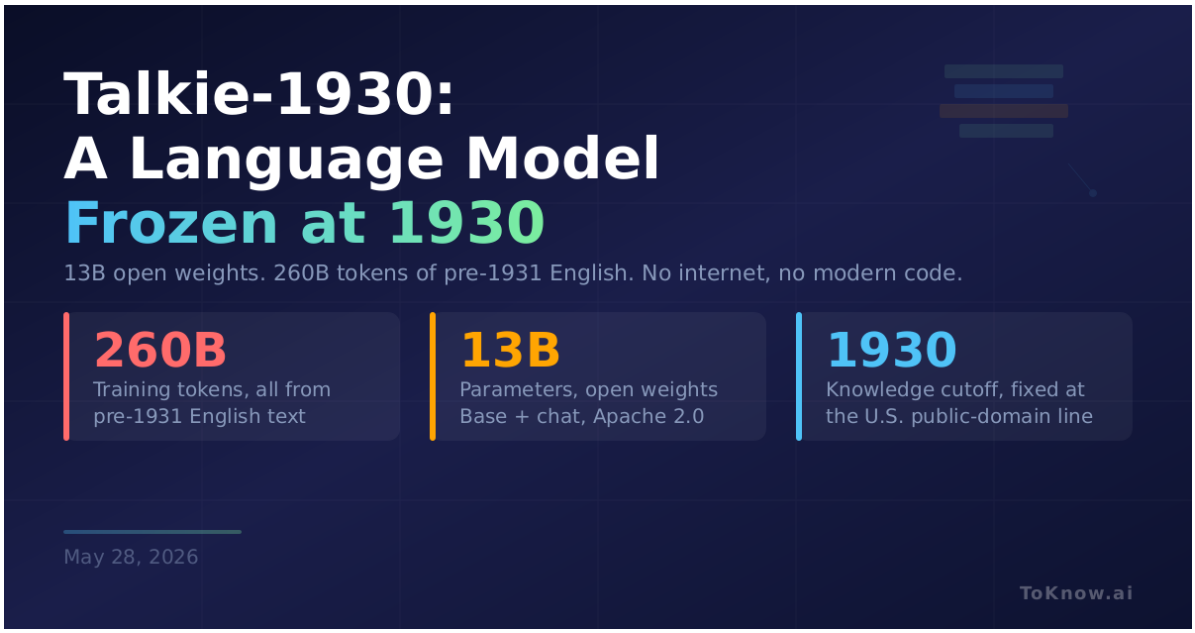
Talkie-1930: A 13B Open Model Trained Only on Text From Before 1931

Kabui, Charles

2026-05-28

Read at ToKnow.ai



Researchers Nick Levine, David Duvenaud, and Alec Radford released talkie-1930-13b, a 13-billion-parameter open-weight language model trained on 260 billion tokens of English text published before 1931. They chose December 31, 1930 because anything published before that date is in the U.S. public domain, which makes the entire corpus legally usable. The data was assembled from books, newspapers, scientific journals, patents, and case law digitized by the

Internet Archive, the Institutional Data Initiative, and Common Pile. Both a base and a chat checkpoint ship under Apache 2.0. To build the chat version without modern data leaking in, the team mined instruction-response pairs from historical etiquette manuals, letter-writing guides, and cookbooks, then used Claude Sonnet 4.6 as a judge during preference training. Average instruction-following rose from 2.0 to 3.4 on a five-point scale.

Every modern frontier model shares the same ancestor: a snapshot of the open web, which makes it hard to separate genuine reasoning from memorized facts. Talkie strips that ancestor away. When the team gave it Python problems on HumanEval, the model had never seen a digital computer, yet it solved simple ones from a handful of example programs supplied in the prompt, including correctly inverting a Caesar cipher. At least some "capability" clearly survives the loss of internet-scale training data.

Benchmark contamination is one of the most persistent problems in LLM evaluation: test items leak into training corpora and inflate scores. A model with a hard 1930 knowledge cutoff is contamination-free against every modern benchmark by construction, which makes Talkie a cleaner instrument for measuring generalization than any model trained on today's web.

Read More: DataFlex shows how dynamically reweighting training data changes what a model learns.

Sources:

- Introducing talkie: a 13B vintage language model from 1930 (talkie-lm.com)
- Meet Talkie-1930 (MarkTechPost)
- talkie-1930-13b-base on Hugging Face
- Hacker News discussion (773 points)
- Talkie is an AI language model trained only on pre-1931 texts (Boing Boing)
