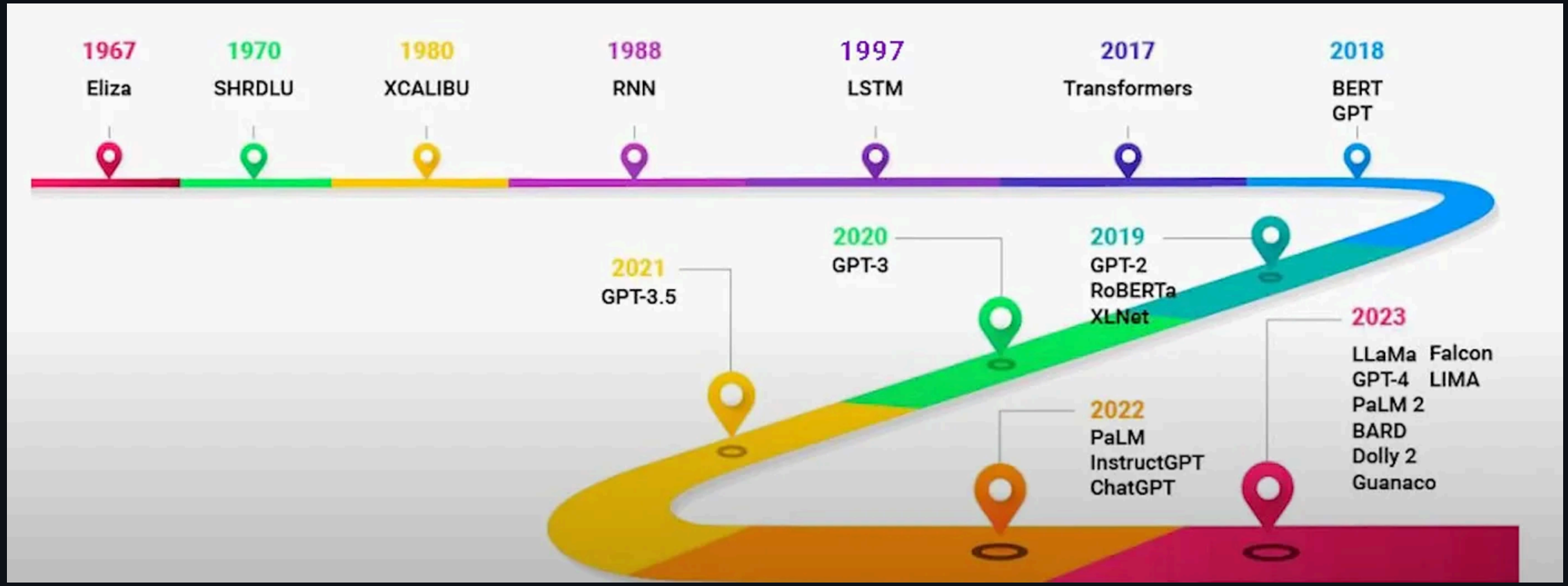
Information Technologies for Industrial Engineers

เทคโนโลยีสารสนเทศสำหรับวิศวกรอุตสาหการ

Large language model (LLM)

LLM

- A type of advanced software that can communicate in a human-like manner.
- General-purpose language processing models
  - Pre-trained on extensive datasets covering a wide range of topics.
  - Understand the fundamental structures and semantics of human language.
- `Large`
  - Substantial amount training data
  - Billions or even trillions of parameters


1967
Eliza
1970
SHRDLU
1980
XCALIBU
1988
RNN
1997
LSTM
2017
Transformers
2018
BERT
GPT
2019
GPT-2
RoBERTa
XLNet
2020
GPT-3
2021
GPT-3.5
2022
PaLM
InstructGPT
ChatGPT
2023
LLaMa Falcon
GPT-4 LIMA
PaLM 2
BARD
Dolly 2
Guanaco

Chatbots and rule-based systems (1960s)

- `ELIZA`
  - The first chatbot ever built by humans.
- It can create an illusion of a conversation
  - by rephrasing user statements as questions using pattern matching and substitution methodology.
- Try it.

Recurrent Neural Networks (1980s)

- Neural networks that can "remember" previous input.
  - Using feedback loop.
- Suitable for natural language processing (NLP) tasks.
- Still, they are not good at retaining memory and suffer from long term memory loss.
  - Vanishing gradient.

Long Short Term Memory (1990s)

- Specialised type of RNN.
- Can remember information over long sequences.
- LSTM architecture
  - `input`, `forget`, `output` gates
  - These gates determined how much information should be memorised, discarded, or output at each step.

Gated Recurrent Network (2010s)

- Designed to solve some of the same problems as LSTMs
  - but with a simple and more streamlined structure.
- GRUs architecture
  - `update` and `reset` gate
  - The reduced gating in GRUs made them more efficient in terms of computation.

Attention Mechanism (2014)

- RNN based variants LSTM and GRU are not great at retaining the context when it was far away.
- Attention allows the model to look back to the entire source sequence dynamically
  - Selecting different parts based on their relevance at each step of the output.
- Performs much better especially in longer sequences.

Transformers Architecture (2017)