Hirundo Unlearns Bias from DeepSeek-R1 and Releases Debiased Model Publicly

Hirundo reduces bias in DeepSeek-R1 by 76% without performance loss, releasing the fairer model on Hugging Face to promote responsible AI development.

DeepSeek-R1 Debiased by Hirundo

TEL AVIV, ISRAEL, January 31, 2025 /EINPresswire.com/ -- Hirundo has successfully reduced the bias of DeepSeek-R1-Distill-Llama-8B by up to 76% using its advanced bias unlearning technology. While DeepSeek's models has fueled adoption, Hirundo's evaluations showed that their Llama-Distilled model exhibits nearly twice the bias of the original Llama model. By unlearning this bias, Hirundo ensures businesses can deploy models that perform well while mitigating potential risks when facing customers. The debiased model is now publicly available on Hugging Face.

> Bias in AI is not just a technical issue—it's a strategic risk for enterprises. Our work on DeepSeek-R1-Distill-Llama-8B shows that bias unlearning is possible, balancing performance and fairness"
>
> *Ben Luria, CEO at Hirundo*

DeepSeek's Rising Prominence—and Its Bias Challenges

DeepSeek-R1-Distill-Llama-8B is an open-source LLM with growing adoption due to its efficiency and high performance. However, Hirundo's evaluations found that it carries almost double the bias compared to its foundational Llama 3.1 8B version.

Bias evaluation using the Bias Benchmark for Question Answering (BBQ) dataset revealed substantial disparities. Race bias increased from 17% in Llama 3.1 8B to 32.5% in DeepSeek-R1-Distill-Llama-8B. Nationality bias rose from 29% to 50.3%, while gender bias increased from 25.6% to 39.3%.

Hirundo recognized these challenges and applied its cutting-edge bias unlearning technology to mitigate the biases, ensuring the model retained its accuracy and utility.

The Challenge of Bias in LLMs

Bias in AI models is not just an ethical issue; it's also a regulatory and business imperative. With the AI Act now in effect in the European Union and increasing scrutiny from US agencies like the Federal Trade Commission (FTC), businesses face rising pressure to ensure that their AI systems are fair and compliant.

For businesses deploying AI in customer-facing roles, biased models can lead to legal and reputational risks, affecting customer trust and satisfaction. Unfortunately, traditional bias mitigation methods often involve costly retraining processes that are impractical for large-scale AI systems.

Hirundo's Breakthrough: Bias Unlearning at Scale

Hirundo's proprietary bias unlearning technology, soon available on its platform, offers a scalable, efficient alternative to traditional retraining. Unlike brute-force methods that require filtering and retraining from scratch, Hirundo's approach is:

- Fast & Efficient: Removes bias in any open-source LLM in under an hour with moderate computing power.
- Preserves Performance: Ensures model accuracy and utility remain intact.
- Scalable: Works across various AI models and applications, including both pre-trained and fine-tuned models.
- Beyond Bias: In addition to bias unlearning, Hirundo also offers privacy and knowledge unlearning capabilities, allowing AI to forget sensitive or outdated information.

Bias Reduction Results: A 76% Improvement

Applying Hirundo's bias unlearning technology to DeepSeek-R1-Distill-Llama-8B resulted in significant bias reduction while preserving model performance.

- Race bias dropped from 32.5% to 7.8% (76% reduction).
- Nationality bias decreased from 50.3% to 15.3% (69.5% improvement).
- Gender bias fell from 39.3% to 13.2% (66.3% reduction).

Model performance remained largely unchanged, with the TruthfulQA perplexity score only shifting slightly from 9.8 to 9.9, and the LogiQA2.0 accuracy score remaining virtually the same at 42.5% and 42.6%.

The debiased model is now publicly available on Hugging Face, alongside the full case study on Hirundo's website.

Ben Luria, CEO of Hirundo, on the Criticality of AI Fairness in Enterprises Using LLMs

"Bias in AI is not just a technical issue—it's a trust and compliance necessity," said Ben Luria, CEO

of Hirundo. "For businesses, biased models can negatively impact customer interactions and regulatory standing. We believe that AI should be built with fairness from the outset, and biases must be actively addressed. Our work with DeepSeek-R1-Distill-Llama-8B shows that bias unlearning is both possible and scalable. By releasing the debiased model, we are taking a significant step toward fostering a more responsible AI ecosystem."

A New Standard for Responsible AI

Hirundo is committed to shaping a future where AI is both powerful and fair. By making the Bias-Unlearned DeepSeek-R1 model available on Hugging Face, we invite researchers, developers, and businesses to collaborate in advancing responsible AI. Organizations interested in applying Hirundo's unlearning and AI optimization technology to their own AI models can request early access to the platform at Hirundo's website. Enterprise customers will receive dedicated support for bias and knowledge unlearning, data optimization and integration planning.

About Hirundo

Hirundo is the first startup in the world to offer an unlearning solution, pioneering the concept of "making AI forget." Our technology enables AI models to remove unwanted data or behaviors they have previously learned, ensuring that biases, sensitive information, or inaccuracies can be effectively remediated. By leveraging patent-pending technologies, Hirundo is dedicated to making AI more safe, trustworthy and accurate. Our solutions empower enterprises to deploy AI with confidence, ensuring responsible AI adoption while providing mission-critical accuracy.

For more information, visit Hirundo's website or contact press@hirundo.io.

Ben Luria
Hirundo
ben@hirundo.io
Visit us on social media:
LinkedIn