study LLM with me

- study goal
- note-share settings
- class 1
  - what is our goal in this class?
    - become a ai- full stack engineer
  - how to study?
    - principle and theorems
    - hands-on practice
    - real world cognition
  - why study ai?
    - Industry consensus is
      - AI will inevitably reshape the world.
      - To reap the benefits of AI, one must enter the field now.
      - What needs to be solved, the technological approach, and the product strategy are all unknowns.
        - when the uncertainty becomes certainty, the importance of coding will rise
  - ai
    - what is ai?
      - what is the difference between ai and non-ai algorithm?
        - one view: only based on machine learning and Neural network.
    - what ai can do right now?
      - ✔️Enhancement of node efficiency.
        - Completing a minute segment of a certain node
      - ❌Complete workflow.
        - Unable to complete a multi-link production chain
      - how can we collaborate with AI right now?/identifying practical application scenarios.
        - 1. Familiarity with profit/output processes and methodologies = Start with familiar domains.
        - 2.1 Enhancing node efficiency = Have AI learn from the capabilities of the most skilled employees and then assist other staff members.
        - 2.2 Enhancing node efficiency = Look for scenarios where input and output are text-based.

            - 3. Think in terms of node scale = Don't aim for large and comprehensive solutions; instead, break down tasks and focus on small tasks and scenarios first.
- LLMs
    - what is LLMs
        - what is the principle of Large Language Models (LLMs)？
            - like GPT (Generative Pretrained Transformer)
                - generally speaking: LLMs are trained on vast amounts of text data(token), learning to predict the likelihood of the next word in a sentence.
                    - what is token?
                        - In the context of language models, a "token" typically refers to a piece of text that has been tokenized, or broken down into smaller, manageable pieces for processing. Tokens can be words, characters, or subwords, depending on the granularity chosen for the task at hand.
                            - what is tokenize?
                                - Tokenization is a fundamental step in natural language processing (NLP), as it allows a language model to understand and generate text by analyzing and predicting these tokens in sequence.
                    - what is this likelihood/possibility?（Not a comprehensive definition）
                        - the model learns the probability distribution of a sequence of tokens. For each input token or sequence of tokens, the model predicts a set of probabilities for what the next token could be, based on what it has learned. These probabilities are indeed stored within the neural network's structure, allowing the model to generate or continue a piece of text by selecting the next token that has the highest probability.
                    - Transformers and other architectures.
                        - transformers
                        - RWKV
                        - Mamba
            - Digital neural networks and biological neural networks are the same in mathematical principles.
                - Ilya Sutskever: i find it both obvious and incredible that a neural network is a digital brain that lives inside a computer (and that actually kinda works)
                - treat AI like a human.
        - how to work with LLMs?
            - Train foundational large models:

  - The whole world only needs few(less 1000) people to do this.
- Construct large model applications
  - All tech professionals, and even everyone should do sth with it!
  - Three modes of working with AI:Embedded, Co-pilot, Agent.
    - AI Embedded: Integrated within systems or devices to enhance functionality with AI capabilities.
    - AI Co-pilot: Works alongside humans, providing assistance and recommendations, augmenting human decision-making.
    - AI Agent: Acts autonomously, performing tasks and making decisions on behalf of humans based on learned or programmed criteria.
  - technical architecture
    - 1 Pure prompt
      - question-response
    - 2 agent+function calling
      - ai question human for extra information
    - 3 RAG(Retrieval augmented generation)
      - like open book open note test
      - The core of RAG and Fine-tuning is the use of a vector database
        - RAG augmenting the model's responses with information retrieved in real-time from external databases or document sets. (not trained by these embeddings)
      - not changing model weights
    - 4 Fine-tuning
      - close book close note test
      - The core of RAG and Fine-tuning is the use of a vector database
        - **Fine-tuning** involves taking a pre-trained base model and further training it with a specialized, often smaller and domain-specific dataset.
      - changing model weights
  - How to choose a technical approach?
    - 1 prepare test data
      - data to measure your LLM is good or not
    - 2 Validate feasibility with conversational applications
      - like testing in a chatbox with ChatGPT.
    - 3 Whether need extra supplement knowledge?
      - Yes == +RAG

  - 4 Whether need to interface with other systems?
    - Yes == +Function calling
  - 5 Whether worth trying fine-tuning?
    - Yes == using historical data for fine-tuning.
    - 3 cases for worth trying fine-tuning
      - 1 Improve the stability of LLMs.
      - 2 With a large user base, token costs are high, necessitating a reduction in inference costs.
      - 3 Increase the generation speed of large models.
  - 6 Delivered
- How to choose your foundational large model?
  - domestic
  - foreign
    - for China mainland
      - to B / to C no foreign LLMs (ChatGPT)
  - open source
    - best for data security