Undergraduate Project

Skill mapping for the construction industry using Knowledge Graph and Generative AI

Submitted by:

CE19B028 – Ankit Sanghvi

Guide

Dr. Nikhil Bugalia

Department of Civil Engineering

Indian Institute of Technology, Madras

August 2022 – May 2023

Table of Contents

Acknowledgement

I would like to express my gratitude to my research guide, Professor Nikhil Bugalia. from the Building Technology and Construction Management Division of the Civil Engineering Department, for his patient guidance, enthusiastic encouragement, valuable connections and useful critiques of this research work. I would also like to thank Professor Anil Prabhakar from the Electrical Engineering Department for his invaluable input and direction in the early days.

I would also like to thank all the corporate training team and staff at Larsen and Toubro and HyperVerge, for providing me with the necessary resources and data to complete the necessary experiments, as well as for assisting me in conducting tests in a correct and precise manner.

I would like to thank Santhosh Loganathan (Faculty at L&T Institute of Program Management) and Gayathri Meka (Program Lead, HyperVerge Academy) for their direction at a time when I was still learning my way around the space and conducting research interviews of users. I also greatly appreciate Shruthi Apalla and Ruchira for aiding me in solutioning review so that the results could be compiled in a timely and organized fashion.

Finally, I would like to thank Jayanth, Subham, and Prabhat, who were always rooting for me and gave me a ton of moral support. The biggest thank you to my brother Pratik and parents - Neelam and Ravi who are the best parts of my day, and keep me going through everything.

Abstract

Keywords: corporate training, education, generative AI, skill mapping, knowledge graph, competency

The rapid pace of technological innovation and the evolving nature of the job market require a dynamic approach to skill acquisition, particularly in the corporate sector. Addressing this need, our project investigates the complex interplay between corporate training, skill mapping, and education, with a focus on leveraging generative AI technologies. Through a comprehensive examination of different professional learning pathways, we construct a comprehensive knowledge graph that captures the multi-dimensionality of skills and competencies required across various industries and roles.

Recognizing the myriad of pathways that exist in the professional learning space, our project explores how to effectively map these paths in order to provide more efficient and directed corporate training. By constructing a knowledge graph utilizing generative AI technologies, our research unearths the potential of AI to comprehend and translate the intricate web of competencies and skills that span across industries and job roles. Our investigation not only establishes the intersections of skills within and between sectors, but also extrapolates the direction of future skills demand, thereby enhancing the capacity to tailor learning interventions to the evolving professional landscape.

The culmination of our work is the development of CourseGPT, an AI-powered chatbot integrated into a learning management system. CourseGPT is uniquely positioned to assist learners by recommending courses aligned with their career goals, aiding in planning academic journeys for specialization, and providing real-time assistance to resolve doubts during the learning process.

The development of CourseGPT embodies a significant step forward in our mission to reinvent the corporate training paradigm. As a result of our comprehensive research into skill mapping, competency-based learning, and education systems, we've been able to engineer an AI chatbot that delivers personalized learning recommendations, immediate doubt resolution, and academic journey planning. Additionally, CourseGPT breaks down language barriers that often hinder learning opportunities, offering support in all Indian vernacular languages and providing voice assistance. As a multifaceted tool for professional development, CourseGPT has the potential to significantly improve the efficacy of corporate training and lifelong learning, marking a major breakthrough in the application of AI technologies to education. By offering a more tailored, responsive, and inclusive learning experience, CourseGPT stands as a testament to the transformative potential of AI in professional learning and corporate training.

Problem

Construction companies at large scale have diversified projects across various climates, demographics, economies & other variables. Hence, the planning of internal educational training programs, in domains such as quality, safety & efficiency is of paramount importance to make the best use of the talent in the company. Also, a promising growth trajectory for an employee is an edifying experience that earns their fidelity to the company, thereby maintaining a strong company culture. According to LinkedIn's 2018 Workforce Learning Report, a whopping 94 percent of employees admitted they would stay at a company longer if it invested in their careers.
Current systems of planning these programs are ineffective in covering the breadth of possible skills & roles, & a set of education plans taken to achieve them. Companies that build internal knowledge management systems are unable to track the effectiveness & ROI of each such initiative due to lack of a system that helps them understand the accumulated learnings in a company from a holistic standpoint - what skills are already present in the company? How can different employees be upskilled with different programs so that they can best manage upcoming projects? Another study conducted by the Association for Talent Development (ATD) found that companies that invested in formalized training experienced a 24% higher profit margin than those that did not.

Overview of domain

Employee training and development is a program that helps to learn a particular skill as well as knowledge to improve employee productivity & performance in their current organization or job role. It develops future performance & helps focus on more employee growth.

Companies that have a proper training and development process can retain more employees, see higher profitability, and have more engaged employees. Furthermore, it helps the organization avoid the costs linked to losing talents. Training and developing an employee doesn't simply help their growth but pushes the company to grow as well. Likewise, it helps employees know that they are valued in the organization.

A formal definition of training and development is "*Training and development is an attempt to improve current or future employee performance by increasing an employee's ability to perform through learning, usually by changing the employee's attitude or increasing his or her skills as well as knowledge*" An organization that nurtures its employee skills and thinks about its growth certainly attracts better talent and sustainability.

Steps to ensure an effective employee upskilling initiative

Objectives and scope

The objective of this project is to help construction industries:

- Understand the gap between the current skill set of employees and plan effective roadmaps for employees to fill roles that are required now or will be required in the near future.
- Plan and track employee education programs more effectively. Track ROI & spend on these initiatives.
- Increase employee retention by ensuring holistic growth of employees

Benefits of planning effective employee education programs

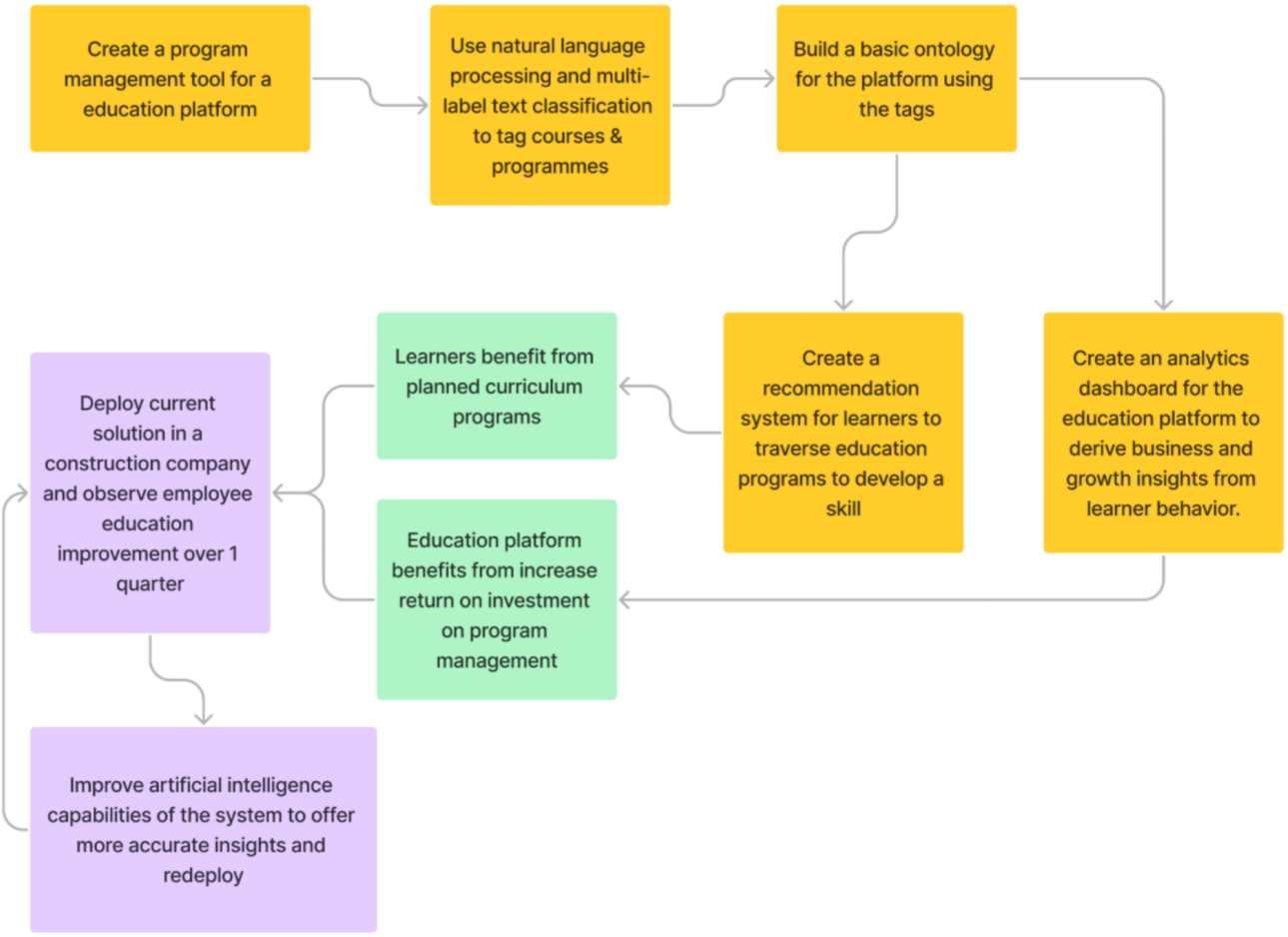
Organizations prosper under continuous learning. The winning advantage is often with the organization that adapts not also changing scenarios but welcomes the development.

Several aspects can be considered trainable under various departments and verticals. However, once the gap is understood, individual training can be designed to affect a solution. Furthermore, a feedback or evaluation system is mandatory to create a

successful training program closely.

Flowchart of methodology

The proposed approach will be carried out in 3 phases
1. Creation of a knowledge management platform and knowledge graph for the education platform - NPTEL
2. Stabilize the system and monitor the benefits of intelligence from learner data
3. Deploy solution in a construction company and improve return on investment in employee education program



Proposed flow for the Project

Timeline

Phase 1	
July - October, 2022	• Create a basic knowledge management platform for IIT Madras academics and deploy. • Observe student behavior and provide valuable analytics to the academic section. • Literature review and Research on natural language processing and knowledge graphs • Work with NPTEL national education team to formulate creation of skill-map over their course content data
Phase 2	
January - March, 2023	• Creation of a basic knowledge graph using NPTEL data • Research on 1 MNC's knowledge management system and integration • Create the proposed solution for the MNC and deploy the system • Build a competency module to track Return on Investment on employee education programs.
April - May, 2023	• Improve and generalize the system to span across multiple companies and platforms • Improve the accuracy of machine learning model • Develop final integrated solution • Draft final project report
Final Submission and Evaluation - May, 2023	

Impact Parameter

According to one industry report, U.S. companies spent over $90 billion dollars on training and development activities in 2017, a year-over-year increase of 32.5 %. Udemy reports that in 2021, employers with wellness programs enjoyed a $3.27 ROI

for every dollar spent and employees gained over 10 hours of productivity. India, as a nation invests INR 13000 crores every year into higher Indian education. The vision is to empower institutes & companies to launch more impactful educational programs.

Phase 1 - July to November, 2022

This phase was further broken down into 5 main sub-phases
1. Case Study - Larsen & Toubro - Site visit & user interviews
2. Literature Review
3. Picking a test sandbox environment of similar nature
4. Development of solution & deployment
5. Users & Initial feedback analysis

Let's now look at each part of the phase in detail

Case Study: Larsen & Toubro

Larsen & Toubro invested INR 14 Lacs in the previous quarter to build supporting software for employee training management. This involved maintenance of an internal Knowledge Management System & uploading content. However, the knowledge ingestion from a user perspective was not worked upon & hence, as the trove of knowledge increased, traversal of the same became a difficult problem

To understand this better, a conversation was organized with the Institute of Program Management (IPM) at Larsen & Toubro Chennai office. The emphasis was to understand the Independent Companies (ICs) in L&T and how the employees in each IC were trained to do their job better. The Graduate Engineering Trainee (GET) Programme is something each fresher in the industry must go through, to understand the breadth of roles in L&T. In this, they undergo a training program - comprising of various roles in various ICs ("job rotation policy), and at the end of 12 months, based on their performance in each IC program, they finally get assigned a role at one

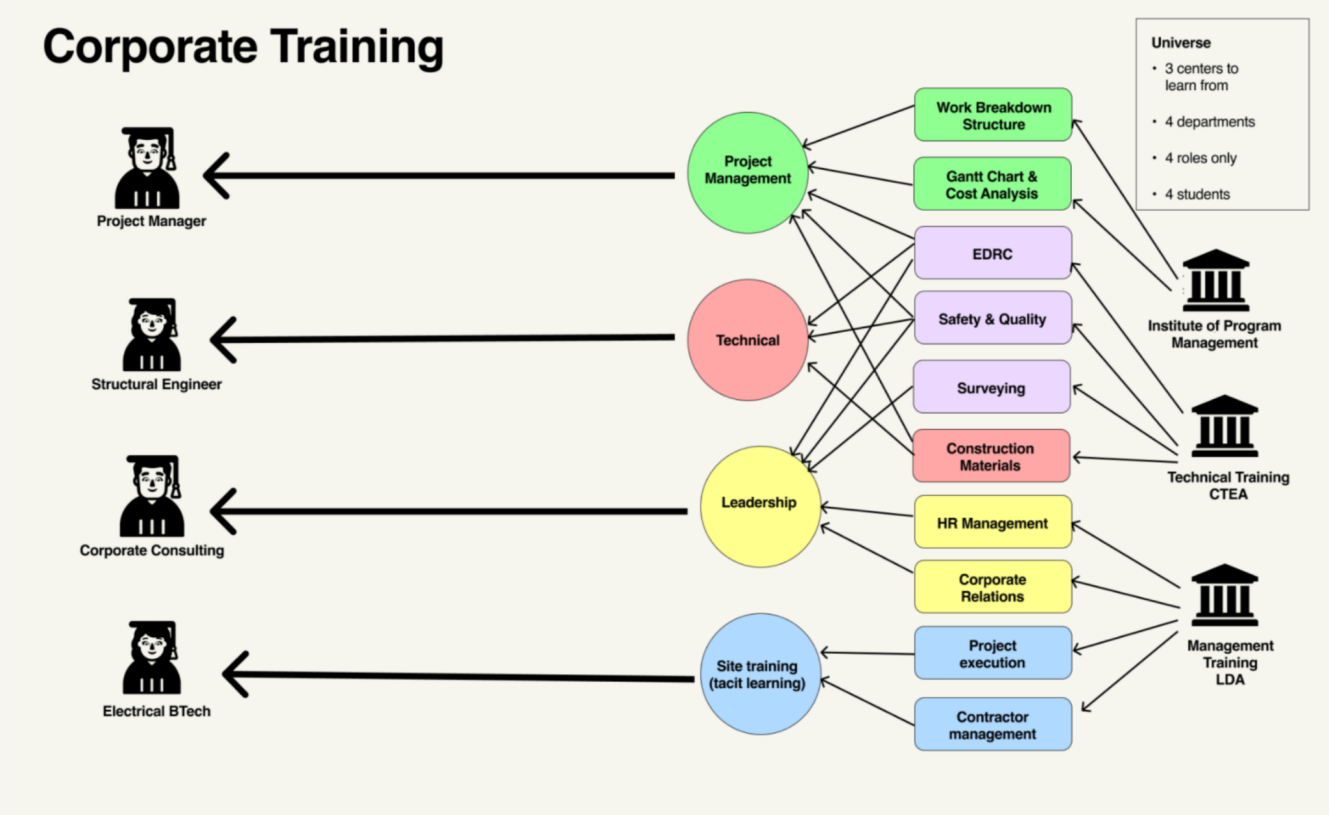
Apart from day-to-day training & improvement programmes, L&T also has 3 academies for specialized learning in 3 key fields

1. **Learning & Development Academy (LDA)** - Lonavla, Maharashtra, India : This academy focuses on leadership, corporate relations & human resource management learning.
2. **Corporate - Technology & Engineering Academy (C-TEA) -** Mumbai, India : This academy focuses on specialized training in engineering solutions - mainly civil, mechanical & electrical engineering

3. **Institute of Program Management (IPM) -** Chennai, India : This institute is the key center for learning project management, estimation & analysis

These 3 centers help edify the growth trajectory of the employees and give them a chance to specialize in a domain of their choice. This is where the "abundance problem" starts to arise.

When the number of courses & programs increases, the consumers don't know what to choose & the providers don't know what & how much to make



Corporate Training in Larsen & Toubro

Faculties, Graduate Engineering Trainees & the IPM Program Lead were interviewed with regards to understanding the top 3 problems that they faced in terms of program management. The answers all resounded with the word "roadmap" & seeing how a set of training courses done today would not only affect the employee's career prospects but also, the company's future workforce efficiency & innovativeness.

The need to reskill oneself is a mandatory requirement to stay relevant & employable in today's fast changing world. The corporate training industry market will reach $487 billion by 2030. The skill mapping software this project provides will help

companies & institutes launch, track & make the best of educational initiatives.

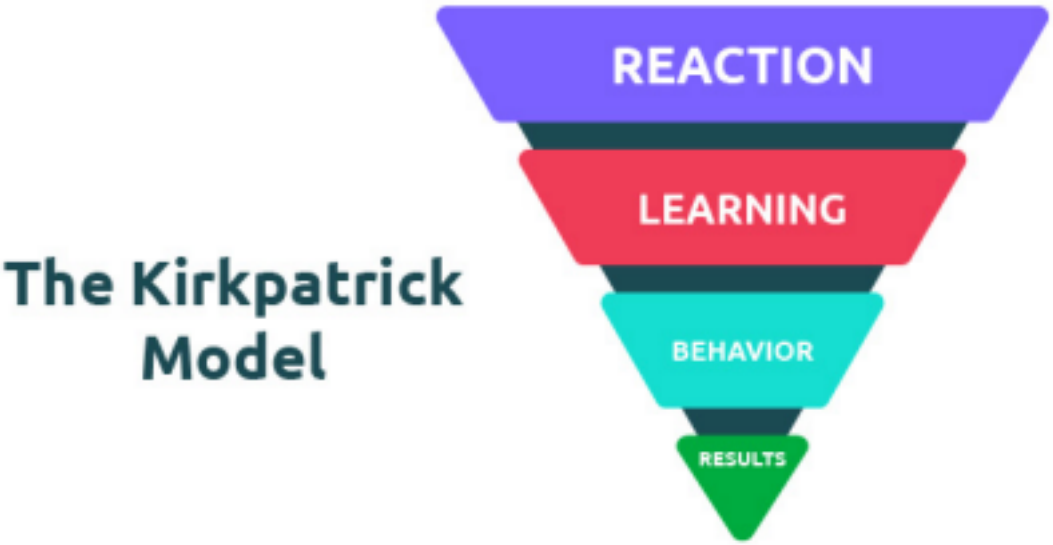
Literature Review

This problem is generalized not just to construction, but to any industry. However, its effects are more pronounced in civil engineering due to the sheer scale of the projects that require
scores of people to undergo training. Thus, literature review was done of several papers in the industry to understand existing approaches. 4 most pronounced papers were:

1. Education-to-Skill Mapping Using Hierarchical Classificationand Transformer Neural Network
    - Human skill forecasting is an important problem for any economy
    - The complexity of education data lies in the diversity in possibilities of educational paths that people take through 1 or a combination of education programs
    - Transformer neural network with Natural Language processing helps glean insights from this information
    - Applications - similarity search, occupation prediction
2. Competency based performance model for construction project managers
    - Unlike functional competencies, which measure performance against predetermined minimum occupational standards, competency-based systems are founded on the key behavioral competencies that underlie superior levels of performance
    - This emphasis stems from a growing recognition of the centrality of the project manager's competency and authority to the performance of projects
    - The methodology used to support this programme ofresearch took its lead from the established "McBer job competency assessment process"
    - Applications - performance management, team deployment & job-matching, recruitment & selection, career development & succession planning
3. Workplace training and generic and technical skill development in the Australian construction industry
    - Employees are generally well aware of the importance of workplace training in their career development and they largely appreciate training as being a critical factor for developing their capacity to perform their roles successfully and to maintain their employability.
    - 2 groups were studied - one from a construction company & the other, comprised of employees from a number of different construction industry-based companies
    - Top skill gaps - leadership skills, industry-specific skills, process & project management skills, managerial skills, communication & interpersonal skills

4. Kirkpatrick Model for Corporate Training Program

- 4 levels - Reaction, Learning, Behavior & Results
- At each level, we have to understand 3 things
    i. Evaluation characteristics & reason
    ii. Evaluation tools & methods
    iii. Relevance & Probability



Kirkpatrick Model of analyzing Corporate Training

The company being studied, Larsen & Toubro, currently maintains job & skill requirements using a competency matrix. Most MNCs have an internal job listings portal where employers list roles & projects & applicants start applying. However, they also wish to connect this with their corporate training system so that there is a skill score awarded to each employee, that makes them more suitable for particular specialized projects. Let us do a deep dive into the Kirkpatrick Model to understand how this is possible

Deep Dive into the Kirkpatrick Model of Corporate Training

The Standard for Leveraging and Validating Talent InvestmentsTM is the Kirkpatrick Model. You can rely on its efficacy since it has developed over the course of more than six decades of use by learning and development experts all across the world.

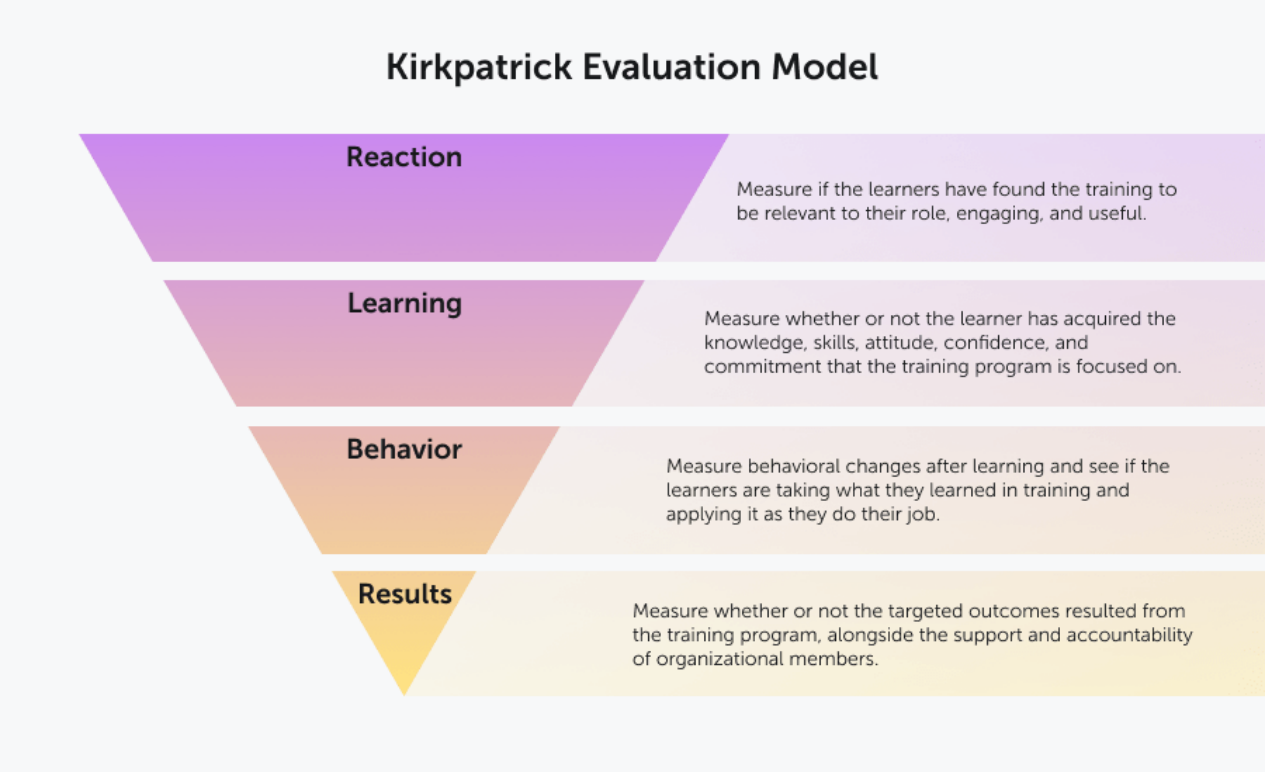
The tried-and-true approach is effective across many industries, including government, the armed forces, business, consultancy, services, and humanitarian. We are convinced that the model will work for you because it can be used with any application. Onboarding, product and programme launches, leadership development, diversity, equity, and inclusion (DEI),
safety, security, and succession planning are all common topics for Kirkpatrick

programme evaluation plans.

Level 1: Reaction

Reaction, which gauges whether learners find the training beneficial, enjoyable, and applicable to their employment, makes up the first level of criterion. This level is typically
evaluated by asking students to review their training experience in an after-training survey, sometimes known as a "smile sheet."

A focus on the learner rather than the trainer is an essential part of Level 1 analysis. The Kirkpatrick Model encourages survey questions that are focused on the takeaways of the learners, even though it may feel natural for a facilitator to focus on the training outcome (such as content or learning environment).



Visual Representation of [Kirkpatrick Model]

Level 2: Learning

Level 2 evaluates each participant's learning based on whether they develop the desired attitudes, knowledge, abilities, and commitment to the course. Pre- and post-assessments should be used to determine accuracy and understanding while evaluating learning, which can be done both formally and informally.

Exams and assessments in the form of interviews are two assessment methods. To minimise discrepancies, a specified, transparent scoring procedure must be decided

upon beforehand.

Level 3: Behavior
Level 3 of the Kirkpatrick Model evaluates if participants were actually affected by the learning and whether they are applying what they have learned. It is one of the most important processes. It is feasible to determine whether skills were understood and whether it is practical to apply them in the job by evaluating behavioural changes.

Evaluating behaviour frequently reveals problems in the workplace. Lack of behavioural change may not indicate ineffective training, but rather that the organization's current
procedures and cultural contexts don't support the best conditions for learning the desired change.

Level 4: Results
Direct outcomes measurement is the focus of the fourth and final level, Level 4. Level Four compares the learning to the business outcomes of the organization—the Key Performance Indicators—that were defined before learning began. Higher returns on investments, fewer workplace accidents, and bigger sales volumes are examples of common KPIs.

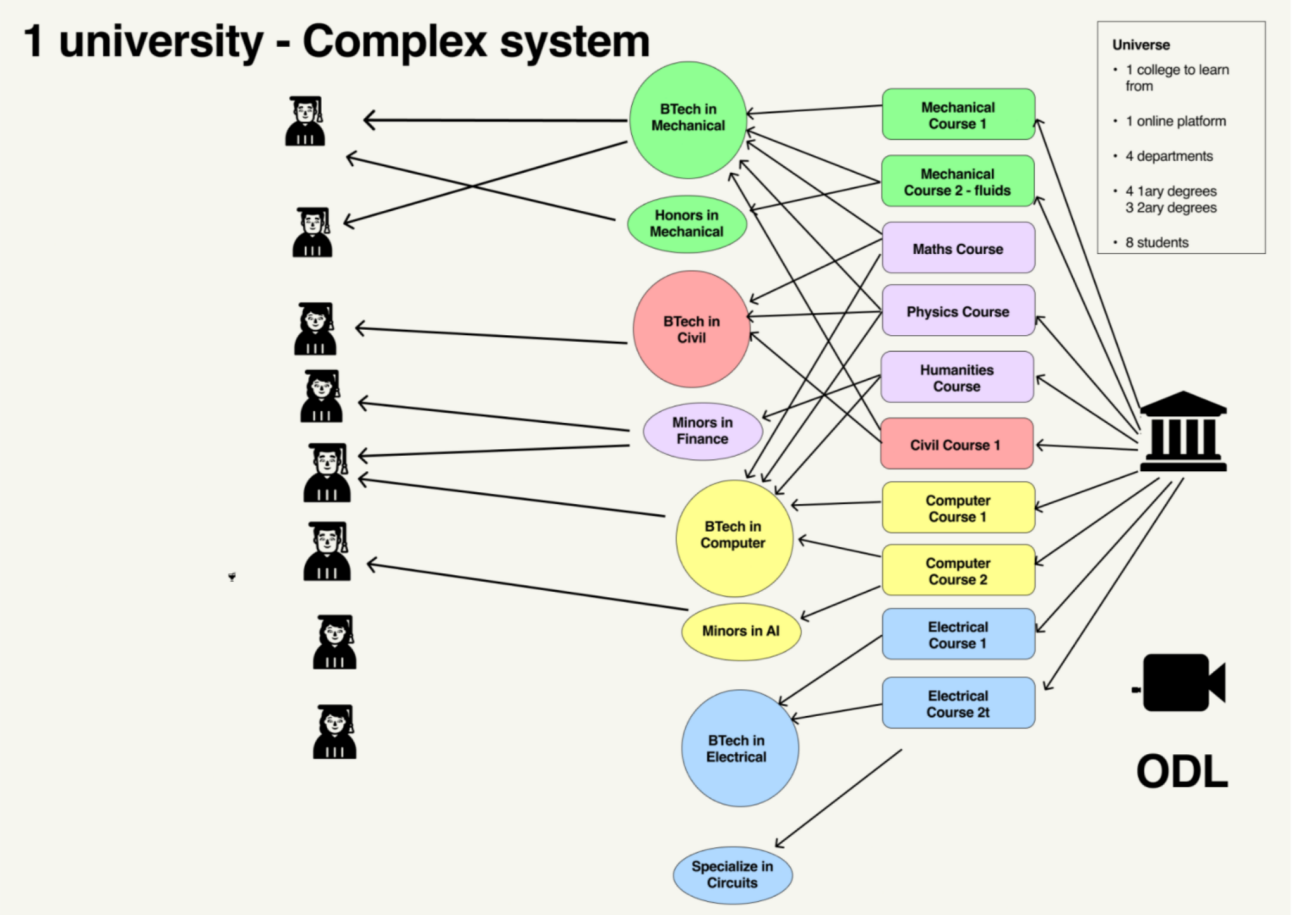
When goals are clearly defined, outcomes are measured, and areas of noticeable influence are found, the Kirkpatrick Model helps establish an actionable measurement strategy. Organizations can evaluate the relationship between each level to better understand the training results by analysing data at each level. As an added bonus, doing so enables organisations to readjust plans and correct course during the learning process.

Picking a test sandbox environment of similar nature

The most similar environment to corporate training is an actual institute - for which IIT Madras was chosen. The hypothesis was 3 pronged:

1. Students are having difficulty in finding course to take - discovery of a course or a path to explore
2. Students not knowing how courses now → affect opportunities later (credit balancing, prerequisite matching, minors & honors degree)
3. Students want to connect with other students who have taken the course before them.

For this, 20 interviews of students from different departments & years were conducted, to understand the problem & validate the hypothesis. (Refer link for detailed user interviews)



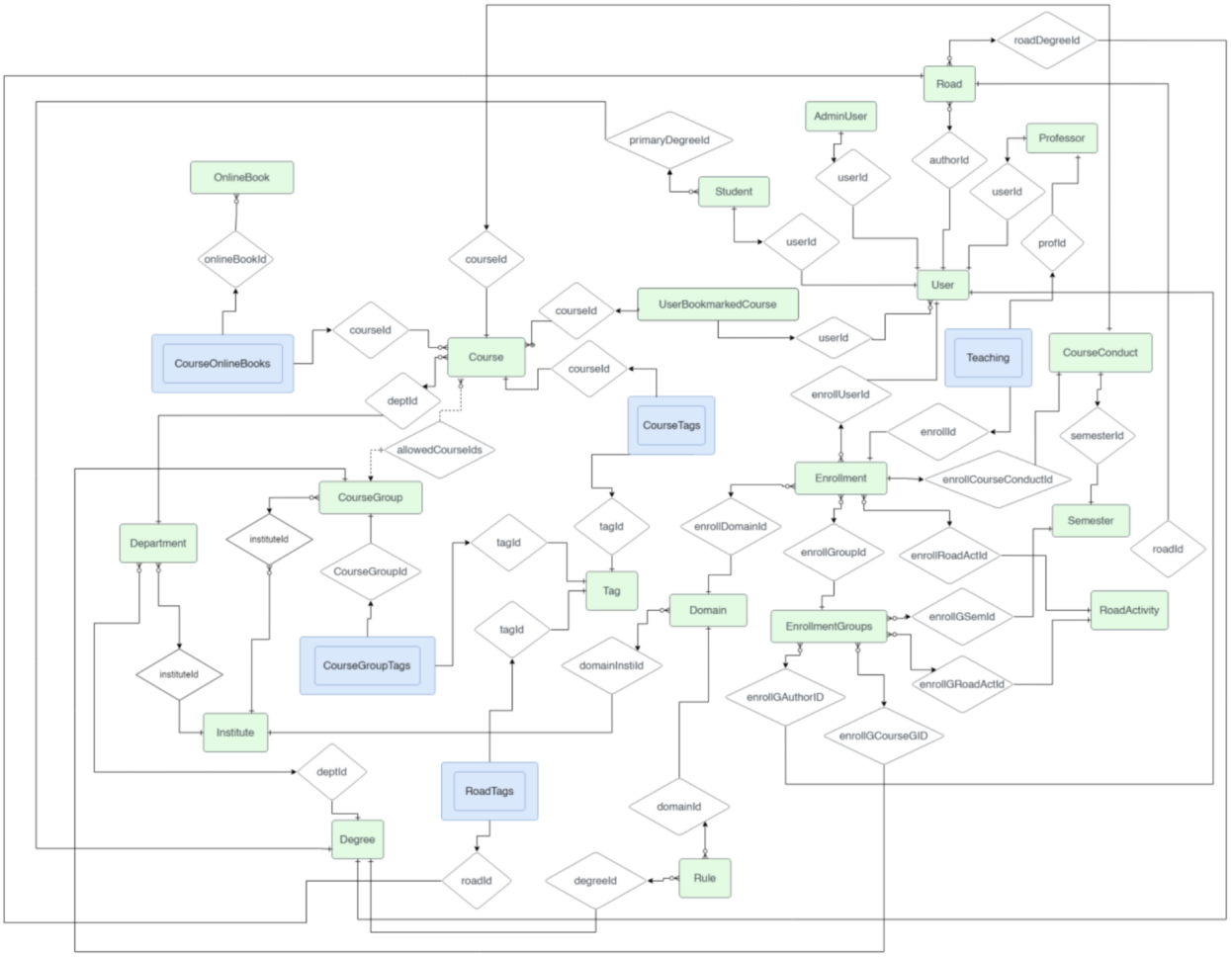
Using IIT Madras's Learning Management System as sandbox for skill mapping

Following this, a set of stakeholders was listed in terms of how the course & academic journey of a student are structured & an Entity Relationship Diagram was conducted to further understand the nature of the interactions & dependencies. Find the detailed Entity Relationship Diagram below.

An entity relationship diagram (ERD) is a visual representation used to model the relationships between entities in a database. It provides a clear and concise way to depict the structure of a database, illustrating how different entities are related to each other and how they interact.

In the ERD below, entities are represented as rectangles, and the relationships between entities are depicted using lines. The entities can be real-world objects, such as people, places, or things, or they can be concepts, such as transactions, orders, or employees. Each entity is defined by its attributes, which are the specific properties or characteristics associated with that entity.

The relationships between entities are crucial in understanding the database design. They can be classified into different types, such as one-to-one, one-to-many, and many-to-many relationships. These relationships define how the entities are connected and the cardinality or multiplicity of the relationship (e.g., one course can have multiple students).

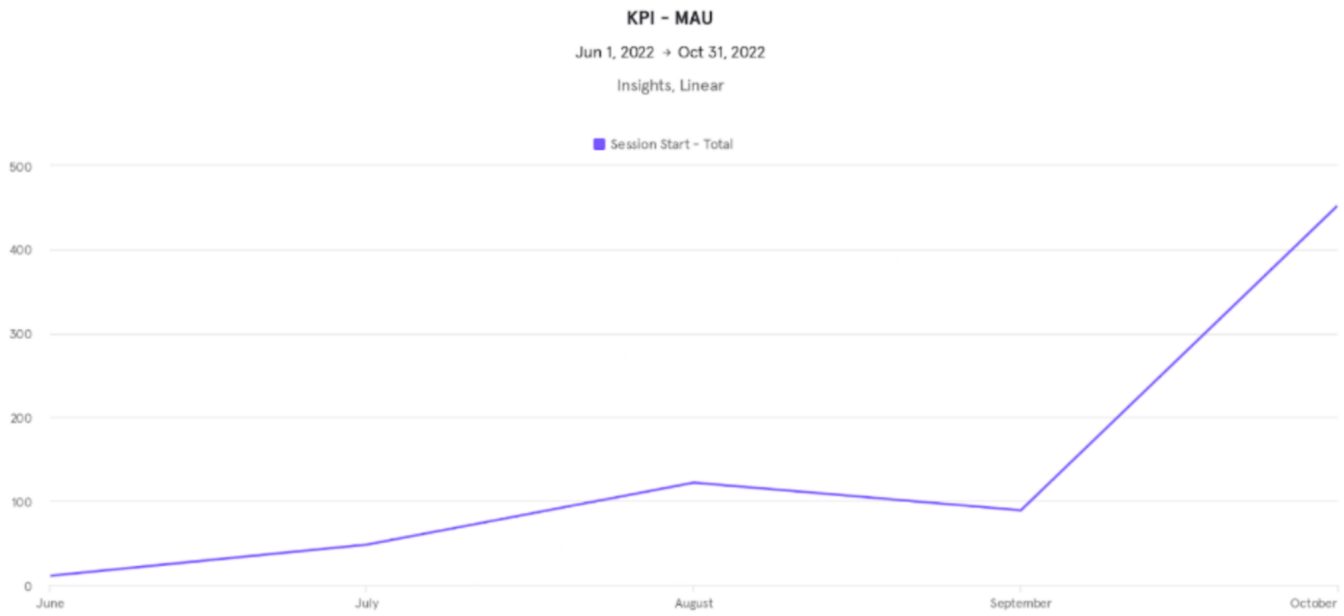


Entity Relationship Diagram of a longer duration knowledge management system

A pilot solution for learners & user feedback

The first solution was launched for students of IIT Madras on August 31st, 2022 by the name CourseMap. CourseMap helps student look at available courses & roadmaps & build their own academic journey. Find a reference video for the usage of the platform at Link. Key features included planning a roadmap across a collection of semesters, understanding in real-time in how specialization paths open up with choices and getting roadmaps peer-reviewed.

The Key Performance Indicator for CourseMap is the students who spend > 5 minutes on the platform for planning their roadmap & discovering career opportunities. CourseMap currently has 473 Monthly Active Users, and this number spikes to 3300+ unique users during the course registration week.



Platform Metric, as of November 28, 2022

A pilot solution for Program Management

Further analysis was done to understand important metrics on educational program management and CoursePlan - a solution for degree management was arrived upon with key features as:

1. Looking at course analytics to understand patterns in student learning
2. Defining & managing different degree programs
3. Create well-defined roadmaps for specialization in degree programs

Reference video for this software - Link

This software was demonstrated to the NPTEL team to improve their service offerings. National Programme on Technology Enhanced Learning (NPTEL) is a project of MHRD initiated by seven Indian Institutes of Technology (Bombay, Delhi, Kanpur, Kharagpur, Madras, Guwahati and Roorkee) along with the Indian Institute of Science, Bangalore in 2003, to provide quality education to anyone interested in learning from the IITs. The main

goal was to create web and video courses in all major branches of engineering and physical sciences at the undergraduate and postgraduate levels and management courses at the postgraduate level.

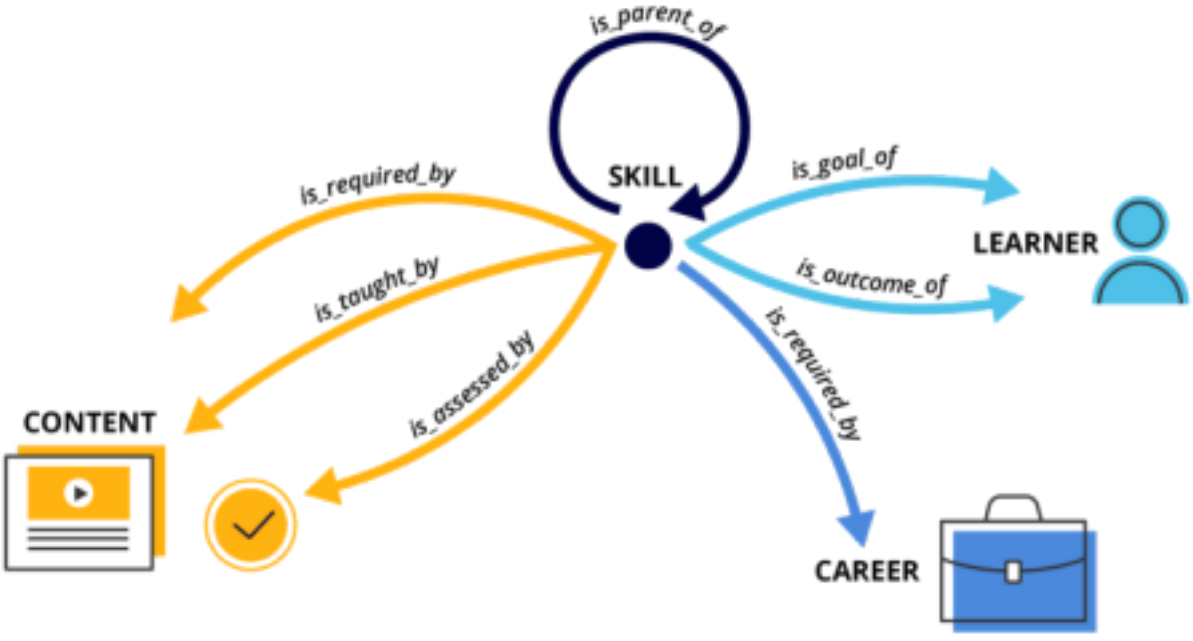
Improving the solution & Comparison Methodology

The solution being proposed in the next iteration of this project is arranging educational opportunities onto a “knowledge graph” aka “ontology” - a graphical data structure that helps glean insights from seemingly unrelated decisions and choices. This would be accompanied by a Software As A Service (SaaS) based tool that helps companies & employees digest the insights from the “knowledge graph” to make more informed decisions

Methodology: Skill mapping using Knowledge Graph

What is a knowledge graph?

A knowledge graph is an interconnected dataset enriched with meaning so we can reason about the underlying data and use it confidently for complex decision-making When you take data and export it into a dynamic structure like a graph, you get a structure that is connected contextually to all of its neighbors and those to their neighbors and so on - graph grows richer with more data → adding context dynamically



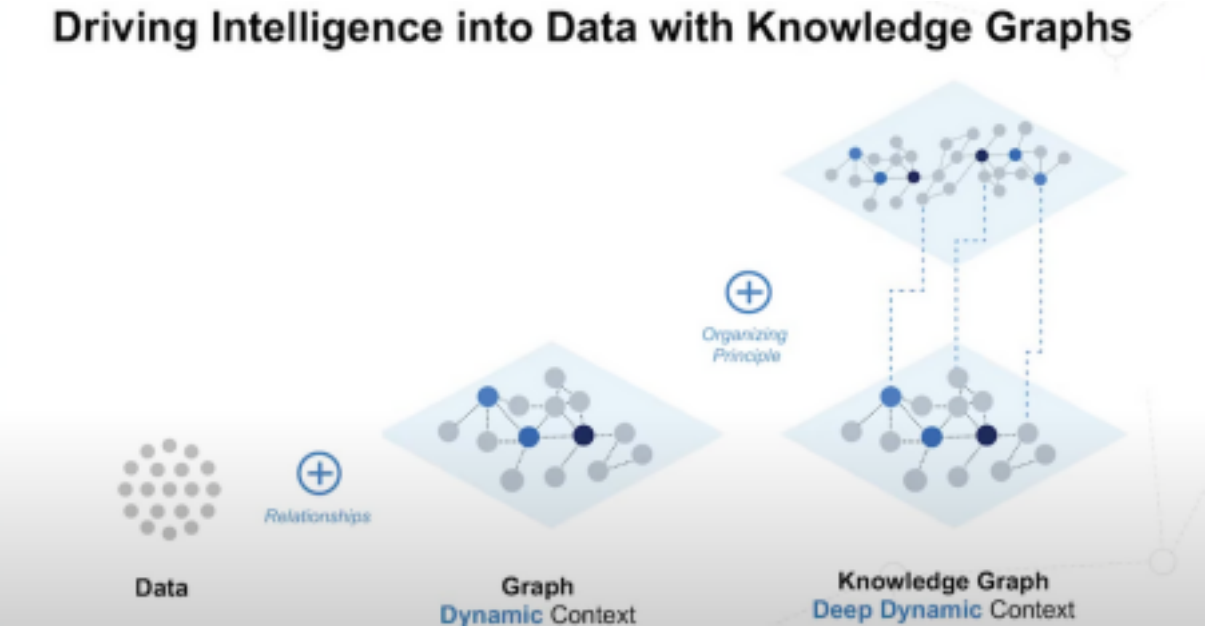
Understanding the connectedness of learner, skill, content and career (Source Link)

Why is a knowledge graph and an accompanying tool a better approach?

Here's how such a system would be more performant, compared to existing solutions.

1. **Lack of connectedness in current systems** - All the methods listed above have studied the process of skill mapping & analyzing workplace training as data points in a table as opposed to harvesting the connected nature of these skills. How 1 skill leads to another & how 2 seemingly unrelated skills fall under the same broader topic that can be created into a productive role for an employee.
2. **Lack of a tool that can help digest insights** - There is a lack of a software tool that companies can actually use to launch & track the Return On Investment of these educational programs. Introducing such a software would also increase employee awareness & literacy about the breadth & depth of each such program

Deriving intelligence by a layered approach



Layered approach of deriving implicit relationships from a knowledge graph

Thus, using an employee knowledge management tool, enables a company to establish the "data" layer, where we mention the company education program data. Next we build the "graph (aka 'dynamic context')" layer on top of this, to see how different programs & courses are connected to each other. Lastly, we build another layer - the "knowledge graph" where we specify rules where an employee would need to complete a set of programs to get

certified in a skill and get a job role in the newly developed skill. Over this, we perform "**semantic search**" to get the answer to any corporate training related question, in an accurate form.

Semantic search using a knowledge graph

Performing semantic search using a knowledge graph offers a powerful and intuitive way to retrieve information that goes beyond simple keyword matching. A knowledge graph is a structured representation of knowledge that captures relationships and dependencies between various entities and concepts. By leveraging the interconnectedness of entities within a knowledge graph, semantic search enhances the search experience by understanding the context and meaning behind user queries.

In a semantic search scenario, the knowledge graph serves as a valuable resource for indexing and organizing information. It contains a vast collection of entities, each with their own attributes and relationships. These relationships can be hierarchical, such as parent-child relationships, or more complex, capturing associations and connections between different entities.

When a user initiates a search query, the semantic search system analyzes the query and extracts its underlying intent. Instead of relying solely on matching keywords, the system examines the relationships within the knowledge graph to infer the most relevant entities and concepts related to the query. This process allows for a deeper understanding of the user's search intent and enables the retrieval of more accurate and contextually rich results.

The knowledge graph plays a crucial role in supporting semantic search by providing valuable context and enabling advanced search capabilities. It allows for entity disambiguation, resolving potential ambiguities by considering the relationships and attributes of entities. For example, if a user searches for "apple," the system can determine if the user is referring to the fruit or the technology company based on the context derived from the knowledge graph.

Additionally, the knowledge graph enables semantic search to uncover implicit relationships between entities. By traversing the graph and examining the connections between entities, the system can identify relevant information that may not have been explicitly mentioned in the user query. This contextual understanding enhances the search results by surfacing related entities, concepts, or even inferred knowledge.

Thus, our the next step in our journey, was to create a the skill map (in the form of a "knowledge graph") and leverage that for semantic search queries, where a user could simply go in and search what they wanted in the language most comfortable to them - and

the system would be intelligent enough to return the best answer and recommendation to the user.

Working with NPTEL team

The National Programme on Technology Enhanced Learning (NPTEL) is a pioneering initiative in India, launched by seven Indian Institutes of Technology (IITs) and the Indian Institute of Science (IISc) in 2003. It is funded by the Ministry of Human Resource Development (MHRD), Government of India, with the vision of providing quality education to anyone interested in learning from the IITs and IISc.

In late November, post the release of our sandbox solutions, we demoed the solution to Dean Academic courses Professor Prathap Haridoss and he started conversations with the NPTEL team at IIT Madras. We started working with Professor Jayakrishnan from the NPTEL team for data procurement and skillmap creation and personalizing NPTEL courses for learners from all over India.

NPTEL, as a platform was facing 4 problems that we identified

1. **Low retention rates** - Although NPTEL was running a Domain Specialization program where a learner could study a sequence of subject courses to gain a topic proficiency, they were noticing only a 3% retention rate (doing more than 2 courses in the same NPTEL domain) of learners. Even, the PR of the NPTEL Stars program was not gaining traction enough as the learning paths of the achievers of NPTEL were not public or helping other learners learn from their precedents.
2. **Student adoption** - 3 problems were faced.
    a. For students, it was difficult to understand which NPTEL courses were matching to their interests and career goals. A recommendation system was much required here, to match the likes of leading online platforms like Coursera and Udemy.
    b. The language barrier of the all-English curriculum was hindering even basic content consumption and course search.
    c. Lastly, doubt resolution was not personalized to each student and hence, dropoffs were observed in students mostly around the 50% course completion checkpoint

Thus, we started onboarding NPTEL data into our skill-map to provide them with AI recommendation, and analytics to improve their topline (link)

Phase 2 - January to April, 2023

Post the release of Large Language AI models in December 2022, we started exploring the space of combining a Generative AI model with our knowledge graph to provide unparalleled intelligence of interconnected systems, served through a basic text-query interface. So, as mentioned above, in this quarter, we focussed on building our knowledge graph, working with an MNC in corporate training and competency calculation and converging on our final platform for personalized learning.

The 1st place to start was to use the data from the sandbox launches at IIT Madras and create a skillmap from the parsed data. From this, we would enable sample creation of a basic interface where students could just go and ask for course recommendations in plain text format.

Building a skillmap for IIT Madras

Understanding the approach

For creating a skill-map, we used the Neo4j graph database, since it offers, out of the box, storage of data entries as nodes having relationships to other nodes in the form of edges. For extracting information from IIT Madras, we used previously available student information since the past 15 years, as well as wrote a custom script to scrape this information from the web and store it in our own database

The Neo4j graph database, owing to its dynamic schema and relationship-oriented structure, can be an incredibly effective tool for constructing a comprehensive skill map utilizing educational course content data from IIT Madras.

The process began by creating nodes for various elements such as courses, skills, topics, and job requirements. Each node represented a discrete entity. For instance, individual courses from IIT Madras formed one set of nodes, the topics within these courses formed another set, the skills acquired from mastering these topics were another set, and so forth.

Following this, we established relationships or edges between these nodes. These relationships were direct and indirect, and they offered immense flexibility in defining the connections between entities. For instance, a course 'C' may be linked to a skill 'S' implying that studying 'C' imparts 'S'. Similarly, a job 'J' may require skill 'S', hence forming another relationship. By connecting these relationships, we can infer that course 'C' equips a candidate for job 'J'.

To demonstrate with an example, let's consider the case of a job requiring knowledge of a "programming language" and a candidate knowing "Python". In the graph, "Python" would be a node under the broader node of "Programming Languages". The candidate node would link to the "Python"

node, establishing their skill set. The job node would be connected to the "Programming Language" node. Consequently, it would be straightforward to identify the overlap and establish the candidate's fit for the job, as the graph data model inherently reflects the interconnectedness of the data.

Moreover, Neo4j's powerful querying language, Cypher, was employed to traverse this graph and provide insightful recommendations or connections based on the relationships within the data. The graph model's inherent ability to express overlaps, hierarchies, and associative relations enables an efficient and intuitive way to map skills and qualifications to job requirements, making it an excellent tool for this purpose. The Neo4j graph database thereby provided a robust and dynamic foundation to construct a comprehensive, insightful, and adaptable skill map using IIT Madras data.

Extracting valuable course tags using Natural Entity Recognition and NLP

First, we took the metadata related to a course such as name, department, textbooks, reference books, course-description, course-content and collated it into a single database for over 600+ courses. Next, to really link the commonalities between topics in courses, we had to extract valuable tags for each course that would help us understand its key concepts and link it to other courses.

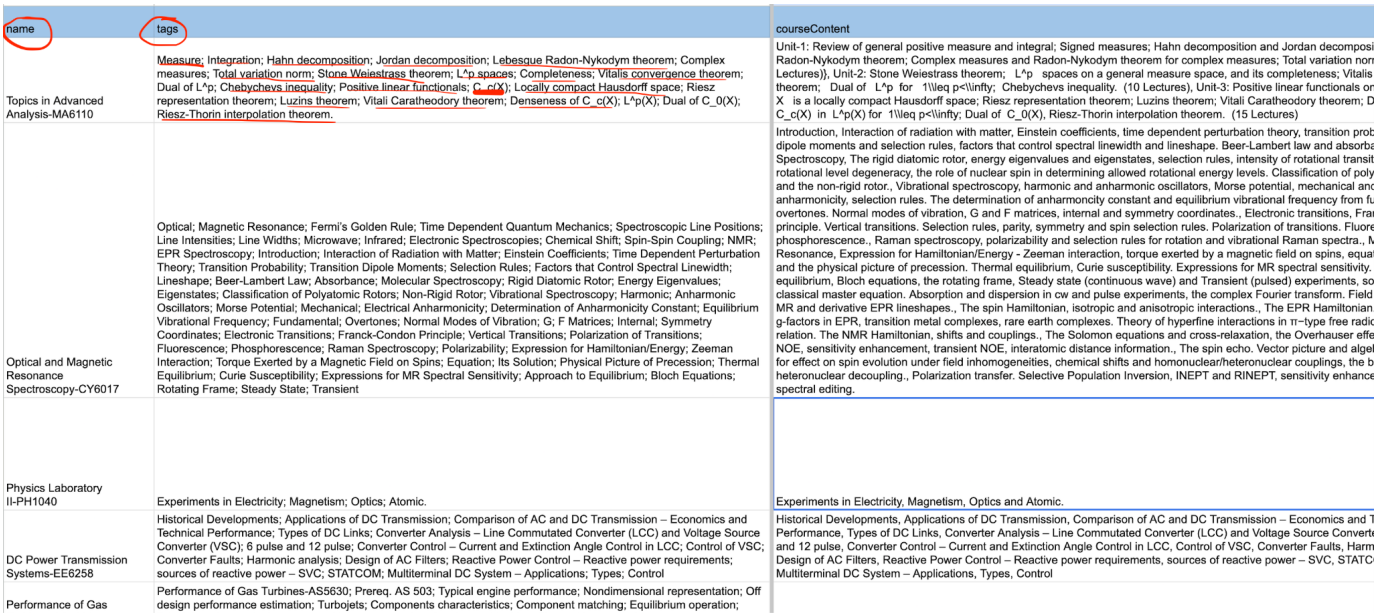
Natural Language Processing, commonly known as NLP, is a branch of artificial intelligence (AI) that focuses on the interaction between humans and computers using the natural language. The ultimate goal of NLP is to read, understand, and make sense of human language in a value-added way. It combines the power of linguistics and computer science to break down sentences into grammatical units, understand the meaning, learn the sentiment, and much more.

Named Entity Recognition (NER), a subset of NLP, is a method used to extract information from unstructured text. This process involves identifying and classifying named entities in text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. NER can be crucial in several applications including information extraction, question answering systems, and machine translation, among others.

Sample NER over a basic text paragraph to extract valuable information

In the context of extracting tags from course metadata, NLP and NER play a pivotal role. The raw course description data provided in the table is unstructured and dense with important information distributed throughout. In this scenario, NLP and NER act as powerful tools to make sense of this data by understanding the context, identifying the key entities, and classifying them.

Take for example the provided course description: "Topics in Advanced Analysis-MA6110." Here, NLP algorithms are first used to parse the text, understanding the syntactic structure of the sentences, and identifying key themes and topics within the text. This includes recognizing key terminologies specific to the course such as "Hahn decomposition," "Jordan decomposition," "Lebesgue Radon-Nykodym theorem," "Complex measures," and "Total variation norm," among others.



name	tags	courseContent
Topics in Advanced Analysis-MA6110	Measure; Integration; Hahn decomposition; Jordan decomposition; Lebesgue Radon-Nykodym theorem; Complex measures; Total variation norm; Stone Weiestrass theorem; L^p spaces; Completeness; Vitalis convergence theorem; Dual of L^p; Chebychevs inequality; Positive linear functionals; C_c(X); Locally compact Hausdorff space; Riesz representation theorem; Luzins theorem; Vitali Caratheodory theorem; Denseness of C_c(X); L^p(X); Dual of C_0(X); Riesz-Thorin interpolation theorem.	Unit-1: Review of general positive measure and integral; Signed measures; Hahn decomposition and Jordan decomposi Radon-Nykodym theorem; Complex measures and Radon-Nykodym theorem for complex measures; Total variation nor Lectures)}, Unit-2: Stone Weiestrass theorem; L^p spaces on a general measure space, and its completeness; Vitalis theorem; Dual of L^p for 1\\leq p<\\infty; Chebychevs inequality. (10 Lectures), Unit-3: Positive linear functionals or X is a locally compact Hausdorff space; Riesz representation theorem; Luzins theorem; Vitali Caratheodory theorem; D C_c(X) in L^p(X) for 1\\leq p<\\infty; Dual of C_0(X), Riesz-Thorin interpolation theorem. (15 Lectures)
Optical and Magnetic Resonance Spectroscopy-CY6017	Optical; Magnetic Resonance; Fermi's Golden Rule; Time Dependent Quantum Mechanics; Spectroscopic Line Positions; Line Intensities; Line Widths; Microwave; Infrared; Electronic Spectroscopies; Chemical Shift; Spin-Spin Coupling; NMR; EPR Spectroscopy; Introduction; Interaction of Radiation with Matter; Einstein Coefficients; Time Dependent Perturbation Theory; Transition Probability; Transition Dipole Moments; Selection Rules; Factors that Control Spectral Linewidth; Lineshape; Beer-Lambert Law; Absorbance; Molecular Spectroscopy; Rigid Diatomic Rotor; Energy Eigenvalues; Eigenstates; Classification of Polyatomic Rotors; Non-Rigid Rotor; Vibrational Spectroscopy; Harmonic; Anharmonic Oscillators; Morse Potential; Mechanical; Electrical Anharmonicity; Determination of Anharmonicity Constant; Equilibrium Vibrational Frequency; Fundamental; Overtones; Normal Modes of Vibration; G; F Matrices; Internal; Symmetry Coordinates; Electronic Transitions; Franck-Condon Principle; Vertical Transitions; Polarization of Transitions; Fluorescence; Phosphorescence; Raman Spectroscopy; Polarizability; Expression for Hamiltonian/Energy; Zeeman Interaction; Torque Exerted by a Magnetic Field on Spins; Equation; Its Solution; Physical Picture of Precession; Thermal Equilibrium; Curie Susceptibility; Expressions for MR Spectral Sensitivity; Approach to Equilibrium; Bloch Equations; Rotating Frame; Steady State; Transient	Introduction, Interaction of radiation with matter, Einstein coefficients, time dependent perturbation theory, transition prob dipole moments and selection rules, factors that control spectral linewidth and lineshape. Beer-Lambert law and absorba Spectroscopy, The rigid diatomic rotor, energy eigenvalues and eigenstates, selection rules, intensity of rotational transit rotational level degeneracy, the role of nuclear spin in determining allowed rotational energy levels. Classification of poly and the non-rigid rotor., Vibrational spectroscopy, harmonic and anharmonic oscillators, Morse potential, mechanical anc anharmonicity, selection rules. The determination of anharmoncity constant and equilibrium vibrational frequency from fu overtones. Normal modes of vibration, G and F matrices, internal and symmetry coordinates., Electronic transitions, Frar principle. Vertical transitions. Selection rules, parity, symmetry and spin selection rules. Polarization of transitions. Fluore phosphorescence., Raman spectroscopy, polarizability and selection rules for rotation and vibrational Raman spectra., N Resonance, Expression for Hamiltonian/Energy - Zeeman interaction, torque exerted by a magnetic field on spins, equa and the physical picture of precession. Thermal equilibrium, Curie susceptibility. Expressions for MR spectral sensitivity. equilibrium, Bloch equations, the rotating frame, Steady state (continuous wave) and Transient (pulsed) experiments, so classical master equation. Absorption and dispersion in cw and pulse experiments, the complex Fourier transform. Field MR and derivative EPR lineshapes., The spin Hamiltonian, isotropic and anisotropic interactions., The EPR Hamiltonian g-factors in EPR, transition metal complexes, rare earth complexes. Theory of hyperfine interactions in π−type free radic relation. The NMR Hamiltonian, shifts and couplings., The Solomon equations and cross-relaxation, the Overhauser effe NOE, sensitivity enhancement, transient NOE, interatomic distance information., The spin echo. Vector picture and algel for effect on spin evolution under field inhomogeneities, chemical shifts and homonuclear/heteronuclear couplings, the b heteronuclear decoupling., Polarization transfer. Selective Population Inversion, INEPT and RINEPT, sensitivity enhance spectral editing.
Physics Laboratory II-PH1040	Experiments in Electricity; Magnetism; Optics; Atomic.	Experiments in Electricity, Magnetism, Optics and Atomic.
DC Power Transmission Systems-EE6258	Historical Developments; Applications of DC Transmission; Comparison of AC and DC Transmission – Economics and Technical Performance; Types of DC Links; Converter Analysis – Line Commutated Converter (LCC) and Voltage Source Converter (VSC); 6 pulse and 12 pulse; Converter Control – Current and Extinction Angle Control in LCC; Control of VSC; Converter Faults; Harmonic analysis; Design of AC Filters; Reactive Power Control – Reactive power requirements; sources of reactive power – SVC; STATCOM; Multiterminal DC System – Applications; Types; Control	Historical Developments, Applications of DC Transmission, Comparison of AC and DC Transmission – Economics and T Performance, Types of DC Links, Converter Analysis – Line Commutated Converter (LCC) and Voltage Source Converte and 12 pulse, Converter Control – Current and Extinction Angle Control in LCC, Control of VSC, Converter Faults, Harm Design of AC Filters, Reactive Power Control – Reactive power requirements, sources of reactive power – SVC, STATC Multiterminal DC System – Applications, Types, Control
Performance of Gas	Performance of Gas Turbines-AS5630; Prereq. AS 503; Typical engine performance; Nondimensional representation; Off design performance estimation; Turbojets; Components characteristics; Component matching; Equilibrium operation;	

*Performing NER over IIT Madras course metadata to extract **key concepts** as tags*

Once these entities are identified, NER comes into play. These identified entities are then classified into a 'tag' category as they represent the core topics covered in the course. This process of identification and categorization helps in systematically organizing the course data, making it easily searchable, relatable, and accessible.

In conclusion, the use of NLP and NER allows for a more structured representation of the course metadata. This paves the way for better course recommendation systems, academic research, and curriculum design. The AI, through these processes, enables a robust and efficient way of connecting learners to the right educational resources based on their needs and goals.

Full dataset of 660 courses and their tagging using NLP is hosted here - link

Once the tags were extracted, it would be easy to link courses to one another, as if nodes in a graph and hence, we proceeded with onboarding the data into Neo4J graph database.

Interesting Findings

Upon onboarding, the IIT Madras course data, we observed 74630 nodes and 100,283 relationships, signifying a 135% interconnectedness metric. Here are the top 3 interesting metrics we found

1. Most popular key concepts taught at IIT Madras - full result at link. Here are the top results

tagName	courseCount
Introduction	101
Applications	59

Design	53
Analysis	48
Classification	42
Stability	37
Research	30
Technology	28
Modeling	26
Optimization	24
Project	24
Simulation	23
Probability	22
Oral Presentation	22
Kinematics	22
Control	21
Dynamics	21
Properties	21
Measurement	21
Special Topics	21
Structure	21
Communication	20
Principles	20
Presentation	20
Thermodynamics	20
Sensors	20
Vocabulary	19
Instrumentation	19
Economics	19

This list solidifies that IIT Madras

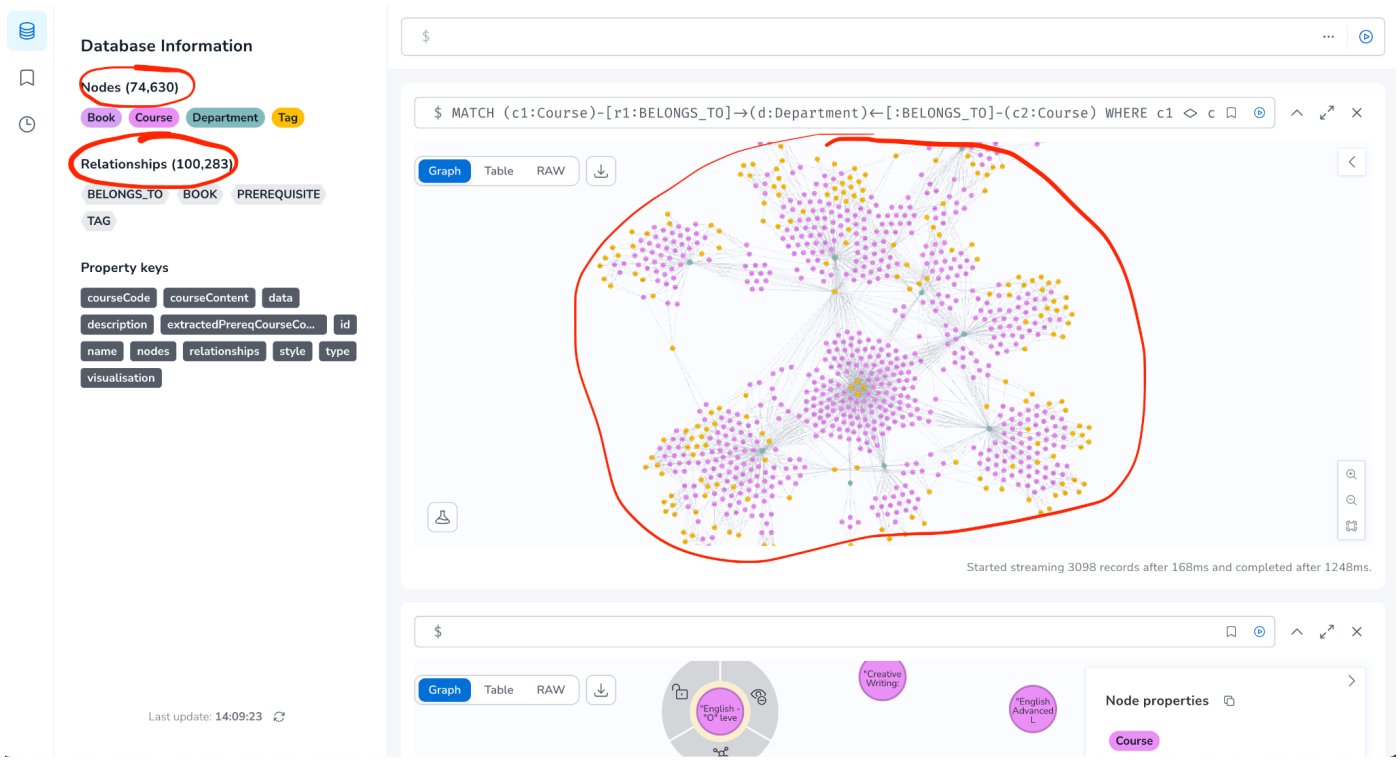
1. Primarily offers introductory courses
2. Most courses are application based
3. No matter the breadth of course, certain common topics such as - research, optimization, technology, project, dynamics, measurement, economics - etc. are important to most courses offered

2. Most connected courses from the entire course list - full result at link. Here are the top results

Course	SharedTagCount	Department
Engineering Mechanics-AM1100C	214	AM
English Advanced Level-HS2210	172	HS
Cultural Studies-HS8640	153	HS
Perspectives in Social Sciences-HS1030	152	HS
Social Psychology-HS4520	151	HS
Financial Management-MP5210	151	HS
World Literature-HS4140	151	HS

This result goes to prove why courses like Engineering Mechanics, Financial Management and English are so core to the entirety of engineering and hence, must be offered to all first year students

3. **Course similarity** - We performed vector embeddings over tags data and were able to determine "course similarity" metric using *cosine* formula. For a given course with a set of tags, we were able to figure out courses that were similar to this. For example, for 1 course "Introduction to `Machine Learning" we were able to get "Deep Learning", "Reinforcement Learning", "Pattern Recognition and Machine Learning" as similar courses. Here's the embedded course data for downloading - [link]



A basic view of the SkillMap as a Neo4J graph

Using the skill-map to help resolve basic course recommendation

The skillmap helped a student understand courses of their choice by offering a basic keyword matching solution.

For eg, if a student went ahead and typed - “How do I become a poet?”, then the query engine would extract valuable tags, such as “poet” from this query and convert it into the following query against graph database - `MATCH (c:Course)-[:TAG]->(t:Tag) WHERE t.name CONTAINS 'Poet' OR t.name CONTAINS 'Poetry' RETURN DISTINCT c.name, [r IN (c)-[:TAG]->(t) | t.name] as tags ORDER BY c.name`

Following this query against the database, we get the following information as output:

Name
African & African-American Literature-HS6300
Creative Writing: Practice and Theory-HS6027
English - "O" level-HS1210
English Advanced Level-HS2210
English Advanced Level-HS2210A
Indian Classics and Cultural Values-HS4020
Poetry-HS5610
Twentieth-Century American Poetry-HS7373

The complete output, with the set of tags can be viewed here - link

Thus, using a skillmap, we were able to tag and derive valuable information from interconnected tags.

The hypothesis that the academic data was interconnected and these connectedness held valuable information could be further utilized for corporate training at a professional level and hence, we partnered with HyperVerge Academy - a company focussed on providing industrial IT training within 6 months to its learners. The goal was to improve their curriculum design and competency calculation flow

Case Study: HyperVerge Academy (HVA)

HyperVerge Academy (HVA) is a professional upskilling school that runs 6-month programs to train people in industrial skills to get employment. They have been in the market since 2019 and have trained over 1000 individuals in technical and soft skills. They needed data intelligence to enhance

their industrial training programs and improve their ROI. Hence, we partnered with them from Jan-March,2023 and helped them understand, evaluate and utilize their academic information to devise better programs.

The reason we wanted this case study was to understand that after solving the value in interconnectedness of academic data, how could we employ the same to calculate the ROI of a company. The answer lied in - “how well does the training enable competency development of their trainees”. A good way to measure the same could be something on the likes of the Kirkpatrick model that we had already discussed earlier.

An extensive company consulting research document that we drafted over the course of January-March, 2023 is available here - link. In this report, we shall only be discussing the key points in our journey with HyperVerge

Understanding the professional upskilling space

Professional upskilling and corporate training represent significant areas of focus in the ever-evolving professional landscape. As technology and market demands rapidly evolve, keeping pace with the latest trends, knowledge, and skills becomes a complex task. The complexity of this challenge arises from the breadth and depth of the subjects that professionals need to master, the speed at which these subjects change, and the unique learning needs of each professional.

The problem is compounded by the reality of a diverse workforce. Different roles within the same organization require different sets of skills. Furthermore, individual learning styles, prior knowledge, and career trajectories necessitate personalized training programs. Traditional training approaches often struggle to meet these needs, leading to less effective training outcomes and employees that are ill-prepared for the rapidly shifting professional landscape.

Solving the problem of professional upskilling and corporate training is essential for several reasons. First, it enables organizations to maintain competitiveness and adaptability in the face of change. By ensuring employees have the necessary skills to perform at their best, businesses can remain innovative and agile. Second, it boosts employee engagement and satisfaction. Professionals value opportunities to learn and grow, and by providing high-quality training, businesses can attract and retain top talent.

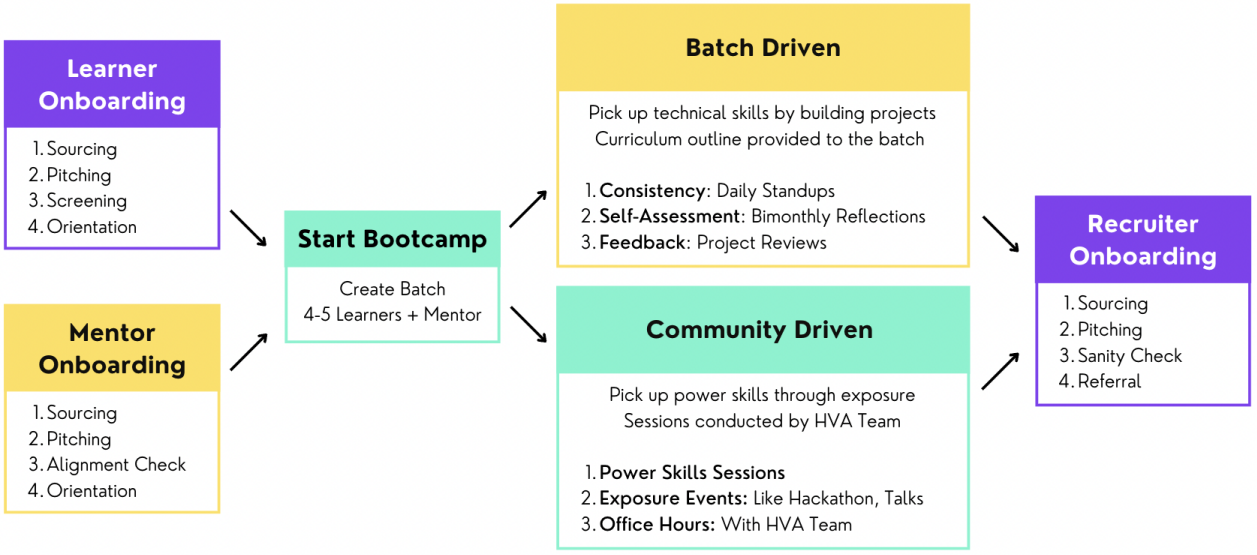
Here are some notable metrics pulled from this survey in the field of industrial training:

- 94% of employees prefer a company that invests in their growth
- 74% of employees feel that they aren’t achieving their full potential due to lack of development opportunities

- 70% of employees indicated that training & development opportunities at a job influence their decision to stay
- 56% of employees would take a course recommended by their manager

Research at HyperVerge Academy (HVA)

HVA is solving the problem of the skill & network gap for young people from underprivileged backgrounds to secure and thrive at well-paying jobs in the tech sector at scale. HyperVerge Academy onboarded 128 learners in 2022. This translates to 25 batches, 23 new mentors. The average learner salary in 2022 was 4.1 LPA. Impact: Increase in household income due to increase in salaries



The flow of operations at HyperVerge Academy

We conducted over 40+ interviews spread across program facilitators (aka PFs - people who design the curriculum, coordinate the student batches and ensure quality learning), mentors (industry professionals who teach IT skills) and learners (students who come to HVA to learn employable IT skills). An extensive log of the same can be found here - link

Major Challenges faced:

The 3 major challenges faced were:

Challenges	Description
Visibility on the program	Low Clarity on: 1. What skills (tech, non-tech) are important to achieve the goal?

	2. What is a learner's roadmap? 3. Where does each learner stand on that roadmap? 4. How can their progress be tracked?
Program Team Bandwidth	1. No. of batches/PF increases 2. Unable to spend 1-1 time with learners at risk 3. Limited analysis of data collected; no streamlining
Engagement and Feedback	1. Low learner interaction with batch, community, and program 2. Prog team time goes into follow ups for feedback, problem-solving

Output of challenges

1. Learner is very dependent on mentor, and prog team for guidance and next steps
2. Mentor/Recruiter can't determine the learner's skill achievement
3. Course Correction is severely delayed
4. Community engagement on slack, outside mandatory calls is extremely low
5. Max time spent on BAU, little/no time on innovation for scale

Outcomes:

1. Quality of Placements is poor, and the average salary is below expectation
2. Learner's job is unstable post placement
3. Learner drop-offs are high
4. Unable to increase program capacity

Proposed Solution - Build a **platform** that addresses the 3 main challenges of ***On-trackness, Program Bandwidth and Engagement*** mentioned above:

The Philosophy behind the Platform:

1. It should promote the concept of **"Zone of Proximal Learning"**- Learner moves from Dependent to Guided to Independent Learning
2. It should be **customisable** as much as possible to keep up with the changing program
3. It should be **personalized** to help each student learn at their own pace, build on their strengths and catch up on their weaknesses
4. Platform should be enabling and non-invasive

What does this Platform need to have?

To address this, the Platform should have views for:

1. **Learners**: to plan and update progress, share assignments, notify completion of milestones, schedule assessments, visualize progress, seek recruiters through their candidate profile, give/take feedback
2. **Mentors**: to monitor learner batch progress, set assignments/tasks, approve milestones, give/take feedback, share learner testimonials
3. **Program Facilitators**: to create new curriculums, setup new skill sets, monitor batch progress through key indicators, give/take feedback, notify/schedule calls
4. **Recruiter**: to post opportunities, see candidate profiles, set up assignments, give/take feedback

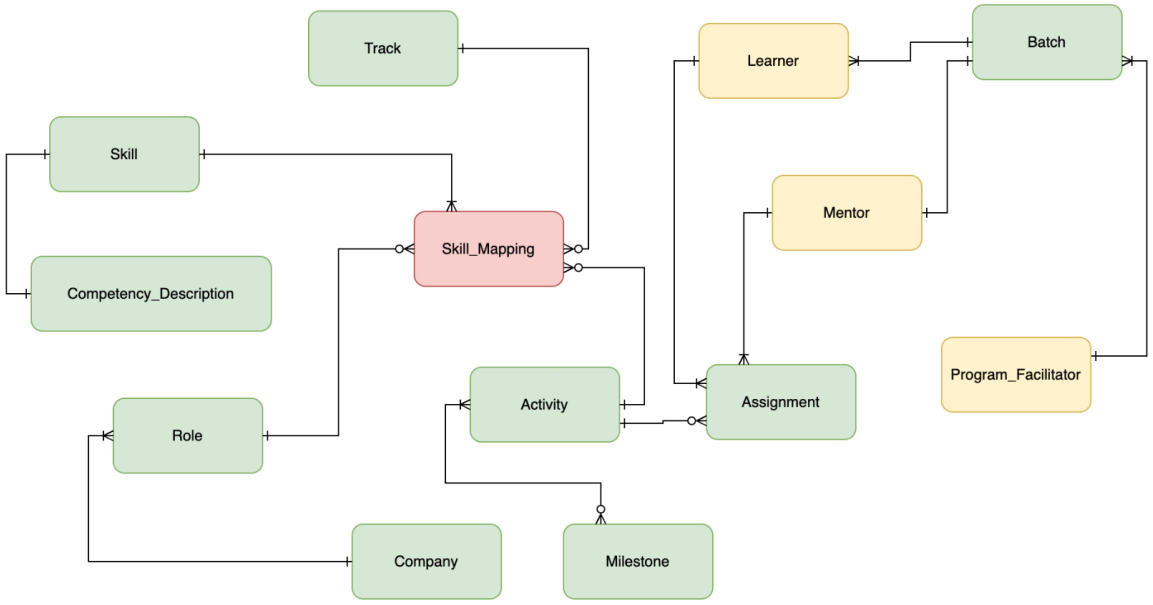
Developing a competency engine and data model

The main goal of an industrial training academy like HVA, is to ensure that the competencies of their learners in the skills they are being trained in, are up to the mark.

Competency is the ***quantification of a skill of a learner*** (eg: Skill = Frontend Dev, Score Points = 50, Competency = 50 in Frontend Development). A competency matrix is a list of such competencies (50 in Frontend Development, 40 in Backend Development. At the end of the day, the competency matrix of a learner getting met, is a North Star. This is what would get the learner plan and achieve their skill and use it for placement. Everyone at HVA - learners, mentors and PFs → must plan, work and track the big picture to move towards it successfully

The following are the data models that we stored and connected in our solution.

Entity Relationship Diagram



The Entity Relationship Diagram for the solution at HVA

1. **Learner** - This is a **"user"** defined for each student at HVA. We shall be storing background information, competency development progress and graduation details.
2. **Mentor -** This is a **"user"** defined for each mentor at HVA. We shall be tracking background infromation, batch health, activities and overall contribution.
3. **Program Facilitator (PF)** - This is a "**user"** defined for the admins of the HVA program. We shall be storing basic user metadata, sourcing data, connected batch information, activites and contribution
4. **Track -** This refers to education curriculums that HVA designs for developing a specialized set of skills in a domain. Examples - "Full Stack Development Track", "DevOps Track". We shall be tracking the curriculum, associated skills, batches and placement metrics
5. **Batch** - This is the "group of learners + 1 mentor + 1 PF" that follows a "Track" at HVA. We shall be storing batch health, associated stakeholder (learner, mentor, PF) details.
6. **Milestone** - This is a module aka "collection of related learning activities/tasks" that a learner gains a skill in a broad topic. Example - "Basics of JS" can be a milestone in the "Web Development Track" at HVA. Other milestones in the same track would include - "UI Development with HTML, CSS", "Intermediate JS", "Advanced Web Dev Projects" etc. We store the associated activities, aggregated skills and related scores associated with each activity inside this milestone.
7. **Activity** - This is the "learning task that a learner performs to develop a skill". It is the child of the Milestone data model. Example - "Basics of JS" can have these 4 activities - 'Read JS Basics blog,

watch introductory YouTube video, Appear for JS mock-interview, Build a JS Todo App". For each activity, we associate a set of skill and a score assigned to each skill. Example - the 'Build JS Todo App can award the learner 2 points in the knowing-JS skill and 5 points in the critical thinking skill. We store the metadata, relevant links, associated skills and scores.

8. **Assignment** - Mentor telling a Learner to do an Activity within a stipulated deadline.
9. **Skill** - It is the basic unit in competency. It is associated to a broad domain (such as "JS" being associated with "Programming"). We store the name, description and domain
10. **Company** - This is the company that a learner interacts with for recruitment. We store the metadata (name, location), competency matrix across roles, average review and placement stats for the company
11. **Role** - This refers to the job defined at a Company, associated with its own competency matrix. We store name, description, competency details, job description, compensation details

Thus, using this data model for competency development, we studied over 30+ existing market solutions that were providing this tool to companies. Here, is the research link for the comparative analysis - link

We realized that no solution was tying competency development to academic activities and industrial output in a direct fashion, as recommended by the Kirkpatrick model. They were either assessment platforms like iMocha or employee productivity and performance tracking platforms. Thus, we devised a simple application using Google sheets, to enable HVA team to monitor students' competencies getting developed across multiple skill sets and help them personalize their tutoring based on the student's scores in each skill - link to application

Using this platform, the mentors and program facilitators had a more numerical insight into the following:

1. What are the skills and level in each skill required for a particular job role (the given example application studies the job role - Full Stack Developer)? These metrics were researched from the internet as well as through consulting with recruitment firms
2. How to design curriculum programs for industrial training such that learners should be able to flexible pick assignments and topics and gain competencies required for a job role
3. How well was a learner learning a set of skills based on the career role they chose? In which skill were they excelling, where were they lacking and by how much?
4. How was a particular batch of learners performing - this would help evaluate the effectiveness of the mentor as well

Thus, by using Kirkpatrick model and simple numerical analysis through a spreadsheet, we arrived at a solution that saved the company around INR 1.4 lakhs and made their processes more efficient

Solutioning and Learnings

There were 2 key learnings during our engagement with HVA:

1. **Personalized learning is the only important unsolved problem** - Every learner has their own personalized style, pace and career paths and need support in different areas. We already knew that skill development data is highly interconnected - this was further validated through the data from the competency engine we built and deployed. Search, recommendation and personalization are the major unsolved avenues - that is the best leveraged use case of the skillmap knowledge graph.

2. **Corporate training academies could directly benefit from education data** - Such professional upskilling academies could design their curriculums, track competency development of learners and personalize mentorship using academic data.
   Hence, the "buffet problem" addressed at the beginning of this project was getting resolved.
   Except for 1 major unsolved part

3. **The uniqueness of each academy** - Each academy was different and their graph of academic data was unique - thus, developing a single system and interface would not work for the users as it would make customization per academy a massive engineering overhead each time any intelligence was required. Thus, we decided to make the system intelligent using Generative AI such that, irrespective of the industrial training ecosystem, just through a single interface, any user could get the best of information out of the graph.

Arriving on an intelligent product with Knowledge Graph

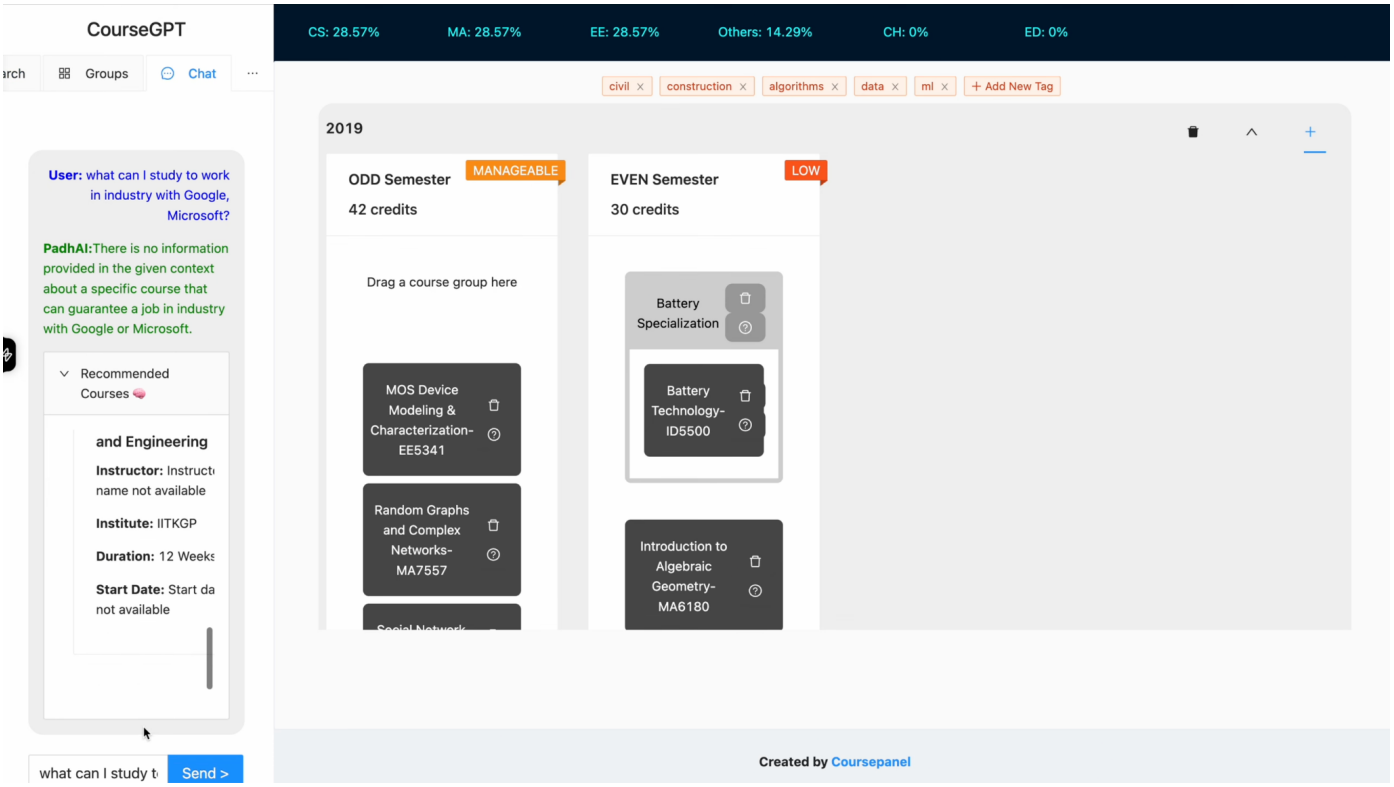
By the beginning of March 2023, we realized that solving use cases one-by-one for each case study - be it NPTEL or HyperVerge or L&T would take time and not generalize well. Hence, we decided to use Generative AI as a technology to prepare the graph and generate appropriate graph queries from the most basic input - such as a voice input from a user in their vernacular language. Compiling all the learnings from NPTEL, IIT Madras and HyperVerge, we trained CourseGPT over NPTEL data - to be the AI sidekick for any student coming to the NPTEL platform for professional education.

Here is a video presentation of the product - link. Going through this video is highly recommended before you proceed with understanding the product, the system and how it works.

CourseGPT

CourseGPT the learner's intelligent partner in online learning. Powered by the latest advancements in artificial intelligence, CourseGPT is an advanced chatbot integrated into any Learning Management System (LMS). It is specifically designed to provide a personalized and seamless learning experience, addressing the diverse learning needs of students and professionals across the globe.

CourseGPT isn't just a chatbot - it's your personal academic advisor, doubt resolver, and learning companion rolled into one. With an incredible range of capabilities that leverage the transformative power of AI, CourseGPT is all about enhancing your learning journey and driving you towards your career goals.



A screenshot of the CourseGPT platform

What does it do?

The following are the key features of the CourseGPT platform in action:

1. **Tailored Course Suggestions:** Whether you're an aspiring software engineer or a digital marketing enthusiast, CourseGPT is your AI sidekick for higher learning. It understands your career goals and preferences, recommending the most suitable courses for you. For instance, if you're aiming to be a software engineer, it could suggest courses in algorithms and machine learning. Leveraging vast amounts of job data and educational resources, CourseGPT can even provide course recommendations for specific industry aspirations, such as working for tech giants like Google or Microsoft.
2. **Strategic Academic Pathway Design:** CourseGPT doesn't just stop at course recommendations. It aids you in strategizing your entire academic journey. By curating an optimal combination of courses available on our platform, it helps you specialize in your chosen field. Whether you're pursuing a major in civil engineering with a minor in computer science, or aiming for a

specialization in battery technology, CourseGPT guides you towards achieving your academic goals. It even takes into account course prerequisites and semester load, ensuring a balanced and manageable study plan.

3. **Multilingual Support:** CourseGPT understands that students come from diverse linguistic backgrounds. To cater to this, it can respond to queries in multiple languages, including vernacular languages like Hindi. This feature ensures that students can make the best use of the platform, regardless of their native language.
4. **Direct Course Enrollment:** CourseGPT not only recommends courses but also allows you to directly enroll in them from the platform. This seamless integration ensures a smooth and efficient learning journey.
5. **Real-time Course Search and Addition:** With CourseGPT, you can search for courses in real-time and add them to your academic plan. If a recommended course piques your interest, you can immediately search for it and add it to your study plan. This feature saves time and enhances the user experience.
6. **Practical Learning Recommendations:** CourseGPT understands that theoretical knowledge needs to be supplemented with practical learning. Therefore, it also recommends platforms like LeetCode and HackerRank for practice, especially for subjects like algorithms. This holistic approach ensures that students are well-prepared for real-world challenges.
7. **Supporting platform management software**: CoursePlan allows colleges to create new courses easily. For example, if there's a demand for a course in a particular domain like economics, the platform enables the institution to launch a course in that area. It also provides analytics for similar previously offered courses, allowing the institution to analyze the performance of those courses and make informed decisions. The platform enables the creation of degrees, which are collections of rules that allow a student to specialize in a particular domain. It also allows for the creation of a curriculum roadmap for these degrees.

How does it enhance the learning experience

Here are 5 reasons why CourseGPT beats most current systems of learning:

1. **Personalized Learning Pathways**: CourseGPT uses advanced AI algorithms to understand each learner's unique learning style, career aspirations, and skill levels. It recommends individualized courses that align perfectly with their goals, unlike traditional systems that offer a one-size-fits-all approach.

2. **Dynamic Skill Mapping**: As industries evolve and skill requirements change, CourseGPT dynamically maps new skills and learning paths. This is in stark contrast to traditional learning systems which often struggle to keep pace with the rapidly evolving job market and technological advancements.

3. **Interactive Doubt Resolution**: CourseGPT incorporates an AI-driven interactive doubt resolution system, which instantly addresses any queries or doubts learners might have during their study. Traditional systems, on the other hand, rely on human instructors, making doubt resolution a slower and potentially inconsistent process.

4. **Language and Accessibility Support**: With support for multiple languages and voice assistance, CourseGPT ensures inclusivity in education. Traditional learning systems often fall short in this aspect, not providing ample support for non-native English speakers or those with accessibility needs.

5. **Strategic Academic Planning**: CourseGPT not only recommends courses but also helps learners strategically plan their academic journeys, suggesting combinations of courses that lead to specializations. This holistic approach to planning is often missing in traditional systems, which focus mainly on course delivery rather than comprehensive academic planning.

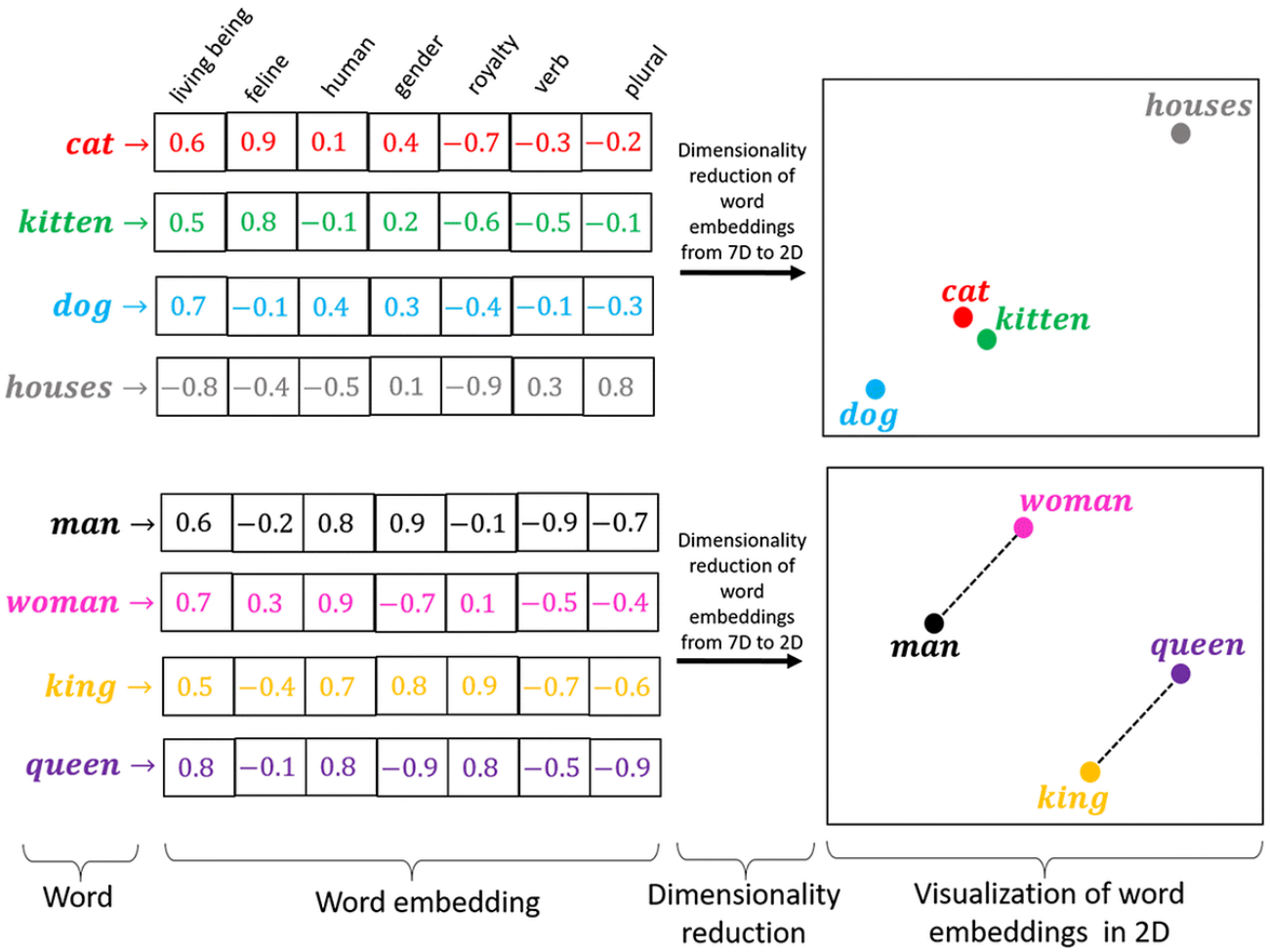
Technical Deep Dive

Understanding Natural Language Processing and Vector Embeddings

Natural Language Processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics. It involves programming computers to effectively process large amounts of natural language data. As part of NLP, Large Language Models (LLMs) like GPT-3, developed by OpenAI, are designed to generate human-like text based on the inputs they receive. They are trained on a diverse range of internet text and can generate contextually relevant text.

However, they do not know specific documents or sources they were trained on and don't have the capability to access or retrieve specific documents or databases during the generation process. Their strength lies in understanding and generating text that closely resembles human-like responses, thus forming a key part of systems like CourseGPT.

Vector embeddings are another crucial component of such systems. In the realm of machine learning, embeddings are a way to convert categorical variables into continuous vectors so as to feed them into a model. They can be used to capture the semantics of words in a language, relationships between items, and even user behavior in an application.



Basic demo of how words are converted into vector embeddings (source)

These embeddings, when stored in a vector database, form the basis of effective search and recommendation systems.

Building a vector database using a knowledge graph

For the purpose of this project, based on the previous training data, we turned our knowledge graph (aka “skillmap”) into a vector database for storing entries and answering queries related to academic data. This was done in 4 steps:

1. **Creating a Knowledge Graph in Neo4j**: The first step involved modeling our data as a graph and storing it in Neo4j. Our nodes represent entities such as NPTEL courses, books, departments, etc., and relationships represented the connections between these entities
2. **Generating Vector Embeddings**: The next step involved generating vector embeddings from our graph. These embeddings are a form of representation that captures the semantic meaning and relationships of your data in the form of vectors in a high-dimensional space. There are several techniques to generate graph embeddings, like Node2Vec, GraphSAGE, etc. This project used the

Node2Vec library in Python to achieve the same. We ran Node2Vec algorithms on our Neo4j graph to generate node and relationship embeddings.

3. **Store Vector Embeddings**: The generated embeddings then needed to be stored for future use. Typically, we'd want a database that is optimized for storing and querying high-dimensional vectors. You could use a vector database like Pinecone, Faiss, or even Elasticsearch with a vector plugin for this purpose.
4. **Querying the Vector Database**: The final part involved using these vector embeddings to make our application smarter. For example, we could now perform operations like **semantic search**, recommendations, or similarity checks by comparing vectors. This involved querying our vector database for the nearest vectors to a given input vector, which gave us the most semantically similar entities to our query.

Scoping NPTEL data into the Vector Database and building a chatbot

To create a system like CourseGPT, we first need to load the relevant data into the vector database. Let's assume we have course data from NPTEL in a CSV format. This data can be processed and converted into vector embeddings using various techniques like Word2Vec, GloVe, or even using transformers-based models. These vector representations can then be loaded into the vector database, which allows us to perform efficient similarity search operations.

The vector database enables us to compare the query vector (which can be a representation of a user's query or a specific course interest) with all vectors in our database, and retrieves the most similar entries. These entries represent courses that are most relevant to the user's query.

Once the entries are returned, we need to translate these course vectors back into a human-readable form. This is where LLMs like GPT-3 come into play. These models can generate contextually relevant, human-like text based on the returned entries. The generated text can be as simple as a course name and description, or as complex as a detailed career path recommendation.

In this way, the synergy of NLP, LLMs, and vector databases leads to the development of an effective system like CourseGPT. Such a system can revolutionize the way we approach professional upskilling and corporate training by providing personalized, contextual, and interactive learning experiences.

Harnessing CourseGPT for workforce development in construction companies

Currently, with the rapid digitization of industries, adopting cutting-edge technologies for corporate training has become a significant trend. For construction companies aiming to maintain a competitive edge, the advent of artificial intelligence (AI) chatbots, specifically ones built using the Generative Pre-training Transformer (GPT) model like CourseGPT, offers vast potential for workforce improvement and career growth.

CourseGPT proves to be a revolutionary tool for corporate training within construction companies. It is capable of facilitating personalized learning, recommending relevant courses, and helping employees plan their career trajectories. The bot's capabilities are underpinned by the robust GPT model that empowers it with a human-like understanding of language, making it a virtual trainer with near-human level competence.

One primary advantage of CourseGPT in corporate training is its ability to provide individualized learning recommendations. In the diverse ecosystem of construction companies where roles range from site engineers to project managers, the learning needs of every individual differ substantially. CourseGPT, using its advanced AI algorithms, can evaluate the learner's current skills, job role, and career aspirations to recommend suitable courses from the learning management system. It can suggest a site engineer to undertake a course on the latest construction safety standards or recommend a project manager a course on agile project management methods.

CourseGPT can also help employees navigate their career paths within the construction industry. By analyzing the employee's skills, job performance, career aspirations, and current industry trends, it can provide insights into possible career trajectories and the necessary skill upgrades. For instance, CourseGPT might recommend a construction supervisor with a knack for technology to consider courses in Building Information Modeling (BIM) for career progression.

Moreover, CourseGPT is equipped to assist in addressing skills gaps in the construction workforce. It can identify areas where employees may need further training and recommend specific courses to help them upskill. This not only ensures that the workforce is constantly evolving with industry standards and trends, but it also promotes a culture of continuous learning and development.

Furthermore, the chatbot, with its 24/7 availability, offers learners the flexibility to seek guidance anytime, anywhere. This is especially useful in construction settings where work schedules can be erratic, and accessing traditional training sessions can be challenging.

Another compelling advantage of CourseGPT is its potential to significantly improve employee engagement. The interactive nature of the chatbot makes learning more engaging and interactive, as opposed to conventional online courses. Employees can ask questions and receive immediate, personalized responses, making the learning experience more dynamic and fulfilling.

In a rapidly evolving industry such as construction, tools like CourseGPT are not merely a nice-to-have but a necessity for companies aiming to remain competitive and relevant in the market. By adopting such innovative technologies, construction companies can ensure a well-trained, motivated, and future-ready workforce.

Further development plans

CourseGPT is in its nascent stages of development and there were are several interesting avenues to be explored that would help improve learning experiences beyond anything we have witnessed before. These are some interesting directions that could be explored with this product:

1. **Adaptive Learning Pathways**: In future iterations, CourseGPT could generate custom learning pathways for users. Based on the student's goals, interests, and previous performance in courses, CourseGPT could recommend a personalized sequence of courses to optimize their learning journey.
2. **Peer Networking Suggestions**: CourseGPT could leverage AI to suggest peer networking opportunities based on shared interests, common courses, or career paths. This could foster more robust discussion forums and peer-to-peer learning opportunities, which can significantly enhance the learning experience.
3. **Deep Dive Sessions**: CourseGPT could schedule deep-dive sessions on topics where many students are struggling, based on an analysis of the questions being asked. These sessions could take the form of additional learning resources, expert-led sessions, or even a Q&A session with the CourseGPT providing detailed explanations.
4. **Integrated Job Market Insights**: CourseGPT could include a feature that draws on real-time job market data to give users insight into how a given course could impact their career prospects. For instance, if a user is considering a course in Machine Learning, CourseGPT could provide data on job demand for ML skills, salary trends, regional hiring trends, and more. This would help learners make informed decisions about their education in the context of their career goals.

Conclusion

The development and evolution of CourseGPT, spread over a 12-month period, represents a significant step forward in the application of advanced technologies such as AI, Natural Language Processing (NLP), and Knowledge Graphs to the critical area of professional upskilling and corporate training.

The journey began with understanding the problem domain, setting the objectives and scope, and constructing the methodology flowchart. The decision to pick a test sandbox environment of similar nature proved instrumental in aligning the development of CourseGPT closely with the real-world challenges encountered in professional learning and development.

Deep dives into models of corporate training, specifically the Kirkpatrick Model, provided insights into the metrics for successful training outcomes. Skill mapping using a knowledge graph emerged as a novel and efficient approach to resolving course recommendations and facilitating an effective learning journey for the students.

Working with the NPTEL team and building a skill map for IIT Madras underscored the potential of this tool in academic environments as well. Implementing Natural Entity Recognition and NLP allowed us to extract valuable course tags and derive meaningful insights from the data.

Moreover, the research conducted at HyperVerge Academy (HVA) enabled a deeper understanding of the professional upskilling space. Facing and overcoming various challenges along the way, we succeeded in developing a competency engine and data model that significantly enriched the capabilities of CourseGPT.

Finally, CourseGPT emerged as an intelligent product, leveraging the power of Knowledge Graphs to recommend courses based on career goals, plan academic journeys, solve student doubts, and inclusively cater to diverse learners. The journey, although filled with complexities and challenges, offered valuable insights into the future of AI-enabled learning and set a foundation for ongoing research and development.

The journey is far from over. There remain many more features to develop and enhancements to make. However, the strides made so far underscore the power of AI in revolutionizing corporate training and professional upskilling. CourseGPT represents a new era of AI-powered learning, driving us closer to a world where personalized, efficient, and inclusive learning is not just a vision but a reality.

References

1. Education-to-Skill Mapping Using Hierarchical Classificationand Transformer Neural Network
2. Competency based performance model for construction project managers
3. Workplace training and generic and technical skill development in the Australian construction industry
4. Kirkpatrick Model for Corporate Training Program
5. Open AI Platform docs
6. Neo4J graph Database documentation
7. Harnessing Large Language Models with Neo4j