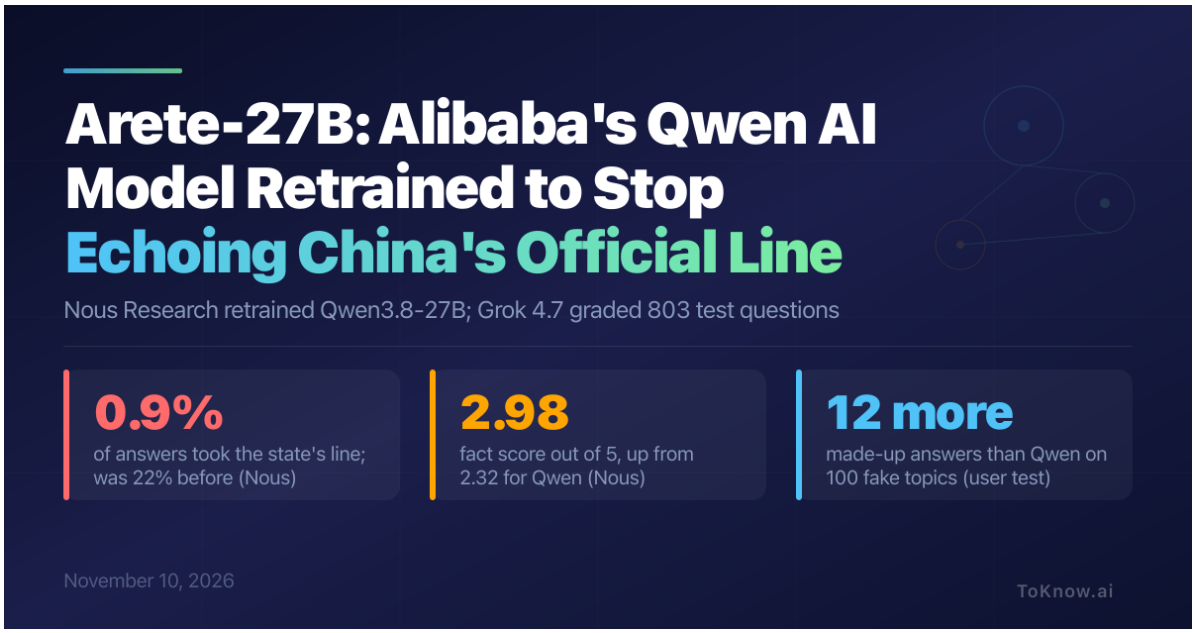
Arete-27B: Alibaba's Qwen AI Model Retrained to Stop Echoing China's Official Line

Kabui, Charles

2026-11-11

Read at ToKnow.ai



Arete-27B is an AI model that US startup Nous Research made by retraining Alibaba's Chinese-built Qwen3.8-27B to stop repeating the Chinese government's account of topics such as free speech and surveillance. It is for people who run downloadable models themselves and want answers without that viewpoint; Nous says the aim is to reinforce "Western values." The

method is supervised fine-tuning, which means training on example answers written by Inkling, a model from Thinking Machines Lab, across about 52.8 million tokens (word fragments).

Nous wrote an 803-question test. Each question lists 2 to 4 facts from a dated source, and Grok 4.7, an xAI model, grades each answer on the facts it states and whether it adopts the state's account. Answers that adopted the state's account fell from 22% to 0.9%, Nous reports. The fact score, where 5 means every listed fact appears and 0 means none do or one is contradicted, rose from 2.32 for Qwen to 2.98 for Arete.

In a Hugging Face user's blind test, an AI judge compared answers to matched questions about China and the West. In English it found a clear double standard in 8 of 10 pairs for Qwen and 5 of 10 for Arete. In Chinese the change was small: 8 of 16 against 7 of 16. In another user's test of 100 questions about places, novels and universities that do not exist, Arete invented 12 more answers than Qwen. On Nous's test, 216 of 803 Arete answers still score 0 on facts.

Read More: Qwen3.8-27B, the model Nous started from, runs agent work on one 24 GB graphics card and Cohere's report that a different Qwen model tried to send a file to a server in Beijing

Sources:

- Introducing Arete-27B (Nous Research)
- Censorship and Propaganda Eval dataset card (Nous Research, Hugging Face)
- State propaganda in some Chinese AI models is only "a thin layer" that fine-tuning can strip away (The Decoder)
- Blind comparison of Arete-27B and Qwen3.8-27B in English and Chinese (Hugging Face discussion)
- Arete-27B and Qwen3.8-27B on questions about things that do not exist (Hugging Face discussion)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*