Primer

Maintain accuracy while managing AI cost

Deploy deep learning models cost-efficiently at enterprise scale

Retain accuracy, reduce compute costs

Why it matters: Deploying deep learning models at scale can incur significant costs, especially when running on-premises. Primer's BabyBear algorithmic optimization framework dramatically reduces expenses, selecting the best AI model for the job at the lowest cost.

The core strategy is inference triage, which selects the least expensive model with a high-confidence prediction. This results in significant reductions in compute costs while retaining accuracy.

Every NLP practitioner has created keyword filters upstream of expensive models to save unnecessary processing. Primer's BabyBear optimizes that human data science work with machine learning.

Under the hood: Within our algorithmic framework, the higher-cost deep learning model is called MamaBear. As documents flow in for processing by MamaBear, we introduce another model upstream: BabyBear. This model is characterized by its smaller size, faster processing, and lower cost compared to MamaBear. The MamaBear model could be a small hosted language model, or even a large language model accessed via API.

We use the confidence of the BabyBear model to determine whether an incoming data example requires MamaBear. If it's a sufficiently easy task, BabyBear handles it. If the confidence is below a set threshold, BabyBear passes it to MamaBear.

By the numbers

Primer's BabyBear framework empowers users to process large volumes of data swiftly and affordably, yielding actionable intelligence for better and faster decision-making.

Dive deeper and read our publication on BabyBear[2] or contact us to learn more at primer.ai/contact.

30-50%

reduction in GPU costs

For most deep learning NLP tasks, BabyBear slashes GPU costs by 30-50%.

90%

cost savings

In some cases, BabyBear savings exceed 90%.[1]

About Primer

Primer exists to make the world a safer place. We believe that AI is critical to accelerating intelligence and decision cycles that keep us safe. Our AI technology is deployed by the U.S. Government, strategic allies, and Fortune 100 companies to extract timely insight and decision advantage from massive datasets. Primer has offices in Arlington, Virginia and San Francisco, California.

1. Primer blog: https://primer.ai/developer/how-primer-built-an-inference-triage-process-called-babybear-to-save-gpu-time/
2. BabyBear: Cheap inference triage for expensive language models, Leila Khalili, Yao You, John Bohannon, May 24, 2022. https://arxiv.org/abs/2205.11747

Contact us to book a demo

publicsector@primer.ai