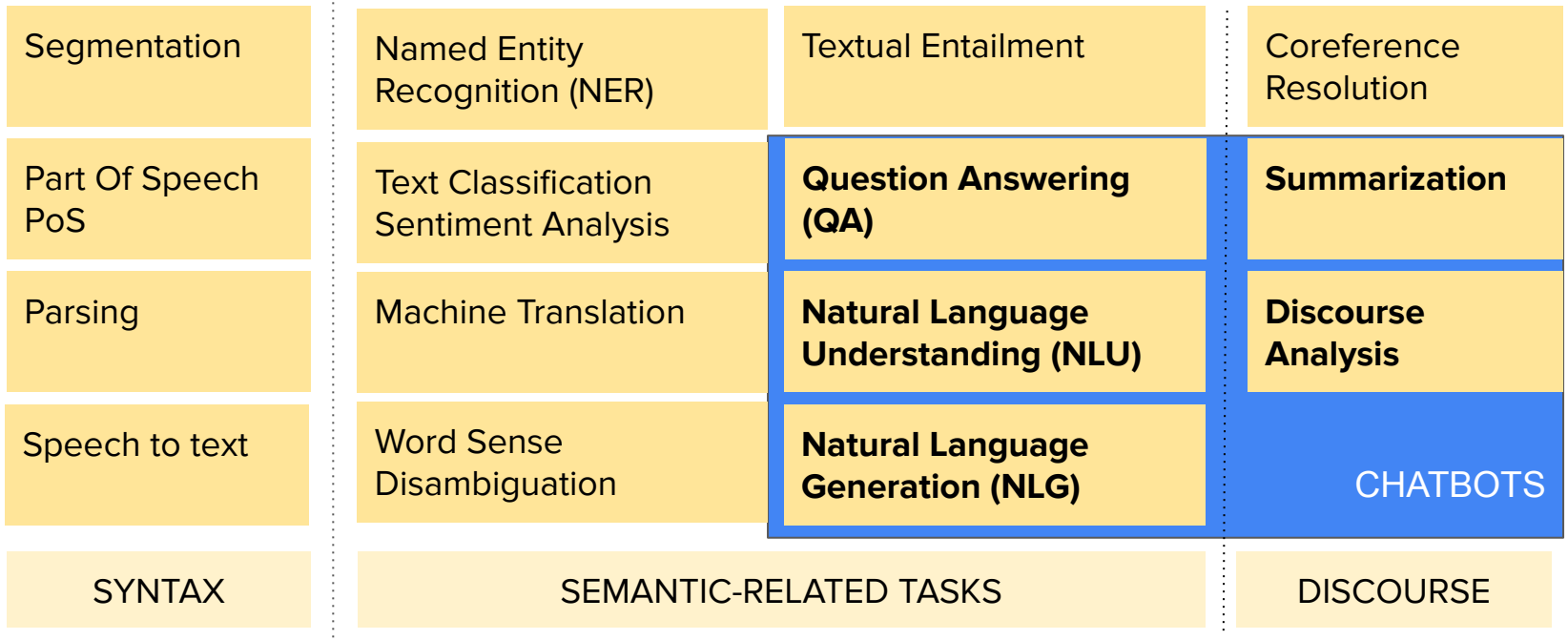
SJK006 - Master in Intelligent Systems

PROMPT DATA ENGINEERING

Learning goals

- What natural language processing (NLP) is?
- What a language model is?
- Transformers: Encoders and Decoders
- Automatic generation of texts and code
- Instructed large language models
- Practice with chatGPT over Data Science problems

Natural Language Processing Tasks



Language models (GPT-family)

GPT is a Transformer decoder aimed at completing word sequences:

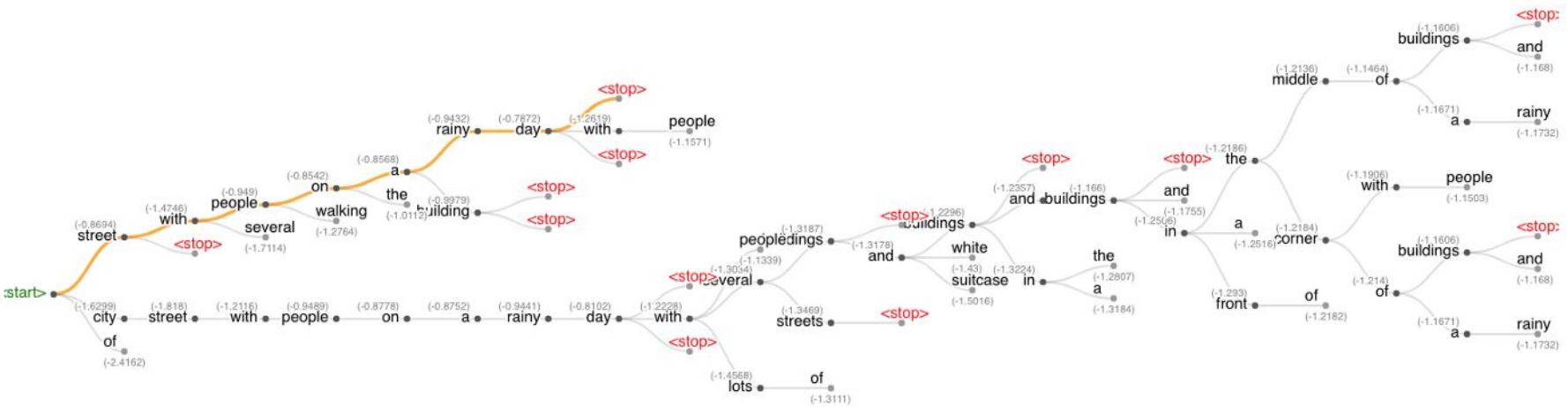
GPT training $\rightarrow p(w_i | w_0, \ldots, w_{i-1}) \rightarrow$ discrete conditional distribution of tokens

REPEAT
Predict the next token with the learnt model, add it to the current sequence
STOP if the sequence has a given length or if the sequence probability drops notably.

How tokens are chosen to generate a highly probable sequence?

- **Greedy strategy**: get always the most likely next token
- **Beam Search**: at each step keep the most likely *k* partial sequences and return the most likely one when the prediction stops
- **Softmax temperature**: we can add a hyperparameter (τ) to the output discrete distribution to smooth it (creativity/hallucination).

Beam Search



$$\text{score}(y_1 \ldots y_t) = \sum_{i=1 \ldots t} \log P(y_i | y_1 \ldots y_{i-1}, x)$$

Beam search decoder with k=3 and max steps as 51

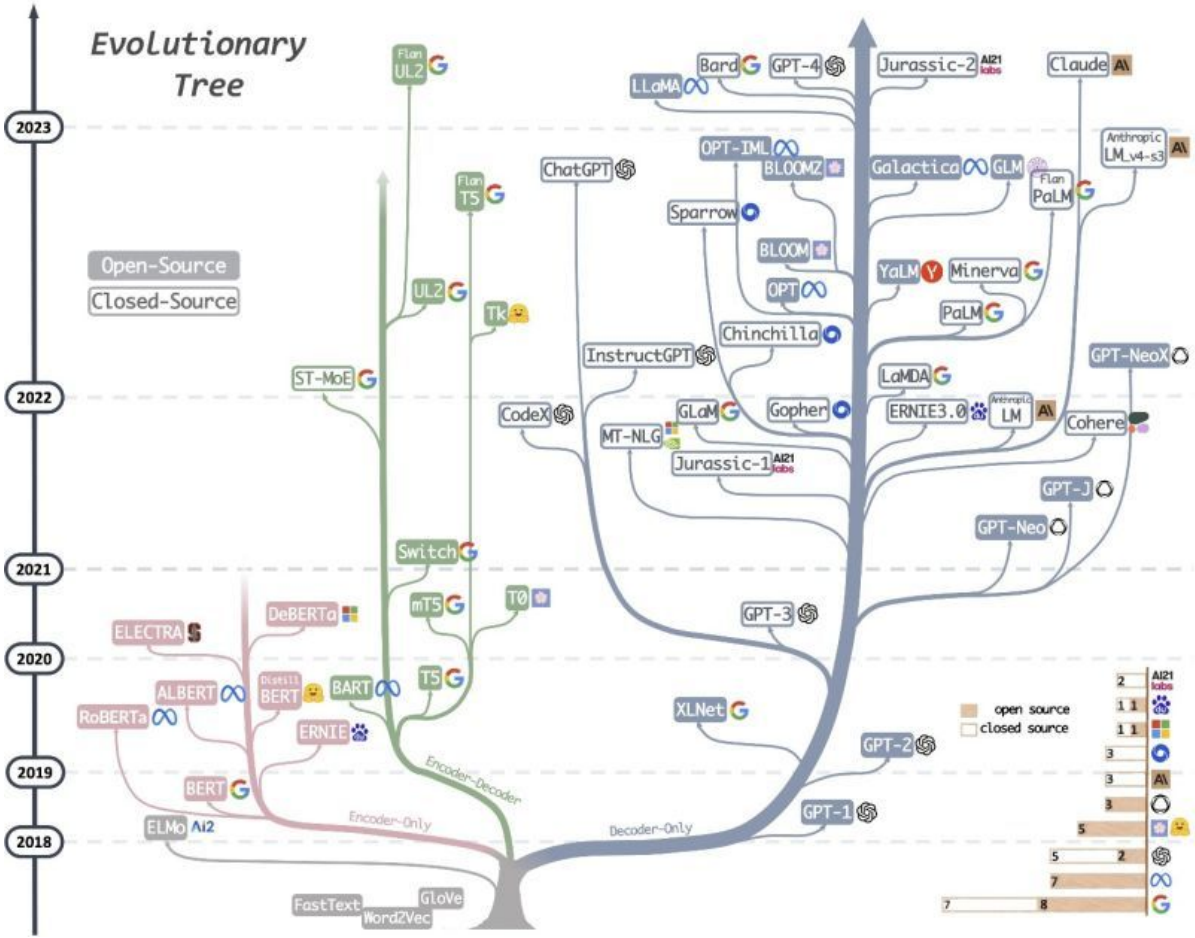
Figure from:
https://medium.com/voice-tech-podcast/visualising-beam-search-and-other-decoding-algorithms-for-natural-language-generation-fbba7cba2c5b

Large language models (LLM)

- Enormous pre-trained models with masking task (decoder-only)
  - GPT-2, GPT-3, GPT-3.5, GPT4 (Open AI)
  - LLAMA 3 (Meta, open source)
  - Gemini, Gemma (Google)
  - Claude (Antrhopic)
  - Falcon (Open source)

https://lmarena.ai/
https://openlm.ai/chatbot-arena/

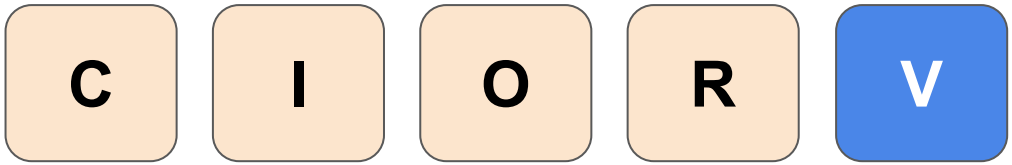
- All of them have more than 10^9 parameters (billions), the size usually is included in the name of the model (e.g. Falcon-7B, Falcon-13B).
- This models capture a lot of semantic and causality patterns of the common sense and specialised domains.
- The main task/purpose of these models is to “complete” an initial text, which is called PROMPT.
- **Instruct models** have been trained to guide the completion according to a series of orders about the length, content, constraints, etc.
- The PROMPT can include a few examples of what we want to get (few shot learning)


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
Flan T5
UL2
Tk
ST-MoE
Switch
mT5
T0
DeBERTa
ELECTRA
ALBERT
DistilBERT
BART
T5
RoBERTa
ERNIE
BERT
ELMo
Ai2
FastText
GloVe
Word2Vec
Encoder-Only
Encoder-Decoder
Decoder-Only
LLaMA
Bard
GPT-4
Jurassic-2
Claude
OPT-IML
BLOOMZ
ChatGPT
Galactica
GLM
Anthropic LM_v4-s3
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
AI21 labs

Instructed LLMs

- Instructed models are LLMs that have been trained to complete instructions (PROMPT)
- chatGPT, Llama-2 and Bard are instructed models.
- Outputs of these models have been supervised by humans through a mechanism called Reinforcement Learning from Human Feedback (RLHF).
- Instructed LLMs are aimed at reduce hallucinations (false predictions) and maximise the alignment with the ethical and legal principles of humans.
- The main limitation of some LLMs are the prompt and output sizes. (~ 4K-8K tokens)
    - We are forced to retrieve the relevant information to be included in the PROMPT
    - We need to integrate LLMs with other existing tools (LANGUAGE CHAINS and Plugins) to complement each other

Basic principles of prompting (chatGPT)



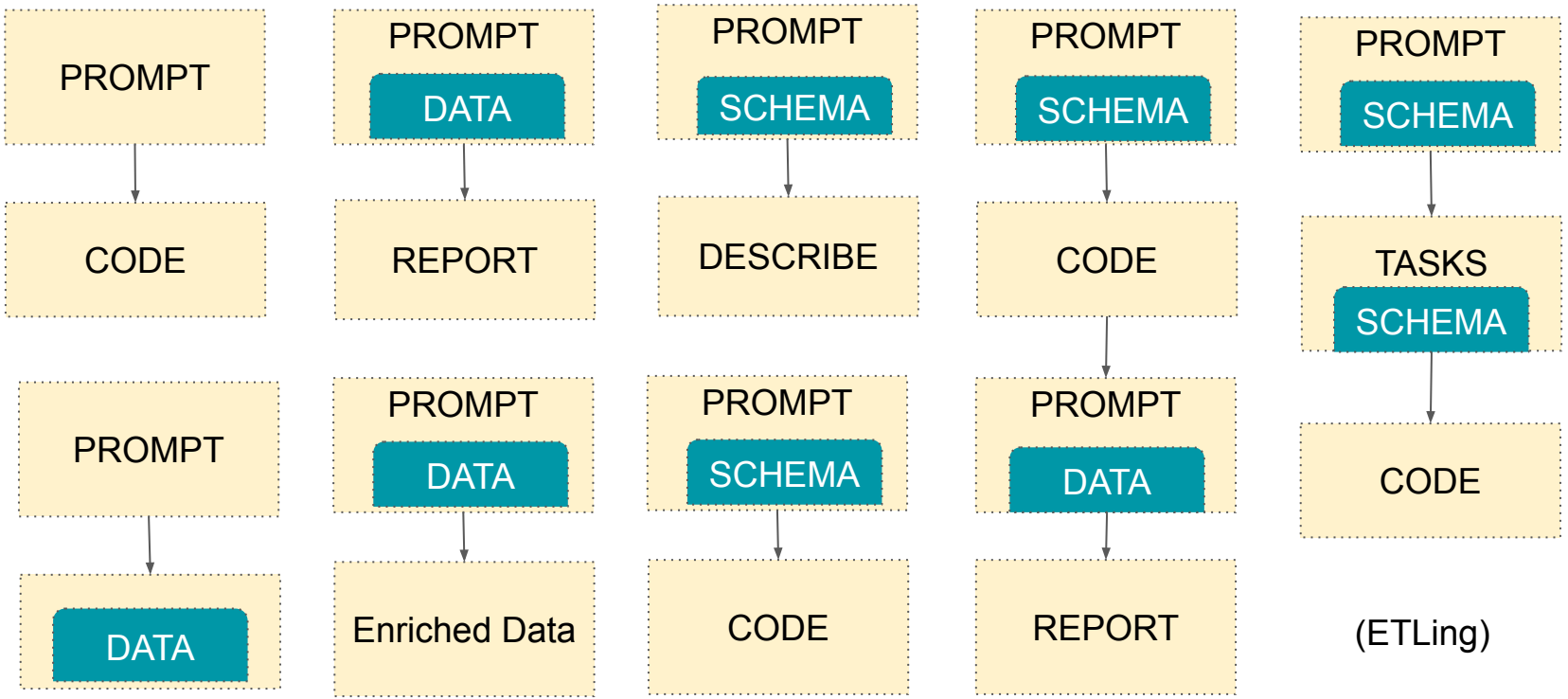
Context: describe without ambiguities the scenario and intent of the prompt

Input: what is being used as input
Output: which is the desired output format (text, JSON, CSV, etc.)
Few shot: write some few examples of input/ouput

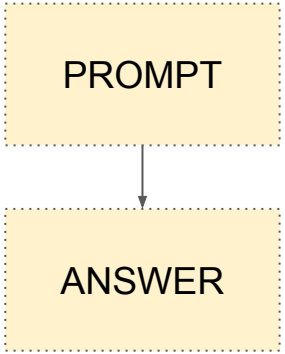
Rephrase: adjust the prompt to be more precise about the intent and the output (errors in the output).
Validate: detect errors, show them to the agent and find alternative outputs.

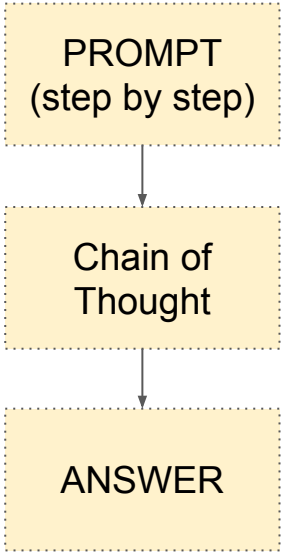
Identify **parameters** to re-use the prompt at scale!

Ways to use prompting for data science

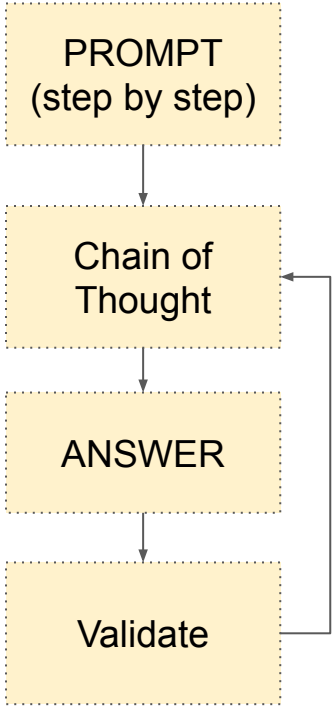


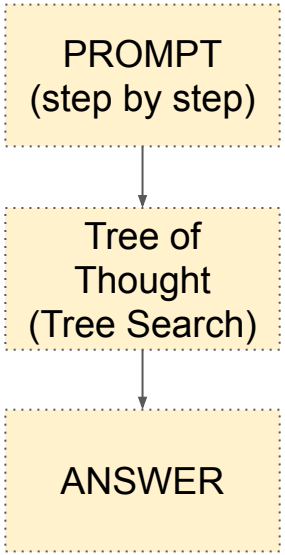
Prompt beyond QA: explain & problem solving



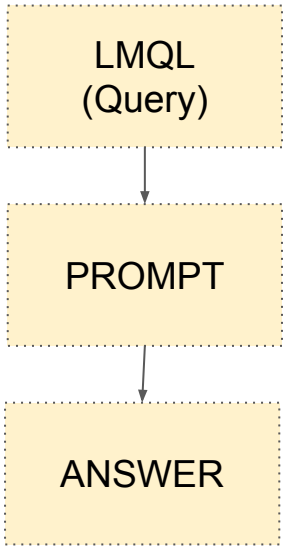


Chain of Thoughts
(step-by-step)





Tree of Thoughts



https://lmql.ai/

LET’S PRACTICE PROMPTING FOR DATA SCIENCE!