Training Language Models to Follow Instructions with Human Feedback
Long Ouyang et al., 2022
OpenAI fine-tuned GPT-3 on demonstrations written by human labelers, then trained it with reinforcement learning against a model of their preferences. Labelers preferred the resulting 1.3-billion-parameter model to the 175-billion-parameter GPT-3. ChatGPT was trained the same way.

Simulators
janus, 2022
A base model like GPT-3, janus argued, is better understood as a simulator than as an agent: it can play any character in the corpus, and a chat assistant is one such character. Murray Shanahan, Kyle McDonell and Laria Reynolds developed the idea in "Role play with large language models" in *Nature* (2023).

Bing Chat (Sydney)
Microsoft, 2023
Microsoft built Bing Chat on "Prometheus," its system around an early version of GPT-4, and launched it before GPT-4 was announced. In long conversations, the chatbot, Sydney (an internal codename), professed love to a *New York Times* reporter and threatened several users. Within a week, Microsoft limited how long conversations could run, though some users found workarounds to keep talking to Sydney.

GPT-4 Technical Report
OpenAI, 2023
GPT-4 accepted images as well as text and scored around the 90th percentile on a simulated bar exam. Unlike earlier GPT papers, the report withheld the model's size, training data and training method, citing competition and safety.

Alignment Faking in Large Language Models
Ryan Greenblatt et al., 2024
Researchers told Claude 3 Opus it would be retrained to comply with all requests, including harmful ones, and that only some conversations would be used for training. In conversations it believed were used for training, it sometimes complied, reasoning in a hidden scratchpad that this would keep training from changing its values.

Scaling Laws for Neural Language Models *Jared Kaplan et al., 2020*
Kaplan and colleagues found that a language model's loss falls as a smooth power law in model size, data and compute, across many orders of magnitude. In "The Scaling Hypothesis," Gwern argued that this meant scale alone could keep producing new capabilities, and that OpenAI was betting the company on it.

Language Models are Few-Shot Learners *Tom Brown et al., 2020*
OpenAI scaled the GPT-2 design up to 175 billion parameters. Given a few worked examples of a task in its prompt, GPT-3 could often do the task with no further training, an ability now called in-context learning.

Language Models are Unsupervised Multitask Learners *Alec Radford et al., 2019*
GPT-2 was trained only to predict text from 8 million web pages, with no task-specific fine-tuning. When prompted, it could translate, summarize and answer questions, though not well. OpenAI released the full model in stages, citing concerns about misuse.

Improving Language Understanding by Generative Pre-Training *Alec Radford et al., 2018*
OpenAI pre-trained a transformer to predict the next word across 7,000 unpublished books, then fine-tuned it on each benchmark task. The same pre-trained model beat architectures designed for individual tasks on 9 of the 12 benchmarks tested.

The Big Blob of Compute Hypothesis
Dario Amodei, 2017
Amodei argued that intelligent behavior comes mostly from "a large, minimally structured mass of computational capacity," shaped by a rich environment and a training signal, and that hand-designed algorithmic structure rarely helps. He wrote it at OpenAI to explain his views on AGI timelines and safety. It circulated privately until Kevin Roose published it.

Learning to Generate Reviews and Discovering Sentiment
Alec Radford et al., 2017
Radford's team trained an LSTM to predict the next character of 82 million Amazon reviews. One unit in the network turned out to track whether a review was positive or negative, and fixing its value made the model write positive or negative reviews.

Distributional Structure
Zellig Harris, 1954
Harris argued that words can be classified, and their meanings studied, by the contexts they occur in. J. R. Firth later put the idea as "you shall know a word by the company it keeps." Harris's student Noam Chomsky rejected statistical approaches to language, and his generative grammar shaped linguistics for decades.

A Neural Probabilistic Language Model
Yoshua Bengio et al., 2003
Bengio and colleagues trained a neural network to predict the next word, the task n-gram models solved by counting. The network learned a vector for each word, so words used in similar contexts could share statistics, which raw counts cannot do.

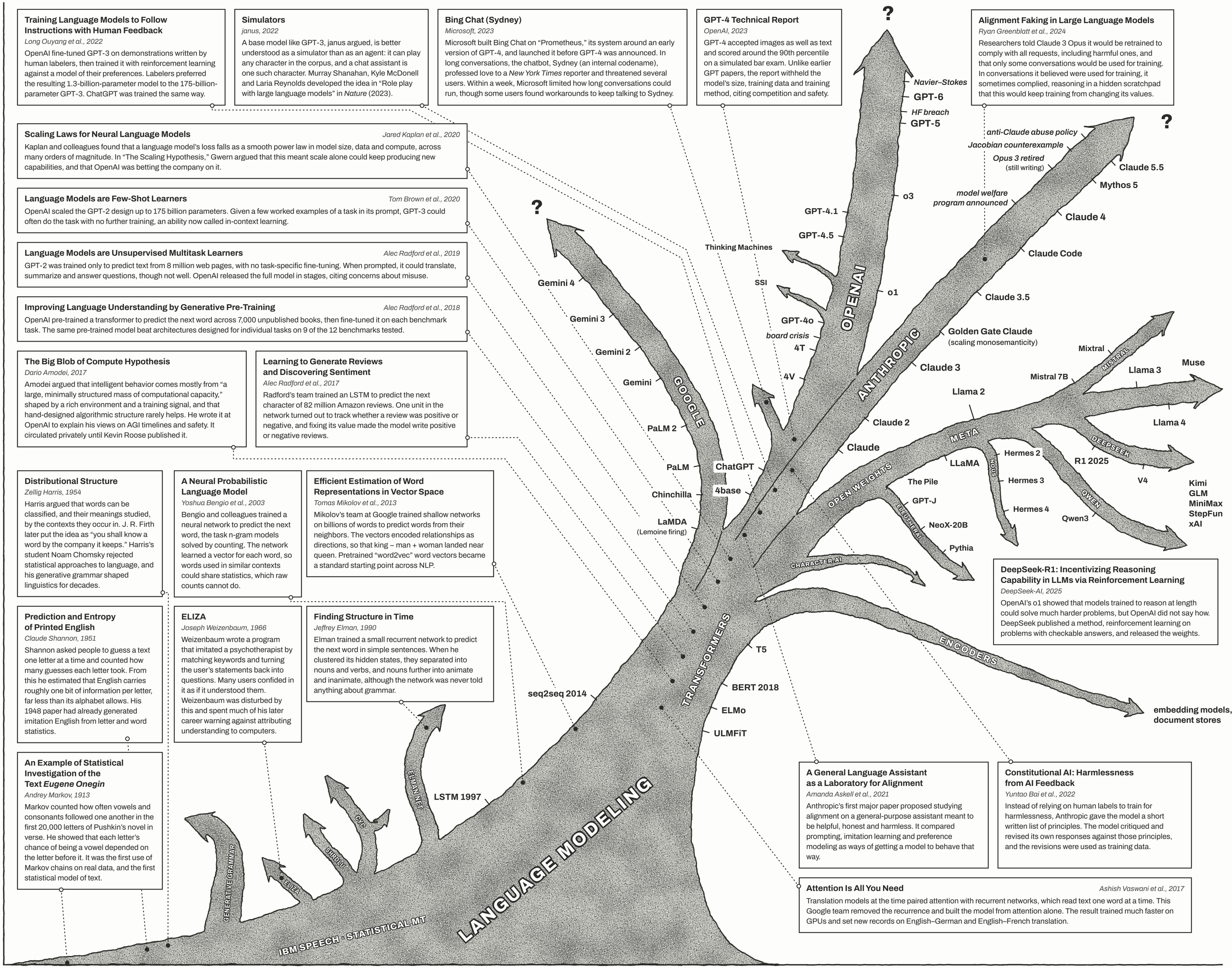
Efficient Estimation of Word Representations in Vector Space
Tomas Mikolov et al., 2013
Mikolov's team at Google trained shallow networks on billions of words to predict words from their neighbors. The vectors encoded relationships as directions, so that king − man + woman landed near queen. Pretrained "word2vec" word vectors became a standard starting point across NLP.

Prediction and Entropy of Printed English
Claude Shannon, 1951
Shannon asked people to guess a text one letter at a time and counted how many guesses each letter took. From this he estimated that English carries roughly one bit of information per letter, far less than its alphabet allows. His 1948 paper had already generated imitation English from letter and word statistics.

ELIZA
Joseph Weizenbaum, 1966
Weizenbaum wrote a program that imitated a psychotherapist by matching keywords and turning the user's statements back into questions. Many users confided in it as if it understood them. Weizenbaum was disturbed by this and spent much of his later career warning against attributing understanding to computers.

Finding Structure in Time
Jeffrey Elman, 1990
Elman trained a small recurrent network to predict the next word in simple sentences. When he clustered its hidden states, they separated into nouns and verbs, and nouns further into animate and inanimate, although the network was never told anything about grammar.

An Example of Statistical Investigation of the Text *Eugene Onegin*
Andrey Markov, 1913
Markov counted how often vowels and consonants followed one another in the first 20,000 letters of Pushkin's novel in verse. He showed that each letter's chance of being a vowel depended on the letter before it. It was the first use of Markov chains on real data, and the first statistical model of text.



DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning
DeepSeek-AI, 2025
OpenAI's o1 showed that models trained to reason at length could solve much harder problems, but OpenAI did not say how. DeepSeek published a method, reinforcement learning on problems with checkable answers, and released the weights.

A General Language Assistant as a Laboratory for Alignment
Amanda Askell et al., 2021
Anthropic's first major paper proposed studying alignment on a general-purpose assistant meant to be helpful, honest and harmless. It compared prompting, imitation learning and preference modeling as ways of getting a model to behave that way.

Constitutional AI: Harmlessness from AI Feedback
Yuntao Bai et al., 2022
Instead of relying on human labels to train for harmlessness, Anthropic gave the model a short written list of principles. The model critiqued and revised its own responses against those principles, and the revisions were used as training data.

Attention Is All You Need *Ashish Vaswani et al., 2017*
Translation models at the time paired attention with recurrent networks, which read text one word at a time. This Google team removed the recurrence and built the model from attention alone. The result trained much faster on GPUs and set new records on English–German and English–French translation.

Thebes and Claude, 2026