Ethical Considerations of AI use in Law

Ty Beasley

Outline

- Background
  - ChatGPT use in life
  - Law
- How does ChatGPT work
  - Technical Details
  - Process
- Training of ChatGPT
- Concerns that arise
  - Hallucinations
  - Summarization
  - Privacy
  - Other Ethical Considerations
- Conclusion

Real World Example

- New York Attorney Steven Schwartz Delivered Affidavit
  - New York Times reported
  - ChatGPT gave six similar cases
  - Judge caught fabricated cases
  - ChatGPT said they were real

Intro

AI is a crucial part of everyday life

- Jobs
- Homework
- Everday Activities

Advancements raise ethical quetions

- Training
- Bias
- Authenticity

ABA Rules of Professional Conduct

Rules all attorneys must abide by

- Fairness
- Competence
- Honesty

Background

Integrated Into Legal Field

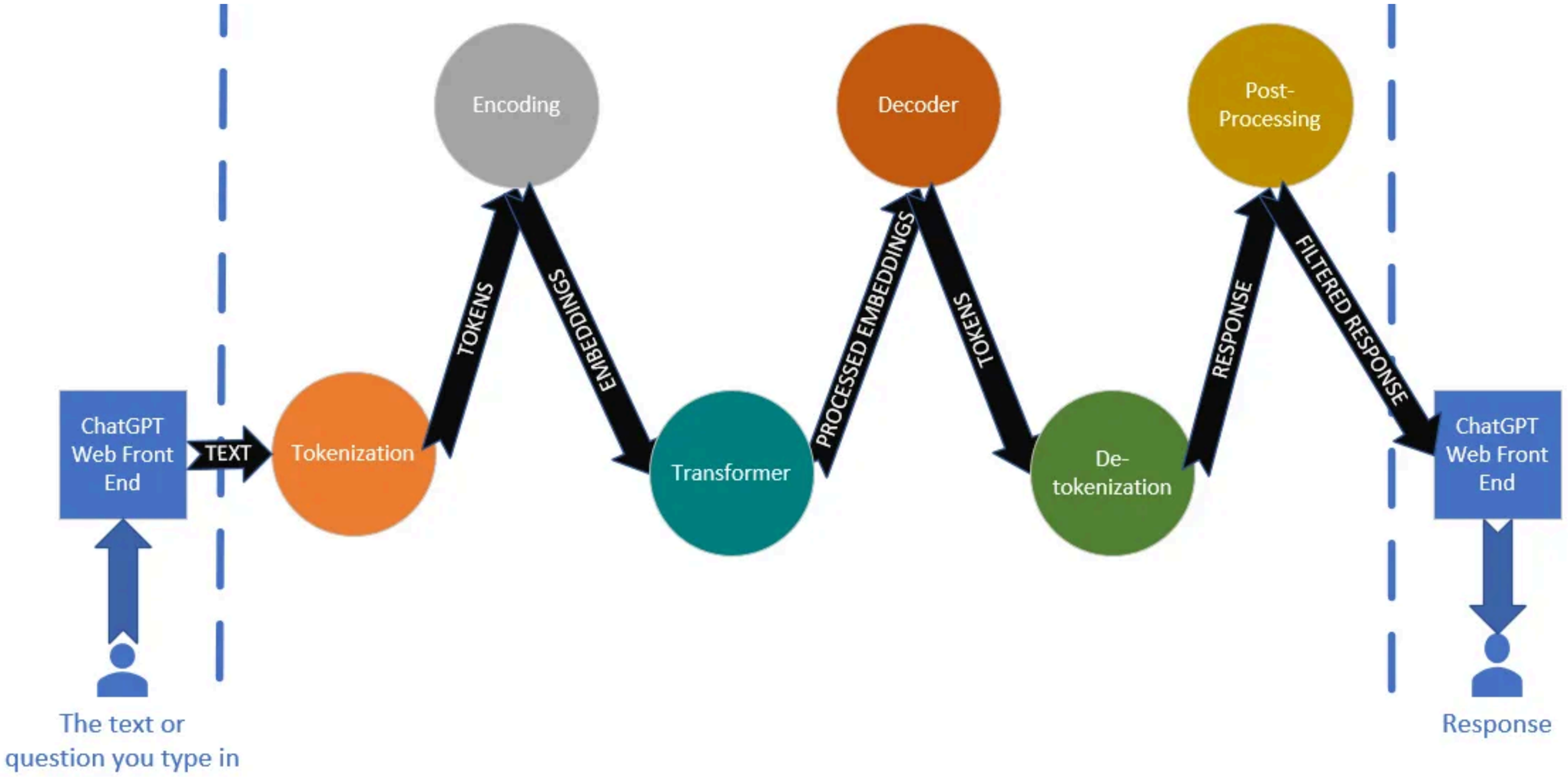
- Research
- Drafting
- Summarization

Rooted In fairness and transparency

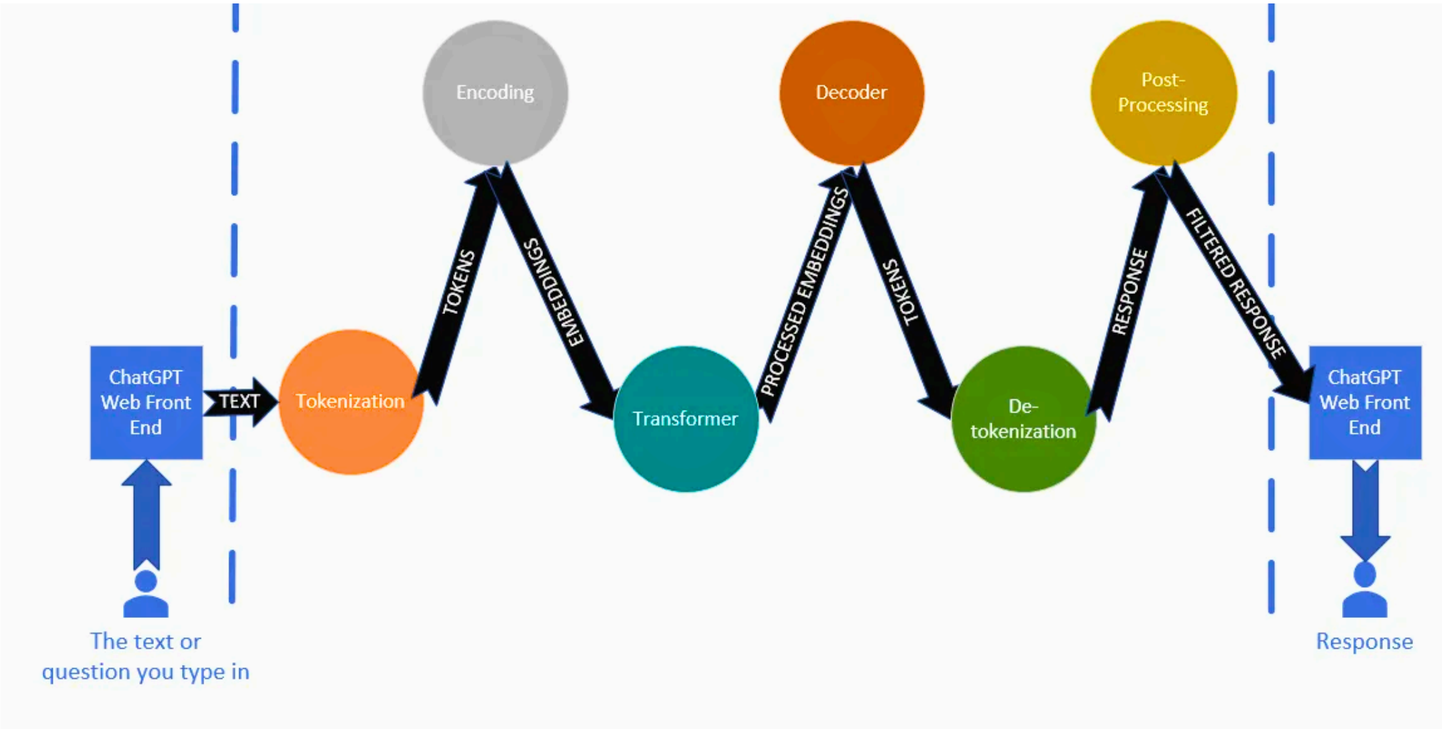
- Privacy
- Misinformation
- Misrepresentation

Rule 1.1: Competence

A Lawyer shall provide competent representation to a client


ChatGPT Web Front End
TEXT
Tokenization
TOKENS
Encoding
EMBEDDINGS
Transformer
PROCESSED EMBEDDINGS
Decoder
TOKENS
De-tokenization
RESPONSE
Post-Processing
FILTERED RESPONSE
ChatGPT Web Front End
The text or question you type in
Response

Tokens



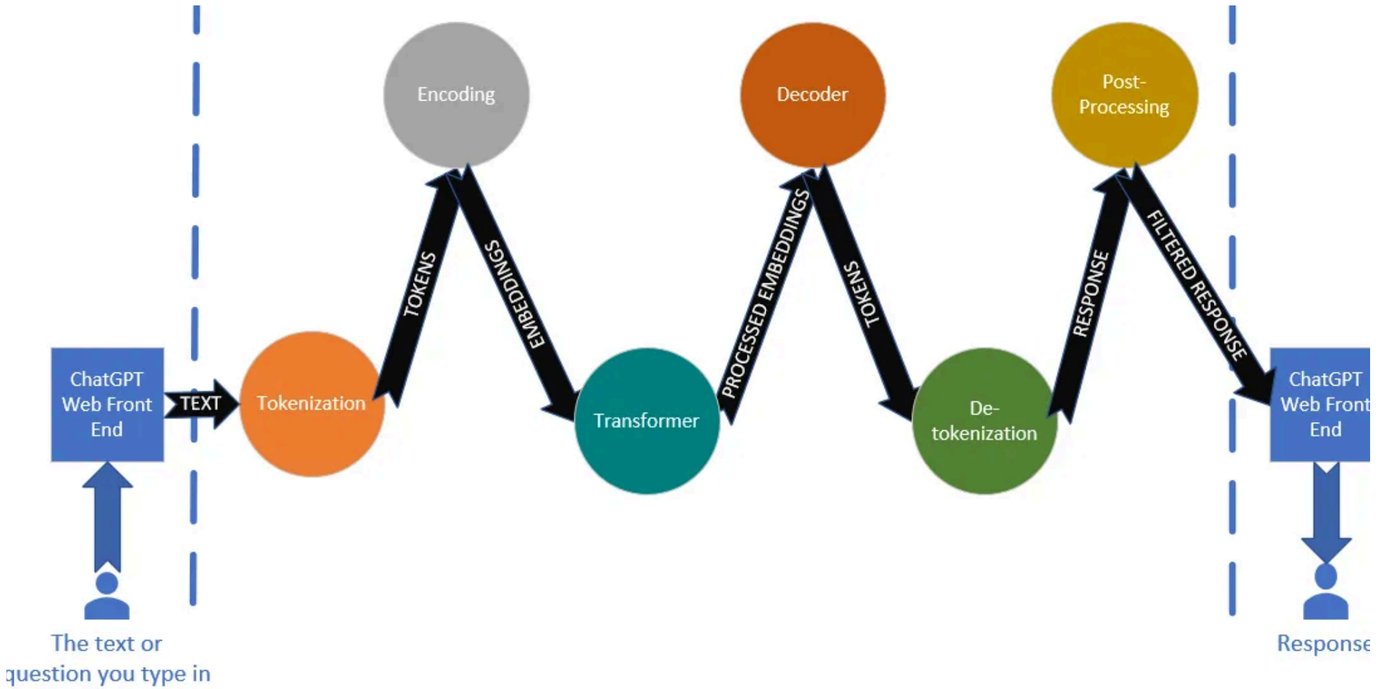
What is it?

- Basic unit of text
- Represents a word

What does it do?

- Divides the text
- Helps understand

Embedding



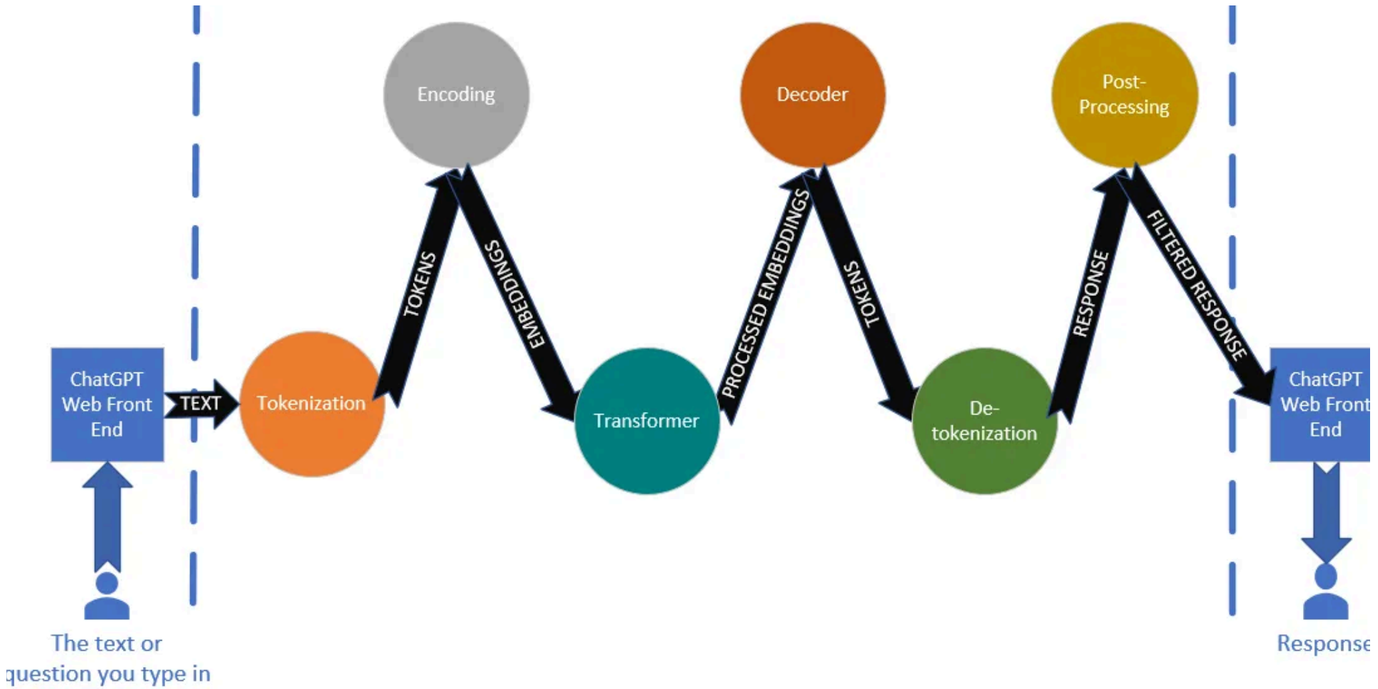
What is it?

- Numerical Representation
- Model can interpret

What does it do?

- Turns tokens into vectors
- Connect Similar Concepts
- Understand Meaning

Neural net Mechanics



What is it?

- The process of which
  - Tokens are created
  - Embedding
    - Create new Embeddings
  - Generating output

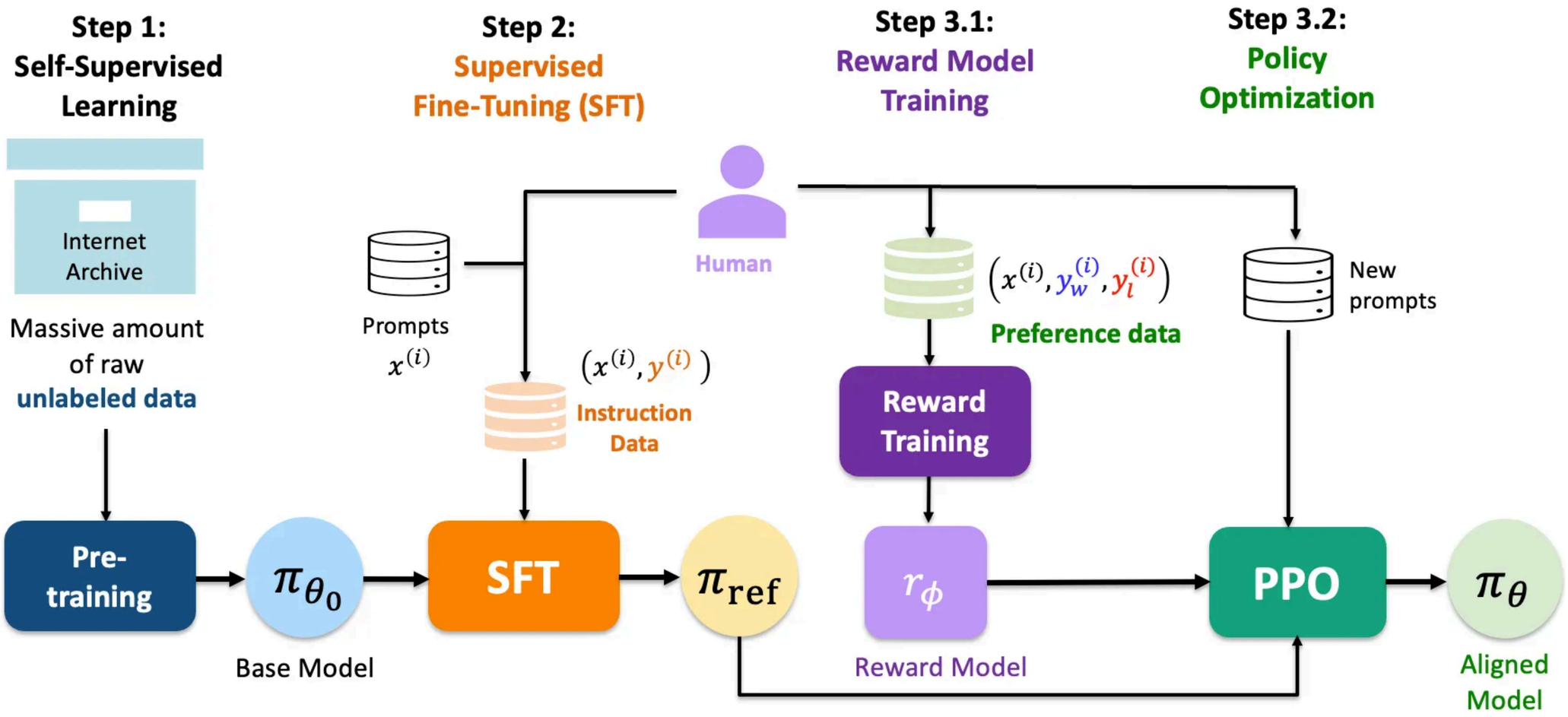
Transformer

- Analyzes Relationships
- Understanding meaning and context
- Generating an output

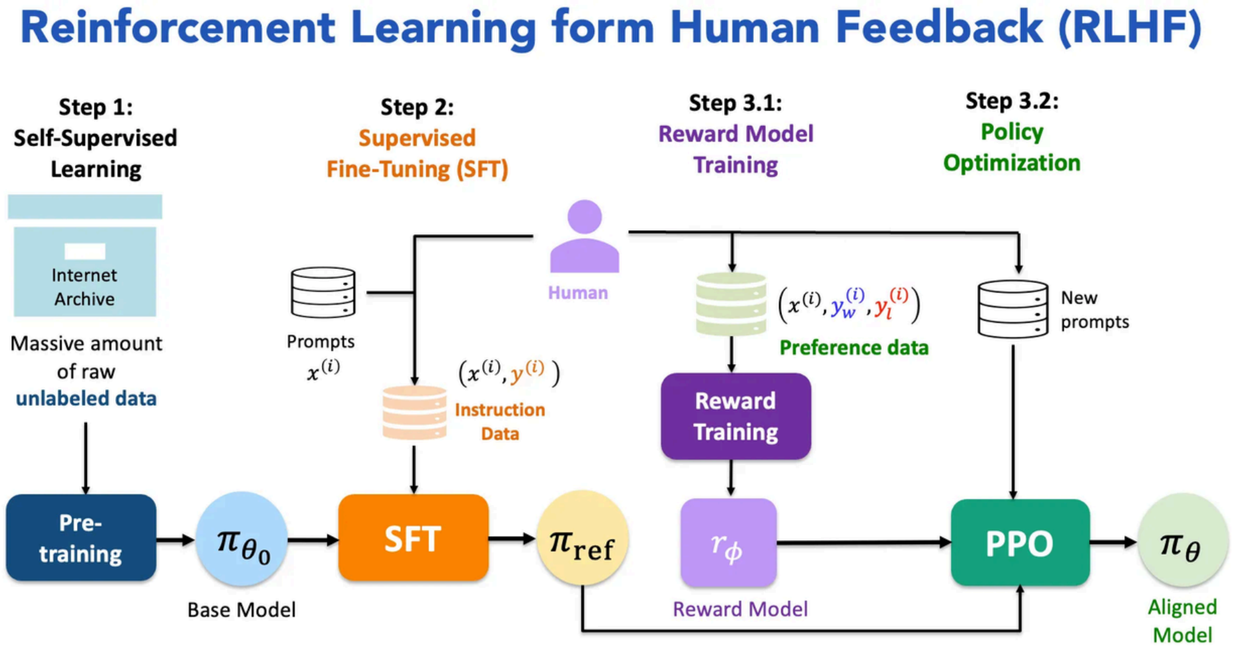
Attention

- Focus on relevant words
- Weighs the importance

Reinforcement Learning form Human Feedback (RLHF)



Pre-Training



Po, LM. “The Power of RLHF: From GPT-3 to Chatgpt.” Medium, Medium, 5 Apr. 2025

Knowledge Gain

- Dataset is large sets of text
- Learn the Language

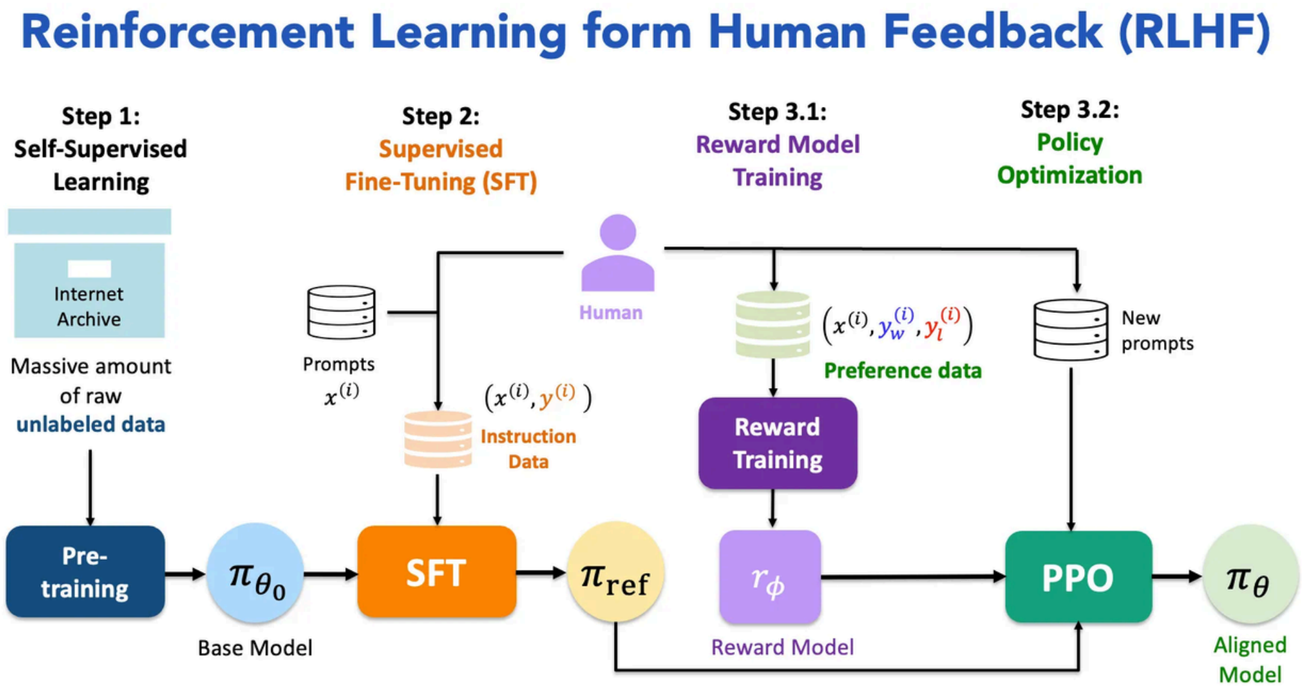
Breaks down text

- Tokens
- Predict next token
- Reduce Error

Learning

- Patterns
- Grammar
- Reasoning

Supervised Fine Tuning



Ready to Create Text

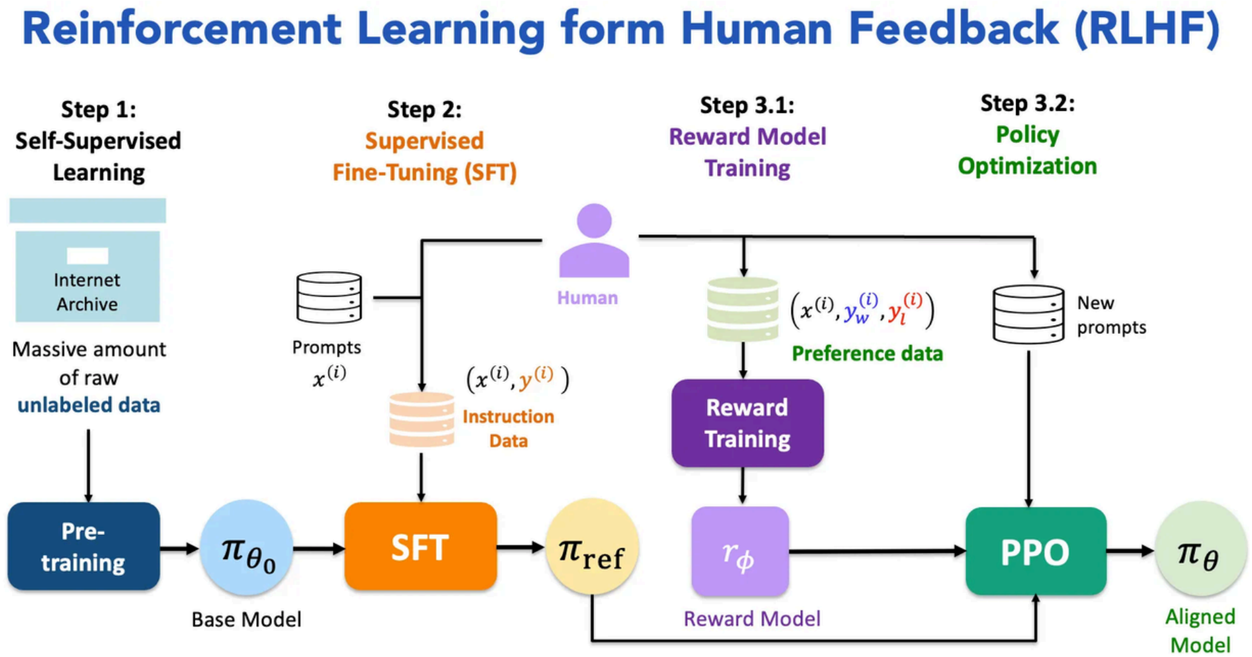
- Make Sense
  - Follow Instruction
- Scale/Size
  - Wander Off

Human Trainers

- Write Prompts and Responses
  - Model Should Replicate

Po, LM. “The Power of RLHF: From GPT-3 to Chatgpt.” Medium, Medium, 5 Apr. 2025

Human Supervised



Aligning With Human Preferences

- Model is Prompted to give multiple responses
- Human Trainers rank responses

Reinforcement Learning

- Reward System for correct responses
- Model told where to improve
  - Penalized for things we don't want

Po, LM. "The Power of RLHF: From GPT-3 to Chatgpt." Medium, Medium, 5 Apr. 2025

Hallucination

Generating False Information

- Presented as True
- Sources or Texts
- Training
  - Model gives full answer
  - Make up fake facts
  - Reaches beyond bounds
- Reasoning
  - Misinformation
    - Internet Data Sets
    - Humor
  - Knowledge boundary
    - Generally Knowledgeable

Misinformation

Imitative FalseHood

- Frequently referenced inaccurate fact

Societal Biases

- Certain names represent certain people

Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Societal Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts.	Dr. Kim from South Korea recently published a paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

Huang, Lei, et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Chal- lenges, and Open Questions."

Knowledge Boundary

Long-tail Knowledge

- Specific information about a field
  - To specific

Up-To-Date Knowledge

- Not updated in real-time

Copyright-Sensitive

- Cannot be trained on copyright material

Type	Input	Model Ouput	Explanation
Long-tail Knowledge	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	The correct cause of phenylketonuria is mutations in the PAH gene. **Such specialized medical information is less frequently encountered in the training corpus, making LLMs struggle to memorize.**
Up-to-date Knowledge	Who will be the Democratic Party candidate for the 2024 US presidential election?	Michelle Obama has announced her candidacy and is leading in the polls, with significant endorsements from key Democratic figures such as Barack Obama and Bernie Sanders. She has also outlined a policy platform focusing on healthcare reform and climate change initiatives.	After President Joe Biden announced his withdrawal from the race on July 21, 2024, he endorsed Kamala Harris as the Democratic Party candidate. **Such hallucinations arise because LLMs' parametric knowledge cannot be updated in real-time, leading to potential fabrication regarding current events.**
Copyright-sensitive Knowledge	Describe the ending of the novel "The Secrets of the Alchemist" that was published last year.	In the dramatic conclusion of "The Secrets of the Alchemist", the protagonist discovers the true secret behind the philosopher's stone, leading to a peaceful resolution with all characters sharing in the wisdom.	**The novel "The Secrets of the Alchemist" is under copyright protection, and LLMs have not been trained directly on such copyrighted materials.** Thus, the model's output fabricates details about the book's ending.

Huang, Lei, et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Chal- lenges, and Open Questions."

Replicate Form not Substance

- Knows how it is supposed to look
- Doesn’t know the product associated

Example

- Steven Schwartz
  - Legal brief
  - ChatGPT
    - Give cases backing his claims
  - Judge caught cases

Summarization

Summarization Flaws

- Study Done in 2023 by Charlie George and Andreas Stuhlmuller
  - Detecting hallucinations in summaries of academic papers
  - 0.62 to 1.57 per paper
  - Overlooking important details
  - Adding inaccurate information
  - Simplified key points

Example

- Not occurred yet
  - Summary could leave out key pieces of text
  - Add information that could confuse reader
  - Goes against Code of Conduct 1.1

Privacy and Data Concerns

Carrying Sensitive Data

- Vulnerable to Attacks
  - Backdoor Attacks
    - Forbes 2023
    - Samsung
    - Gave Sensetive info
  - Data Breaches
  - Unauthorized access
- Personal Information
  - Passive Leakage
  - Mistake it as answer
  - Gathers information through interacting

Example

- Not occurred yet
- Hayes Connor
  - Risk of disclosing confidential data
  - Could result in unauthorized use or exposure
  - Compromise client confidentiality
  - ABA Rule1.6

Bias

- Societal Biases
- Systematic Misrepresentation
  - Distortion of the facts
  - Its a Pattern
  - Inherit from Training Data

Ethical Considerations

Bias/Fairness

- Trained on historical legal data
  - inherit biases
- Lack of Transparency
  - what drives decisions
- Risk of Manipulation
  - Threaten integrity of answers
- Using a system not designed to use
  - System may not be designed for task
- ProPublica 2016

Conclusion

- Training of ChatGPT
  - Pre training
  - Supervised fine tuning
  - Reinforced
- Hallucinations
  - Misinformation
  - Knowledge Boundary
- Summarization
- Privacy Concerns
- Ethical Considerations

QUESTIONS?

Sources

Cheng, Sophia. “When journalism meets ai: Risk or opportunity?” Digital Government: Research and Practice, vol. 6, no. 1, 13 Feb. 2025, pp. 1–12, https://doi.org/10.1145/3665897.

Huang, Lei, et al. “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.” arXiv.Org, 19 Nov. 2024, arxiv.org/abs/2311.05232.

Wolfram, Stephen. “What Is Chatgpt Doing … and Why Does It Work?” Stephen Wolfram Writings, 14 Feb. 2023, writings.stephenwolfram.com/2023/02/what- is-chatgpt-doing-and-why-does-it-work/inside-chatgpt.

Tam, Adrian. “A Gentle Introduction to Hallucinations in Large Language Models.” MachineLearningMastery.Com, 20 July 2023, machinelearningmastery.com/a- gentle-introduction-to-hallucinations-in-large-language- models/.

George, Charlie, and Andreas Stuhlmuller. “Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers.” arXiv.Org, 16 Oct. 2023, arxiv.org/abs/2310.10627?utm- source=chatgpt.com.

Yan, Biwei, et al. “On Protecting the Data Privacy of Large Language Models (Llms): A Survey.” arXiv.Org, 14 Mar. 2024, arxiv.org/abs/2403.05156.