open source LLMs are freely available for modification, distribution and use while the closed source models are proprietary in nature and they have limited access. Customization and flexibility is again open source models offer high levels of customization and flexibility. Open source models offer greater transparency as the source code, model architecture, free trade weights are publicly available.

For support the closed source models often come with a structured support system. and ready to deploy features. For collaboration open source models encourage a collaborative approach to innovation. For data security closed source models may come with enhanced data security features and cost wise open source models are usually less expensive than closed source models. So, equip with the AI workforce.

So, this will give you an idea of the type of skills or talents you will need to manage such AI projects in your organization. So, you have something called data scientists, data engineer, DevOps and cloud engineer, software developer, data analyst, business analyst, solutions architect, legal and risk and compliance, data security and protection. So, this will give you an idea of the type of resource you need to manage your LLM projects. They could be in house or they could be you can collaborate with you can outsource that some organizations have this set of people they will do the work for you.

So, that is why those models are three options are buy, build or partner. So, these are the three models which we work. So, buy is when you are buying the complete product and when only the support will be provided by the vendor and build is if you want to build you really have to employ so many skills in your organization. Shielding with a responsible AI.

Trust and performance risks, hallucinations in LLMs lead to erroneous responses and erode user trust. So, this is one of the major reason which is preventing scaling up of JNI models because of this hallucination factor because you do not know what the output will be. Bias and toxicity risks, security and privacy risk, regulatory compliance and copyright The copyright risk is a big risk because these models have been trained on huge amount of literature, information whatever intellectual properties available in the internet. Now your output, so your model is using those knowledge and if somebody raises a complaint that

my copyrighted article has been used to train your model and because the output is referring to that particular whatever and has not taken my permission. So, that could be a major copyright issue. Same is true for your paintings and other work of art or literature or even music audio files. So, they have all under copyright. So, you have to be very careful about this copyright issue.

So, enterprises must stay informed about AI governance, ethics, policies and regulatory provisions. You cannot say that I was not aware of it or I bought the model and the model has been trained on this. So, I was not involved etcetera etcetera. If you are using it then you are responsible for whatever the output is delivering. Ethical risks, enterprises must navigate ethical concerns related to job loss, technology misuse, for example deep fake.

So, be careful you have to be responsible for whatever the output is being generated by the tool you are using for your organizations for all of these features. The recommendations say to suppose you want to think country like India, what India should do? Access to training data and marketplace. is key for development of AI systems. The government support would be needed to ensure that researchers, enterprises and startups have access to structured and unstructured datasets.

As we know data is a vital thing starting thing. So, can we have the government helping all the researchers, developers with access to large volumes of datasets. Deployment of JNI systems as public goods. So, India has a history of developing successful digital public goods such as India Stack, Aadhaar, UPI, etc. And building on that success the government may consider developing and deploying JNI algorithms as public goods.

So, government of India is already sponsoring a public cloud and buying all the GPUs because otherwise getting GPUs is also a big challenge because of the huge demand and short supply. So, the government is doing for that and investments are expensive. So, if government has a cloud for users to take it on rent. Securing critical digital infrastructure. So, availability of computational infrastructure like I was talking about the cloud government owned cloud.

Securing the technology supply chain is imperative for the development of deployment of AI because you need a cloud otherwise how do you start and you cannot always build one. So, that will be what we have very important barrier to that and access to talent and public funding of R&D. So, funding is required sponsorship is required to develop to come up with the innovation and initiatives by the developers and scientific community. To conclude this session, the government initiatives laying foundation of growth, some of the challenges to realize the full potential of AI can be low intensity of AI research for various reasons, some of which we have discussed right now.

Core research in fundamental technologies, transforming the core research into market So, that is where the scaling-up thing is becoming a bottleneck. Inadequate availability of AI expertise, workforce, and skilling opportunities. So, the talent agenda. So, do we have adequate talent or adequately skilled talent?

High resource cost and low awareness of adopting AI in business processes. Awareness, unclear privacy, security, and ethical regulations. So, we need to have regulations like GDPR and others in the European communities, even with AI, even in the US. So, we also have to adopt such regulations to control the whole market. Unattractive intellectual property regime to incentivize research in AI.

So, intellectual property, as I was talking about, is also another. Hindrance kind of thing to incentivize research in AI because you have to be careful about that. So, with that, I would end this session. Thank you very much.