Northwestern Kellogg

Northwestern ENGINEERING

MBAi 448 | Winter 2026

Transformers

Today

- Assignment 4 / week 5 takeaways *10 minutes*
- Transformers, so what? *10 minutes*
- What is the right AI to use? *20 minutes*
- Quick break *5 minutes*
- Transformers case studies *25 minutes*
- Assignment walkthrough *20 minutes*

Assignment 4 / week 5 takeaways

Transformers

Week 5 takeaways

- **Transformers fundamentally changed how machines understand language—and beyond.**
  Unlike earlier models that processed text sequentially, transformers use attention to evaluate all parts of an input at once, enabling far better handling of context, ambiguity, and long-range dependencies. This architectural shift underpins today's large language models and is now extending to images, audio, and video.

- **"Attention" is the business-relevant breakthrough.**
  Attention mechanisms allow models to focus on what matters most in an input (e.g., key words in a customer request or critical clauses in a document). For businesses, this enables more accurate classification, summarization, routing, and generation—core capabilities behind copilots, chatbots, and automation tools.

- **Different transformer architectures map to different business tasks.**
  Encoder models (e.g., BERT) excel at understanding and classification, decoders (e.g., GPT) at generation, and encoder-decoder models (e.g., T5, BART) at transformation tasks like translation or summarization. Choosing the right architecture is a strategic product decision, not just a technical one.

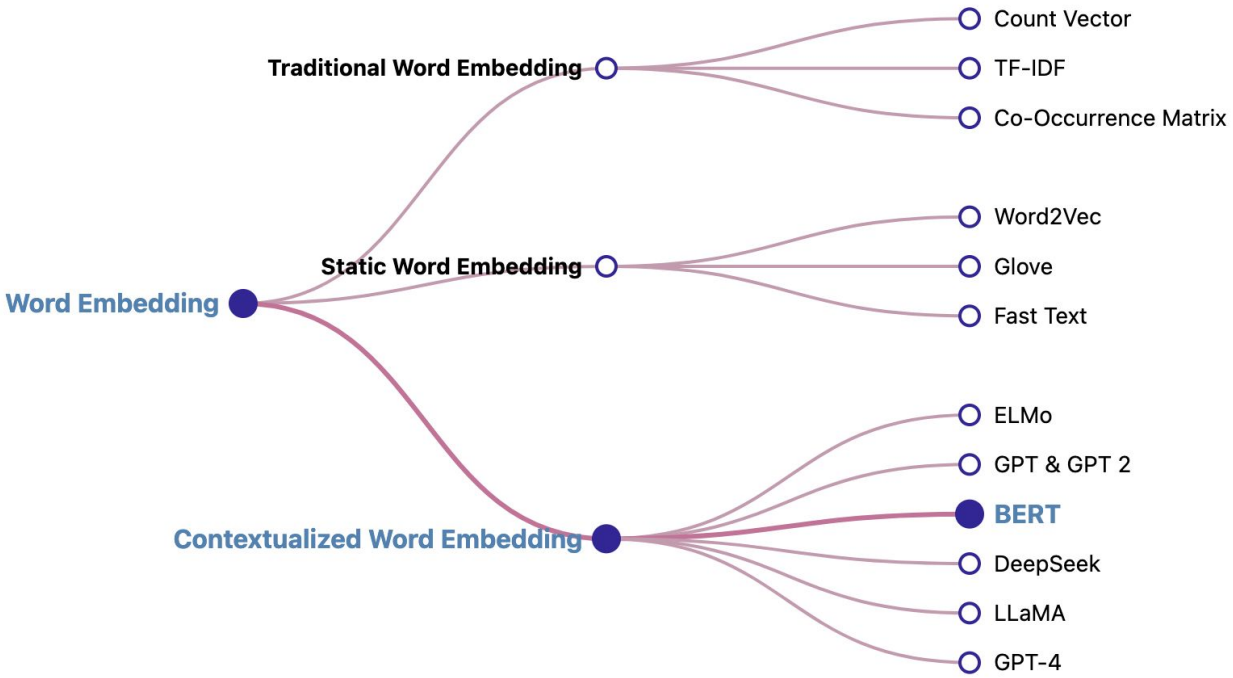
Week 5 takeaways

- **Self-supervised learning ~~unlocked scale—and lowered dependency on~~ *enables deep learning at scale in the absence of explicitly* labeled data.**
  Transformers learn from massive amounts of raw data by predicting missing or reordered information, eliminating the need for expensive human labeling. This is why AI capabilities scaled so rapidly—and why data access, not algorithms, is often the key competitive advantage.

- **Benchmark performance doesn't equal business readiness.**
  While models can achieve "gold-medal" scores on standardized tests, real-world deployment *tasks rarely resemble standardized tests. Further, the models themselves are but statistical learners and so implementations must account for* ~~introduces risks:~~ hallucinations, bias, ambiguity, and data drift. Effective applications require guardrails, evaluation frameworks, and thoughtful human-in-the-loop design.

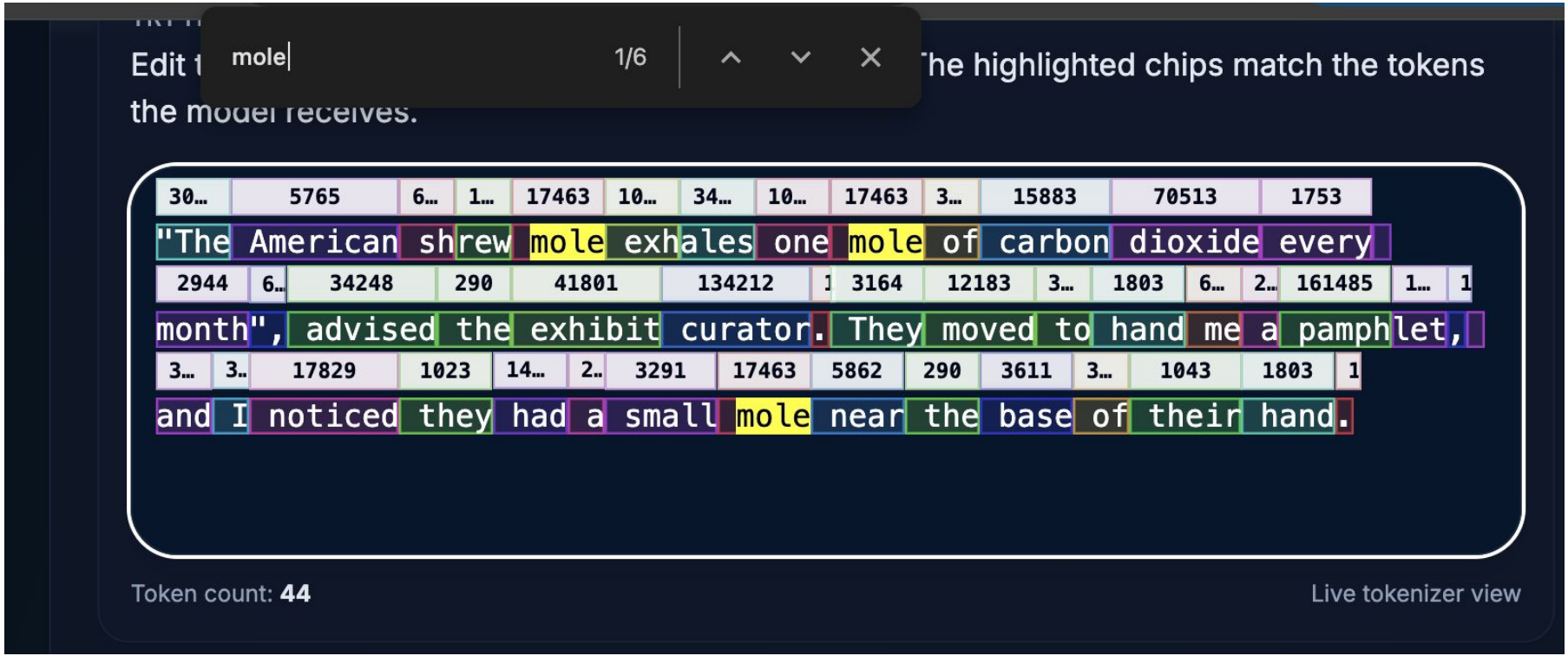
- **The strategic value lies in applications, not models.**
  ~~Transformers are best understood as general-purpose reasoning engines.~~ Competitive advantage comes from how firms integrate them into workflows—routing decisions, drafting responses, summarizing knowledge, or powering agentic systems—not from building the largest model.

Transformers did not invent word embeddings



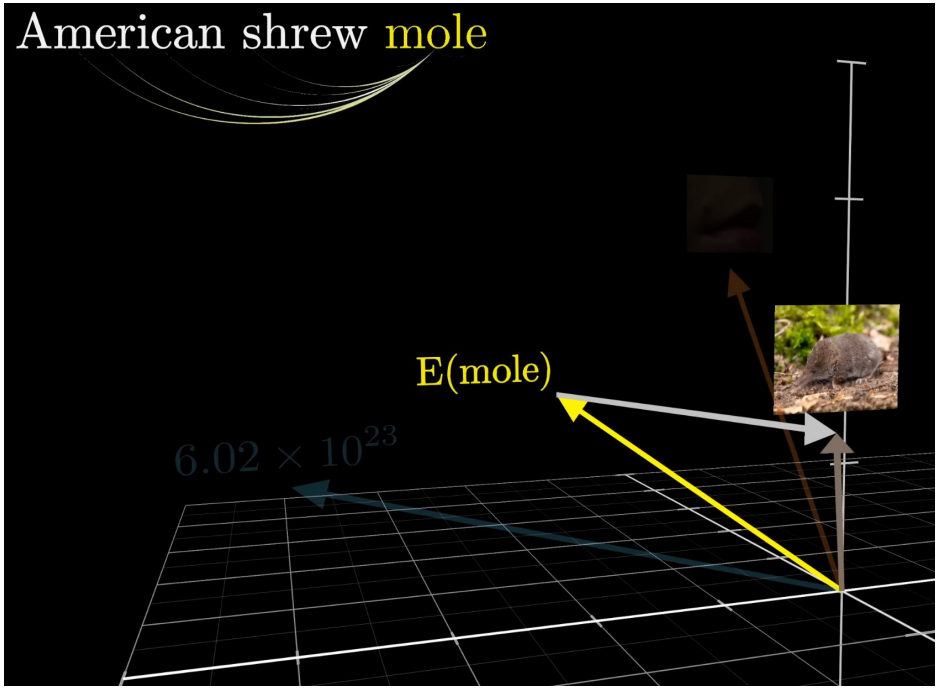
From https://huggingface.co/spaces/hesamation/primer-llm-embedding

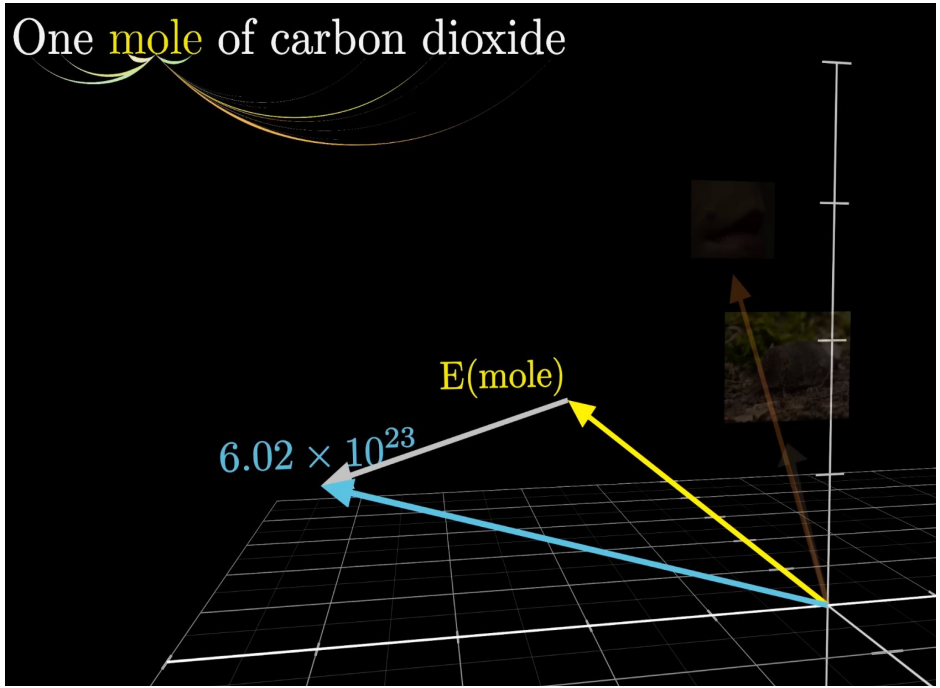
Transformers represent words as tokens



From https://gpt-tokenizer.dev/

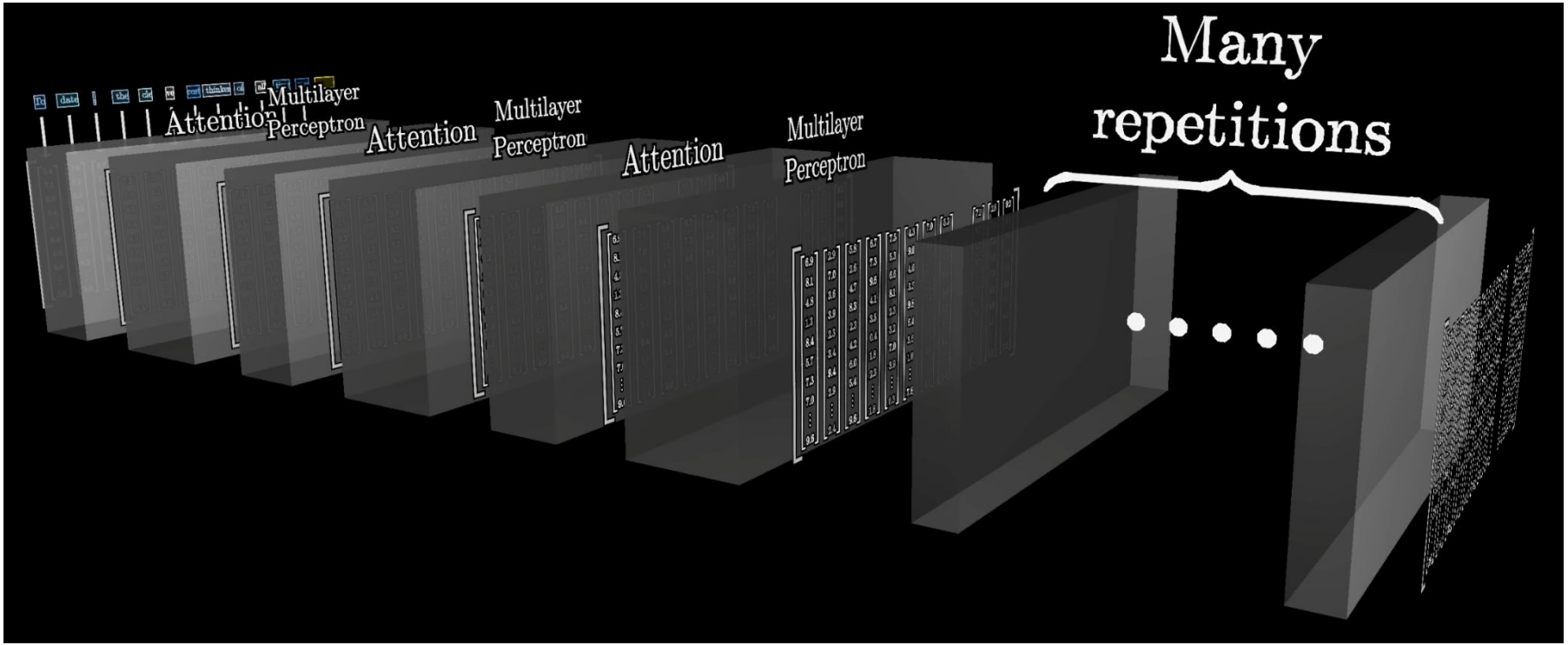
Different word senses use the same token?



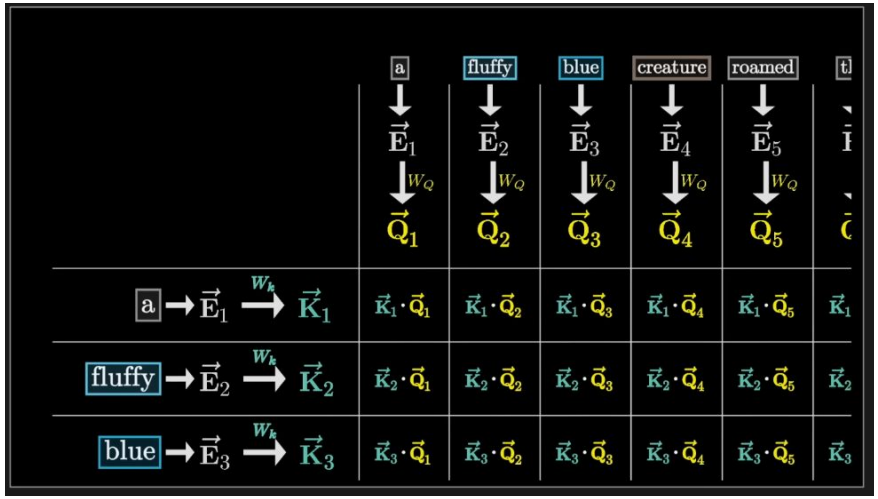


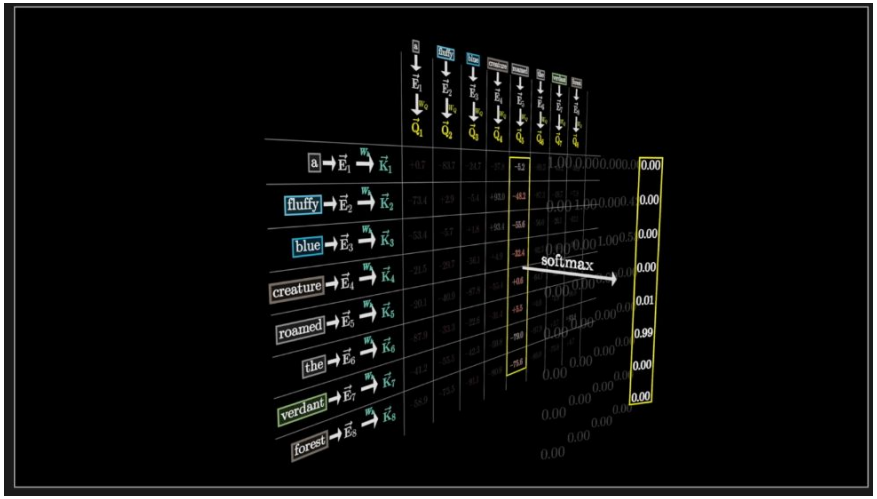
From https://3blue1brown.substack.com/p/visualizing-attention

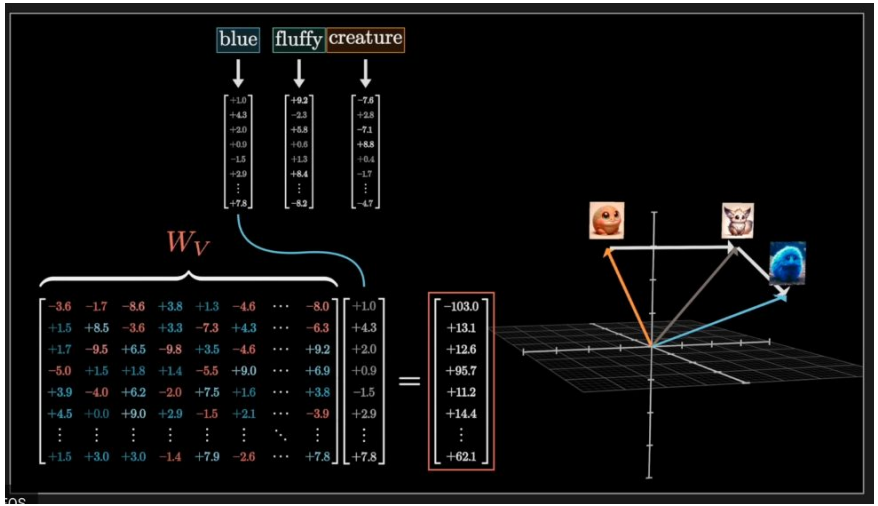
Large models have capacity to learn a lot

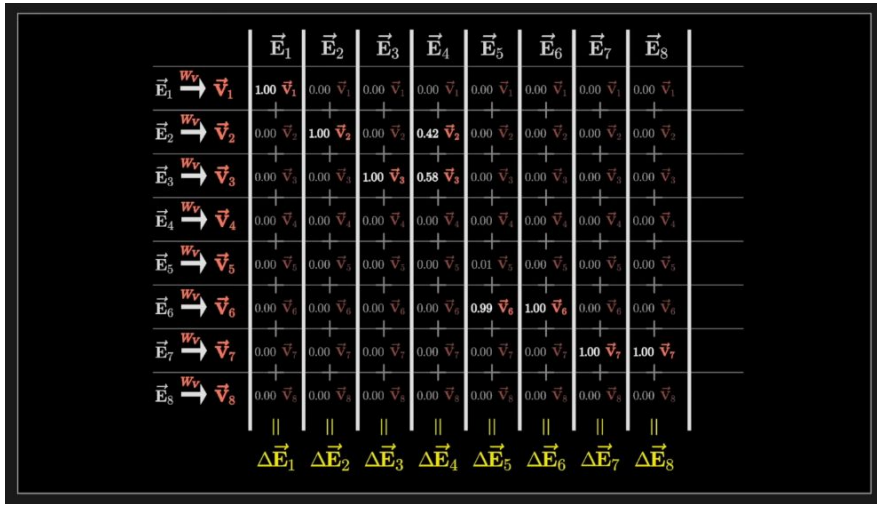


From https://3blue1brown.substack.com/p/visualizing-attention



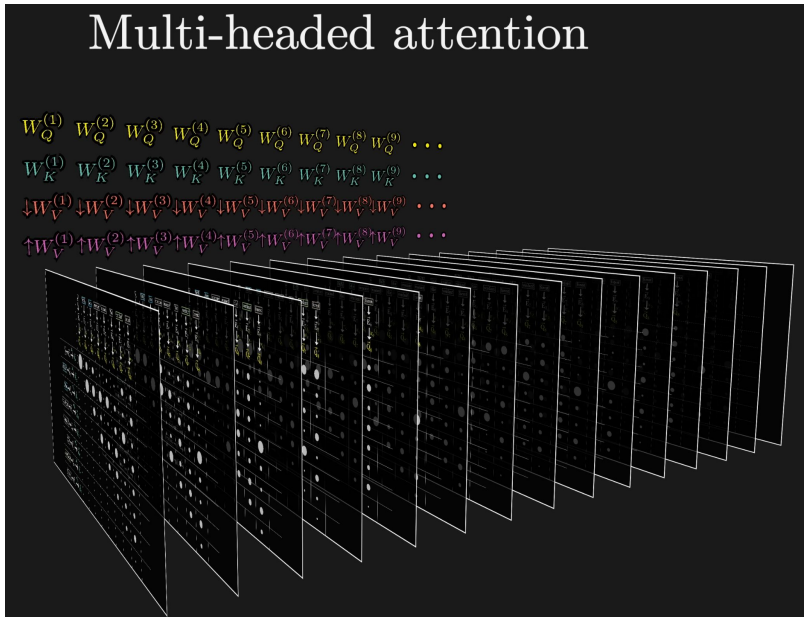


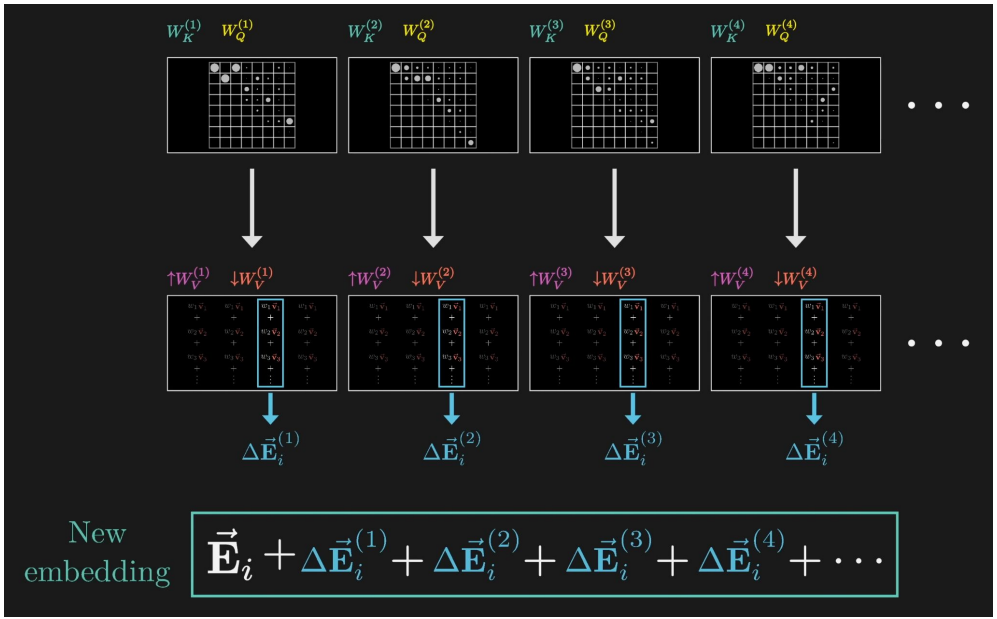




From https://3blue1brown.substack.com/p/visualizing-attention

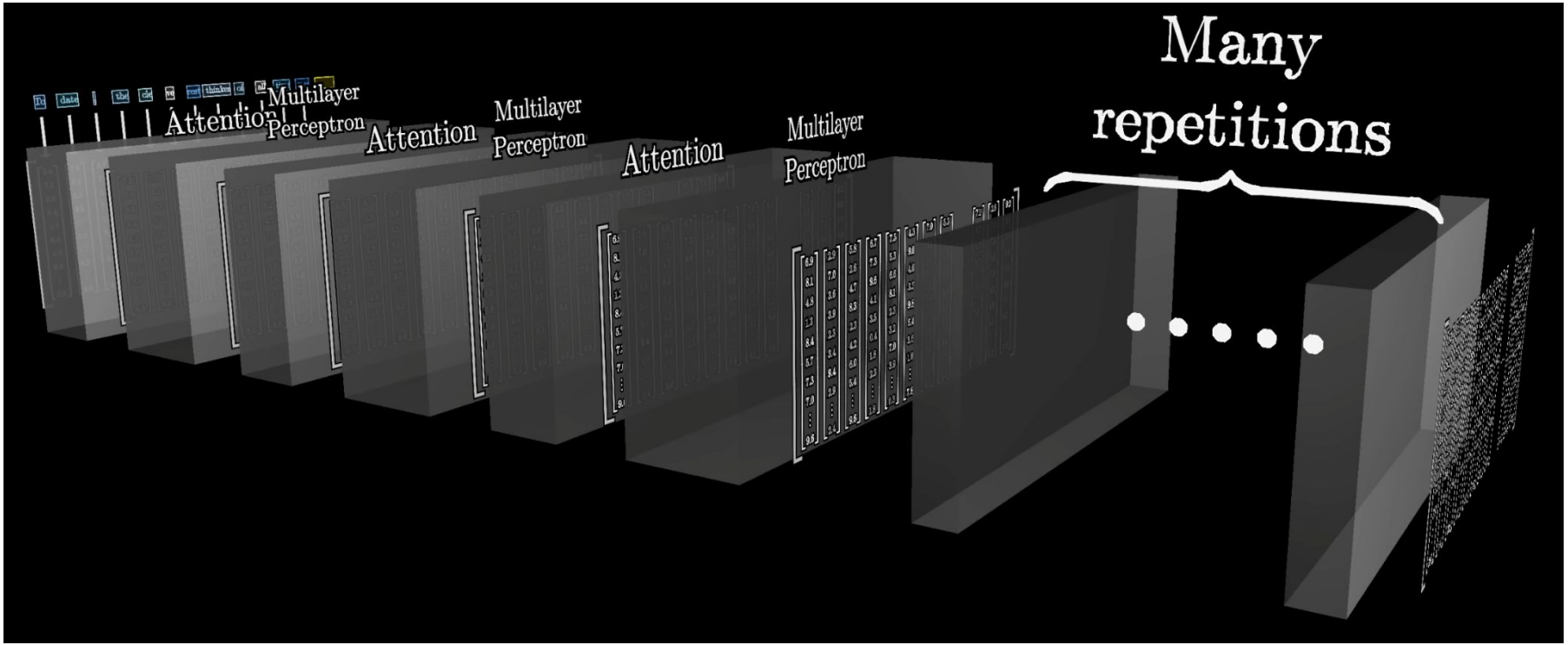
Multi-head attention learns different relations





From https://3blue1brown.substack.com/p/visualizing-attention

Attention is not all you need



From https://3blue1brown.substack.com/p/visualizing-attention

Transformers, so what?

Transformers

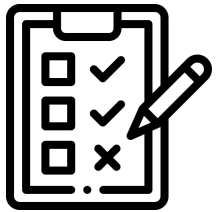
AI enables a variety of discrete capabilities

Capability	Task	Examples
Classify / Label	Assign inputs to predefined categories	Type, triage, diagnose, etc.
Extract	Identify a specific part of an unstructured input	Segmentation of image, audio, etc.
Group	Divide the input into distinct subsets	Cluster, detect anomalies, etc.
Score / Predict	Estimate a continuous numeric value	Forecast, quantify, rank, etc.
Search	Explore to find the best options for the input	Retrieve, route, match, etc.

Different types of AI work with different data

Type

- Structured (tables, defined fields)
- Text (documents, messages, descriptions)
- Images, audio, video, sensor data
- Multimodal



Quality

- Volume / size
- Representativeness
- Labels
- Accessibility

Real world constraints inform the tech choice

Constraint	**Consideration**
Explainability	Must decisions be justified to regulators, customers, or internal stakeholders?
Latency	Real-time (milliseconds)? Interactive (seconds)? Batch (minutes/hours)?
Cost	What can you spend per prediction? On infrastructure? On development?
Availability	Do you have enough examples? Can you acquire or create more?
Resources	What does your team have the expertise and capacity to build and maintain?

Deciding what AI to use when requires knowing your task and data circumstances

1. Can you specify the rules?
2. Can you identify the features?
3. Is there a pretrained model for your data type?
4. Is there data you can use for fine-tuning?
5. Do you have the resources to train a model from scratch?

Transformers enabled the training of larger, more general pretrained models

1. Can you specify the rules?
2. Can you identify the features?
3. Is there a pretrained model for your data type?
4. Is there data you can use for fine-tuning?
5. Do you have the resources to train a model from scratch?

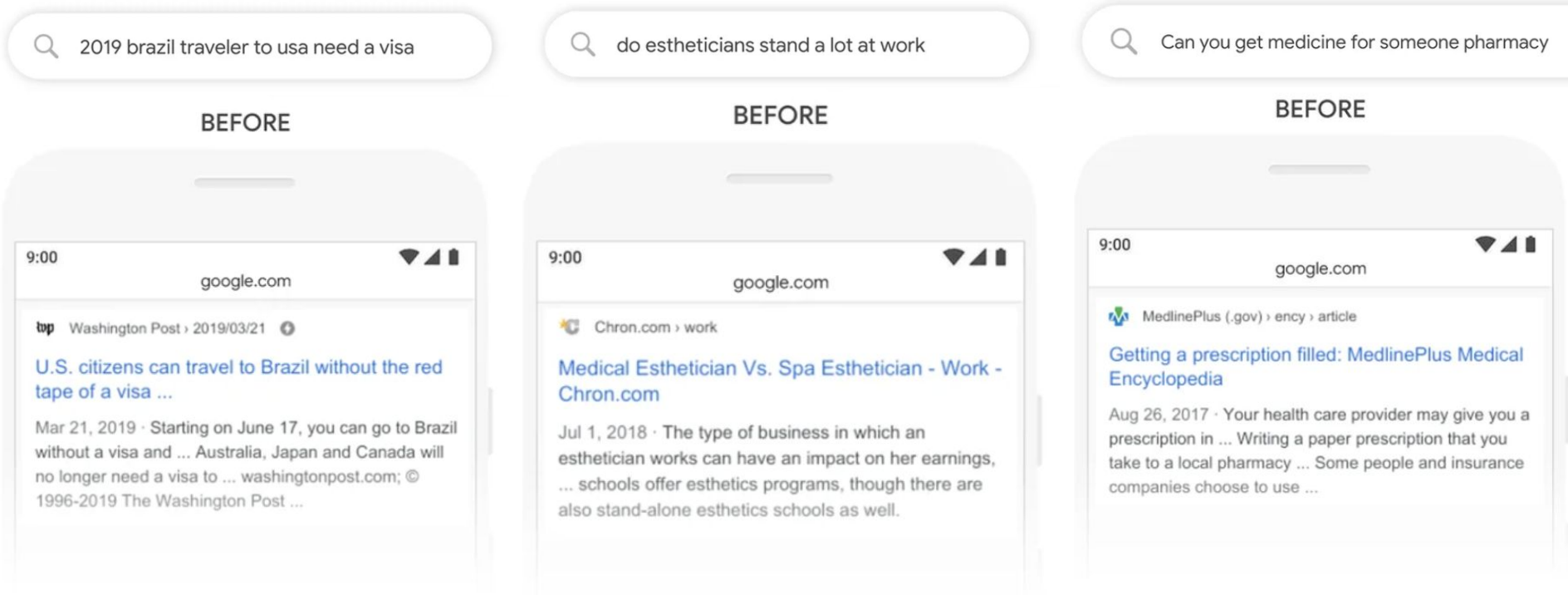
Transformers enabled the training of larger, more general pretrained models

1. Can you specify the rules?
2. Can you identify the features?
3. Is there a pretrained model for your data type?
4. Is there data you can use for fine-tuning?
5. Do you have the resources to train a model from scratch?

Transformers use cases

Transformers

Case Study: Google Search



From https://blog.google/products-and-platforms/products/search/search-language-understanding-bert/

Case Study: Google Search

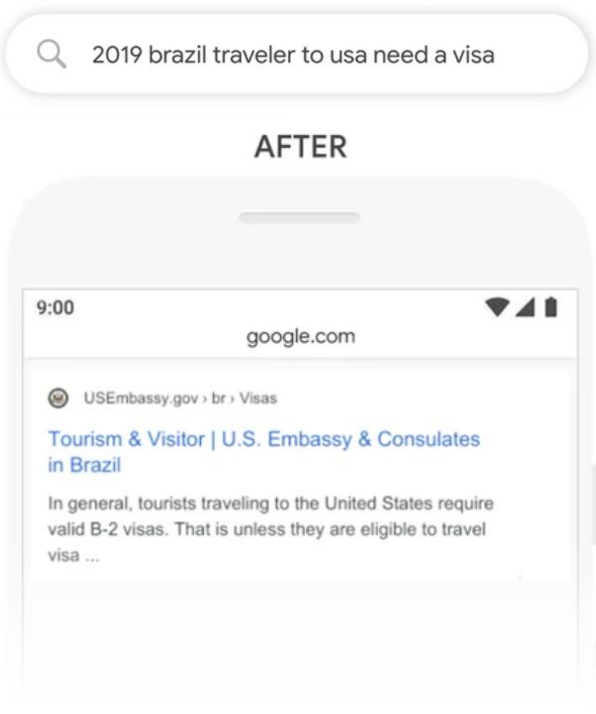
Cracking your queries

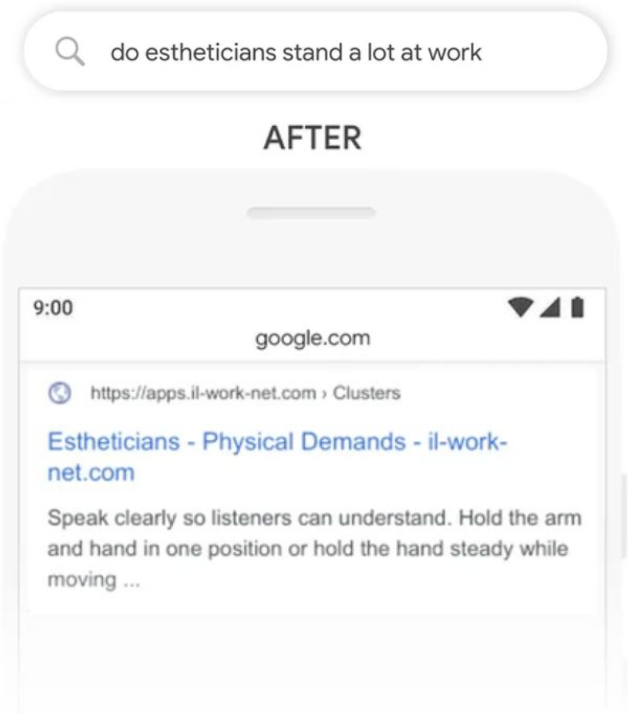
So that's a lot of technical details, but what does it all mean for you? Well, by applying BERT models to both ranking and featured snippets in Search, we're able to do a much better job helping you find useful information. In fact, when it comes to ranking results, BERT will help Search better understand one in 10 searches in the U.S. in English, and we'll bring this to more languages and locales over time.

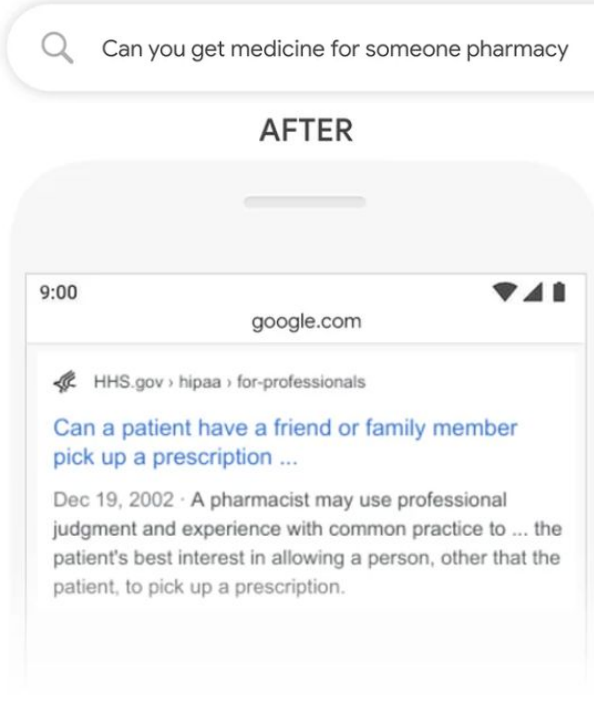
Particularly for longer, more conversational queries, or searches where prepositions like "for" and "to" matter a lot to the meaning, Search will be able to understand the context of the words in your query. You can search in a way that feels natural for you.

From https://blog.google/products-and-platforms/products/search/search-language-understanding-bert/

Case Study: Google Search







From https://blog.google/products-and-platforms/products/search/search-language-understanding-bert/

Case Study: AirBnb

Inspiration for your next trip

South Lake Tahoe

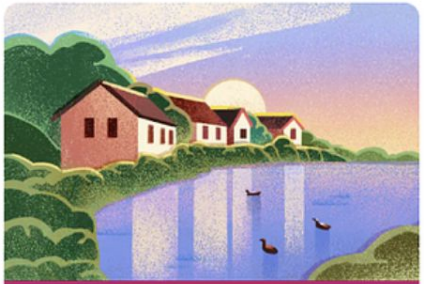
160 miles away

Paso Robles

154 miles away

Arnold

117 miles away

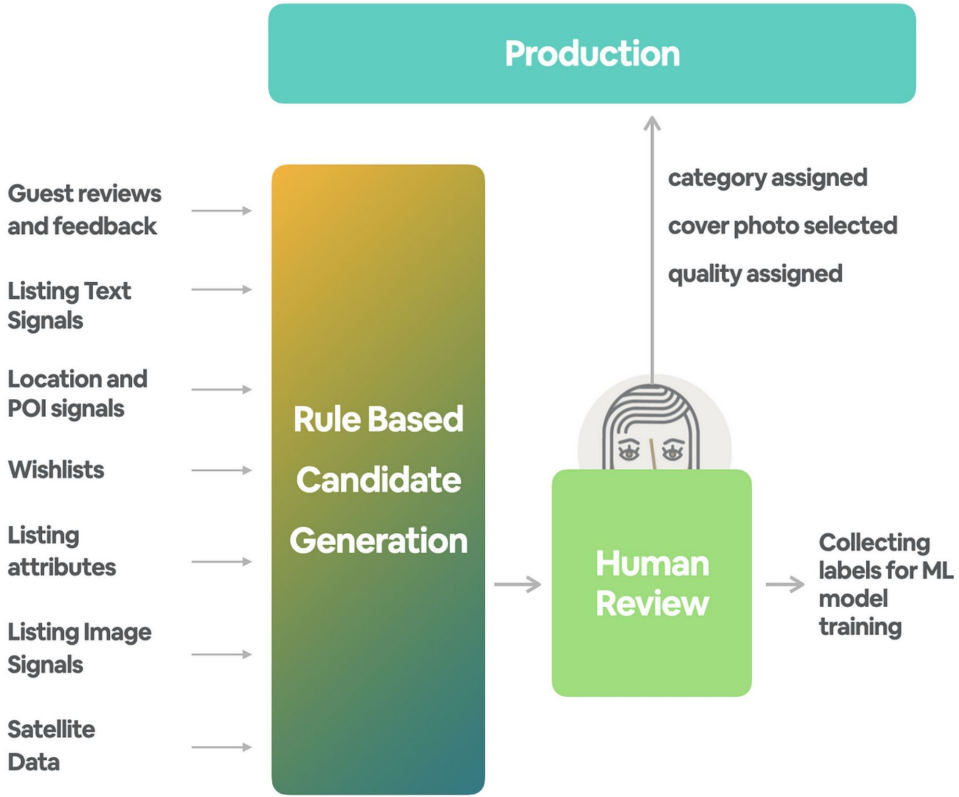


Carmel-by-the-Sea

65 miles away

From https://airbnb.tech/ai-ml/building-airbnb-categories-with-ml-and-human-in-the-loop/

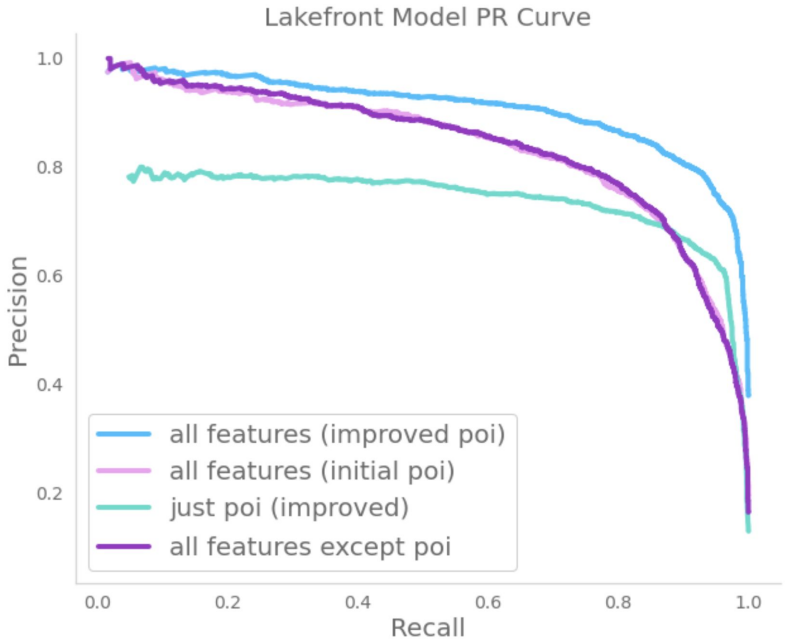
Case Study: AirBnb

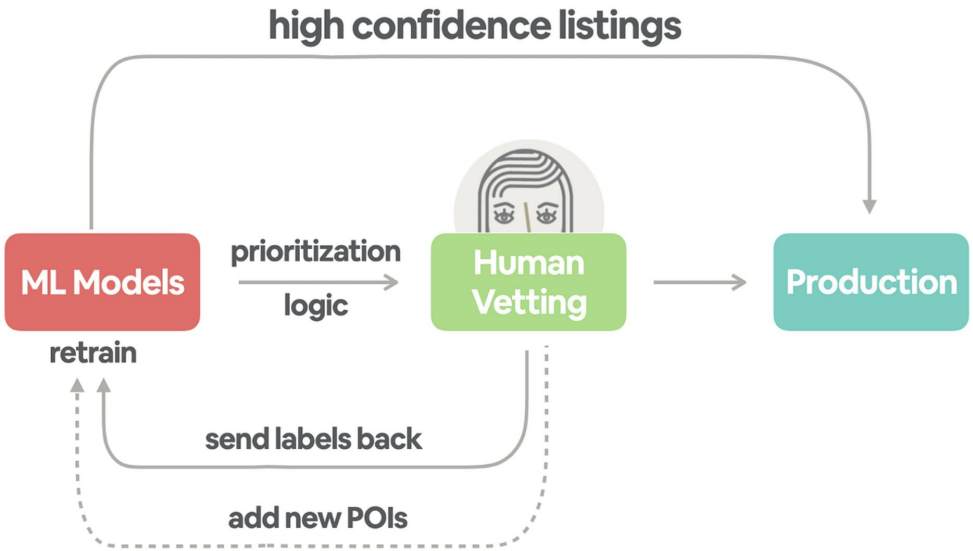


"With the Summer launch just a few months away, we knew that we could not manually curate all the categories, as it would be very time consuming and costly. We also knew that we could not generate all the categories in a rule-based manner, as this approach would not be accurate enough. Finally, we knew we could not produce an accurate ML categorization model without a training set of human-generated labels. Given all of these limitations, we decided to combine the accuracy of human review with the scale of ML models to create a human-in-the-loop system for listing categorization and display."

From https://airbnb.tech/ai-ml/building-airbnb-categories-with-ml-and-human-in-the-loop/

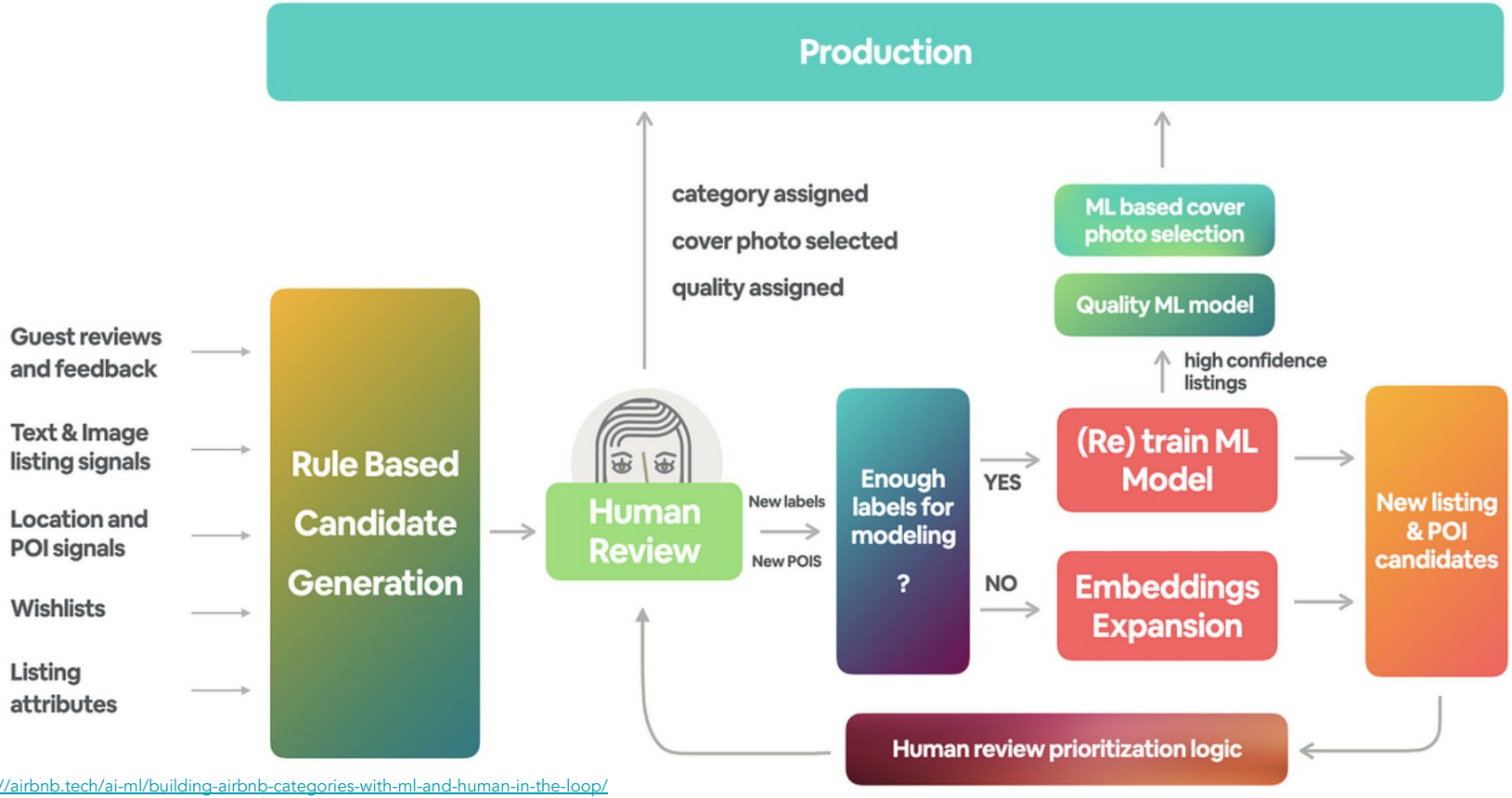
Case Study: AirBnb





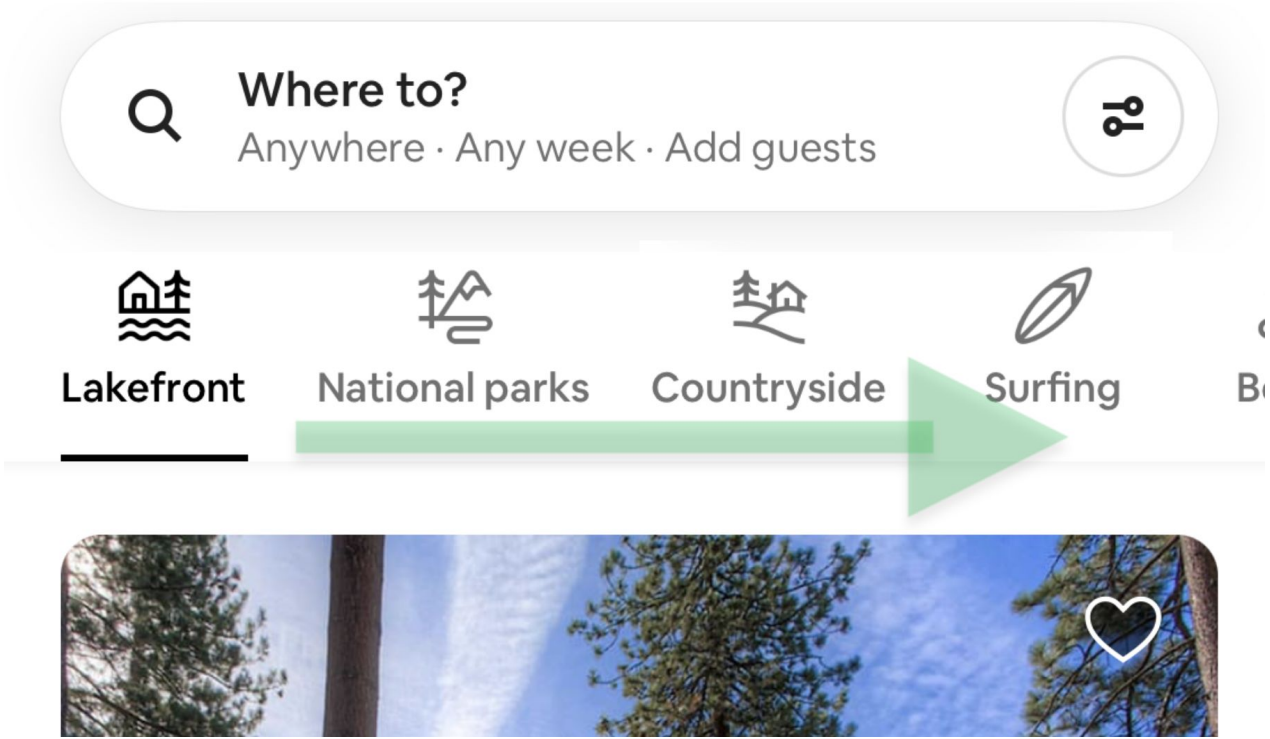
From https://airbnb.tech/ai-ml/building-airbnb-categories-with-ml-and-human-in-the-loop/

Case Study: AirBnb



From https://airbnb.tech/ai-ml/building-airbnb-categories-with-ml-and-human-in-the-loop/

Case Study: AirBnb



From https://airbnb.tech/ai-ml/building-airbnb-categories-with-ml-and-human-in-the-loop/

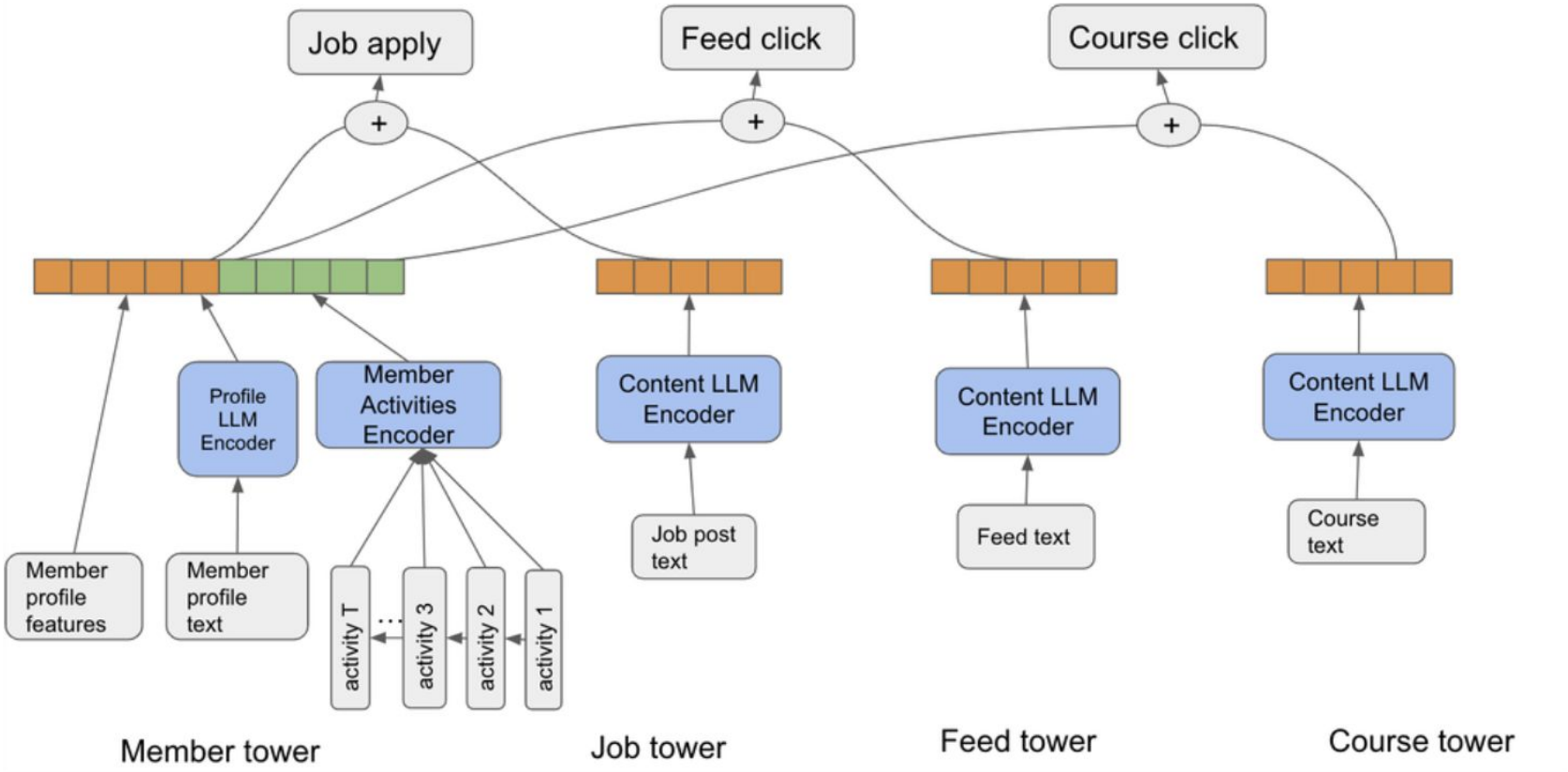
Case Study: LinkedIn

Problem

- LinkedIn wants to recommend Jobs You Might Be Interested In (JYMBII), but discerning which these are is a challenge

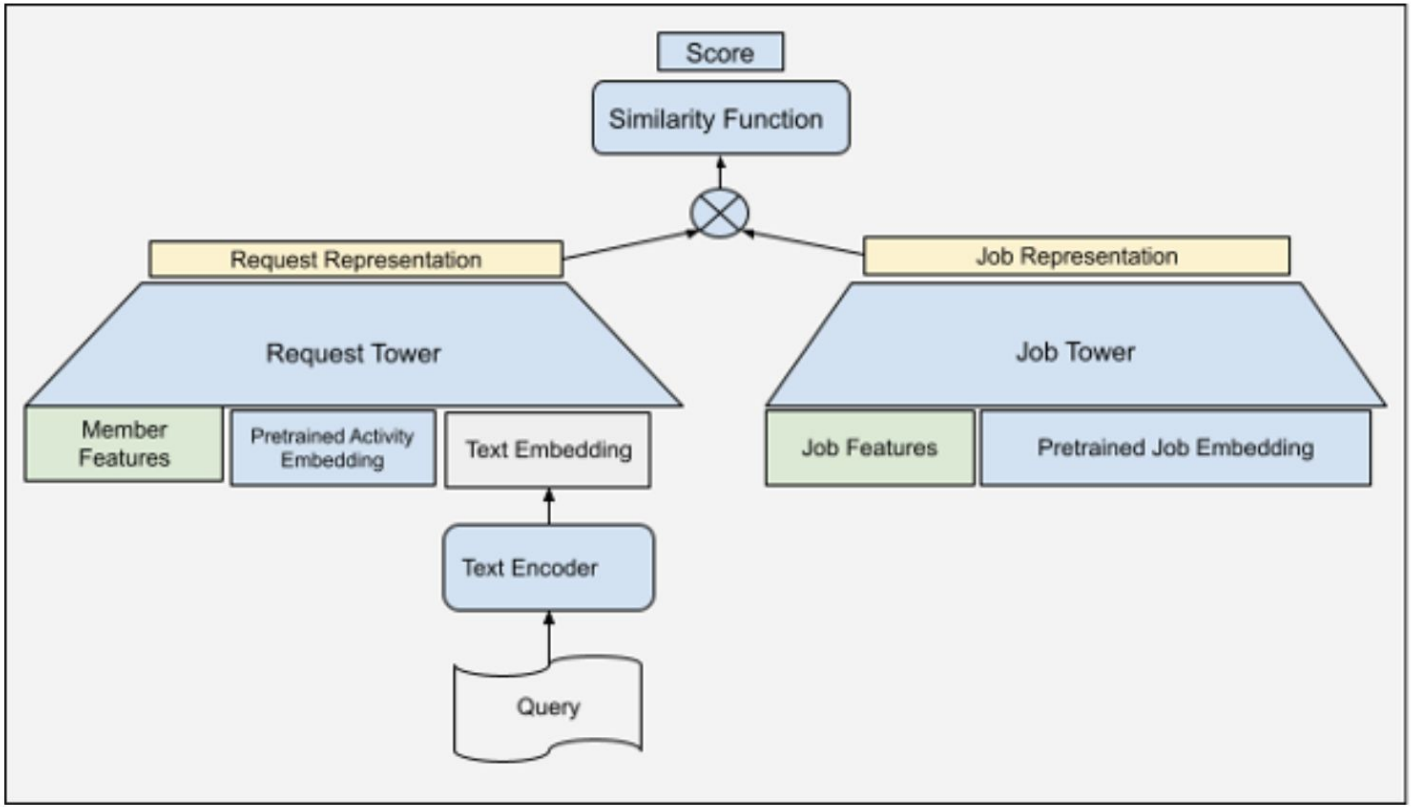
From https://www.linkedin.com/blog/engineering/platform-platformization/using-embeddings-to-up-its-match-game-for-job-seekers

Case Study: LinkedIn



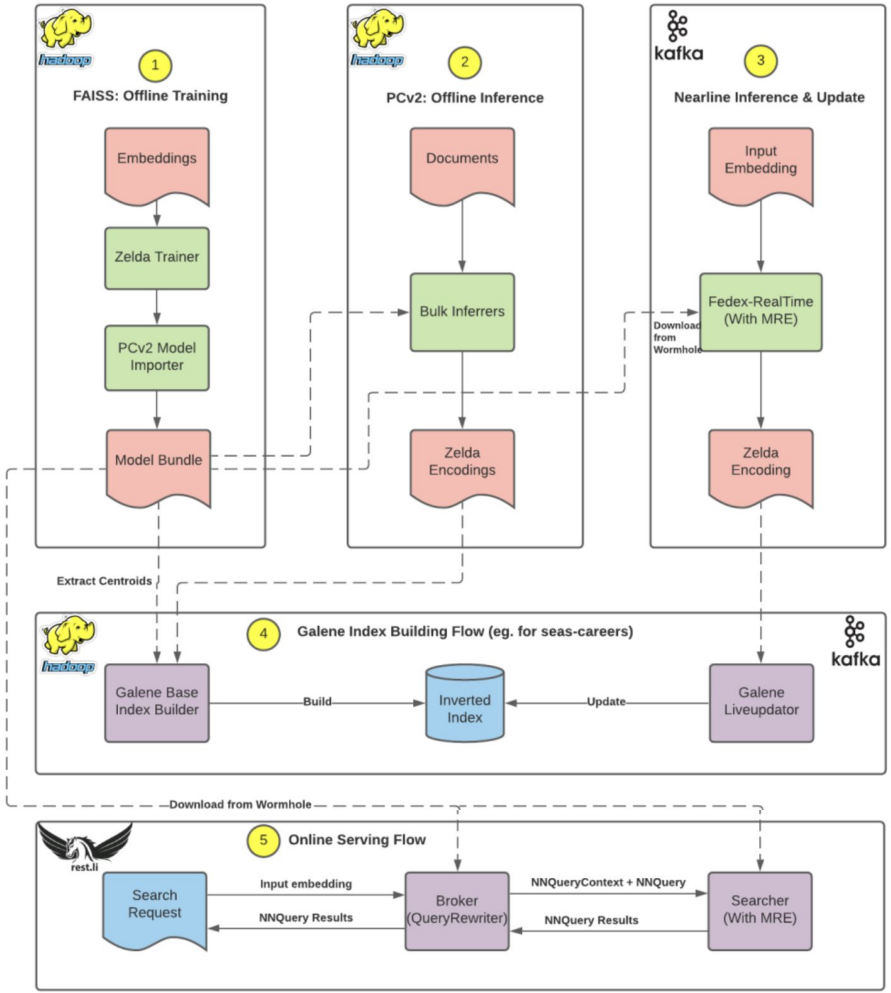
From https://www.linkedin.com/blog/engineering/platform-platformization/using-embeddings-to-up-its-match-game-for-job-seekers

Case Study: LinkedIn



From https://www.linkedin.com/blog/engineering/platform-platformization/using-embeddings-to-up-its-match-game-for-job-seekers

Case Study: LinkedIn



From https://www.linkedin.com/blog/engineering/platform-platformization/using-embeddings-to-up-its-match-game-for-job-seekers

What is the right AI to use?

Transformers

Loan origination

Grant review

Ask an AI model

Transformers

Getting AI to think with you, not for you

Stay mindful of bias

Question solutions

Expand your perspectives

Spot weaknesses or vulnerabilities

Start with your own ideas

https://www.mccormick.northwestern.edu/magazine/fall-2025/getting-ai-to-think-with-you-not-for-you/

Assignment walkthrough

Transformers