AI PROGRESS

AI MODEL TRAINING: SEPARATING FACT FROM FICTION

MYTHS VS. FACTS

MYTH: Training AI models on copyrighted content is illegal.	**FACT:** U.S. copyright law allows for transformative uses under the fair use doctrine. Two U.S. courts have found that generative AI training is transformative and a fair use.
MYTH: AI models copy and store copyrighted content.	**FACT:** AI models learn from patterns and statistical relationships in data. They do not need to store or retrieve original data to create new content.
MYTH: Using complete, unchanged works is never fair use, no matter the application.	**FACT:** Courts have long held that using full, original works can be fair use if it's for a new purpose and doesn't replace or compete with the originals.
MYTH: AI-generated responses are just regurgitated words and data.	**FACT:** AI models generate new content. Memorization of training data is rare and may be due to imprecise training practices.
MYTH: If an AI output seems similar to original works, the AI model must be copying from that material.	**FACT:** Seemingly similar responses can come from common language patterns or discussions of similar ideas and facts — not direct copying.

IS USING COPYRIGHTED MATERIALS LEGAL? YES — WHEN IT QUALIFIES AS FAIR USE

U.S. copyright law — particularly the fair use doctrine — provides clear protection for innovators and creators who use existing works in **transformative ways**, meaning the new uses differ from the original works' purposes and do not substitute for them in the marketplace.

For nearly 185 years, U.S. courts have applied fair use, preventing copyright from obstructing the advancement of new technologies — for everything from the VCR to anti-plagiarism software to internet search engines. **AI is no different.**

✓ VCR

✓ SEARCH ENGINES

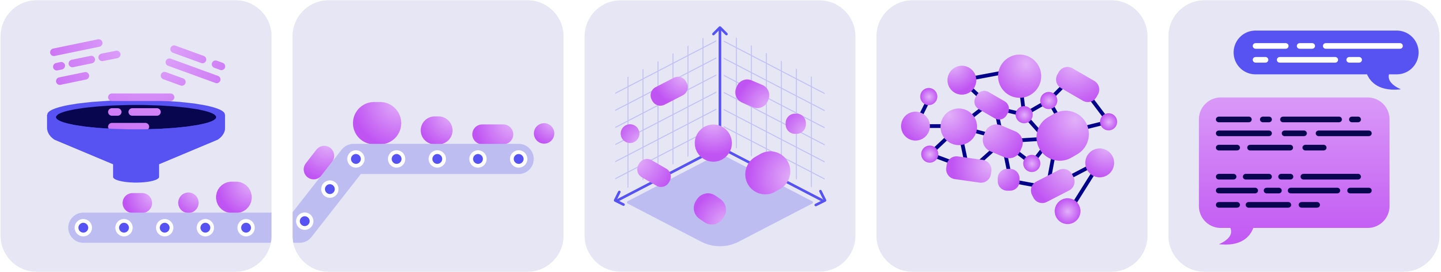
✓ DIGITAL LIBRARIES

✓ ANTI-PLAGIARISM TOOLS

HOW AI MODELS LEARN

Generative AI models — particularly large language models (LLMs) — do not need to memorize or store the datasets they're trained on. Instead, models break down the vast amount of content they are processing into smaller parts called tokens, which are small units like words or characters. From there, each token is assigned a numerical representation, called an "embedding," that captures its meaning and statistical relationship across the training dataset.

Training AI models on broad, diverse datasets is essential for building reliable, accurate systems that deliver real-world benefits.



RAW TEXT is filtered and processed → **TOKENS** are created to represent words or characters → **EMBEDDINGS** are created by assigning a number to each token → **MODEL TRAINING** interprets relationships between embeddings → **RESPONSE** generates answers from each interpretation

HOW AI MODELS GENERATE RESPONSES

LLMs transform words and their relationships into vectors — points in a multi-dimensional space that form a kind of map the model uses to understand and generate language. This mathematical map doesn't store the original content — instead, it focuses on the relationships between words, which is how it can generate new sentences.

For example, if a user asks, "**Do queens live in a castle?**", the model references the vector representations of key words like "queen" and "castle" to identify their contextual relationships.

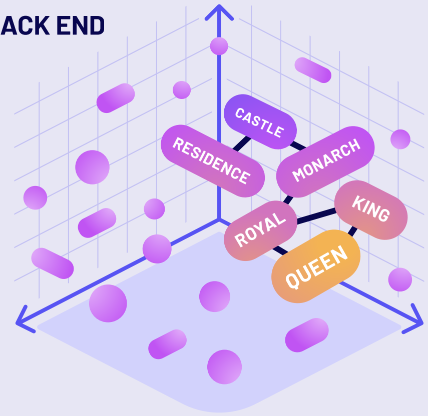
Because "**queen**" is mapped closely to words like "**king**" and "**royal**" in its training data, and "**castle**" is strongly associated with "**monarchs**" and "**residences**," the model recognizes that these concepts frequently occur together. Using this learned pattern, the model predicts the most probable response: "**Yes, as royals, kings and queens often live in castles.**"

FRONT END

Do queens live in a castle?

Yes, as royals, kings and queens often live in castles.

BACK END



WHY IT MATTERS

Training AI models on publicly available data fuels innovation and holds the promise to solve society's biggest challenges. Restricting that access would limit progress across key areas and put the U.S. at a competitive disadvantage.

BENEFITS OF AI

- ✓ Enables faster breakthroughs in health care and life sciences
- ✓ Supports national defense, cybersecurity, and threat detection
- ✓ Powers tools that assist educators, researchers, and scientists
- ✓ Encourages entrepreneurship and boosts economic growth