Subject: Re: Checking In & Recent AI Developments

---

Dear Luke,

Thank you for your thoughtful message. I’m glad to hear you’re settling in well at Amazon and experimenting with those new AI tools. Things here in Berkeley have been busy but rewarding—I’m currently supervising two Ph.D. students on projects in explainable AI and preparing a proposal for a workshop on AI ethics later this summer.

Your summary of recent advances is spot-on. From my perspective:

1. Generative Models & Prompt Engineering

The leaps in contextual understanding you mentioned with GPT-4 and its successors mirror what we’re seeing in the lab. We’re especially intrigued by few-shot and zero-shot capabilities, and we’re running controlled user studies to measure how reliably these models can adapt to novel domains. I’d be happy to share our preliminary findings if you’d like to compare notes.

2. Foundation Models in Industry

The evolution of Bedrock and Titan models at AWS is indeed exciting. In our group, we’ve begun benchmarking Titan against open-source models—evaluating latency, throughput, and fine-tuning cost. It might be worthwhile for you to look at our benchmarking scripts; they could help streamline your team’s evaluation process.

3. MLOps & Model Monitoring

Your point about drift detection resonates strongly. We recently integrated Google’s TFDV (TensorFlow Data Validation) with a custom alert system to flag distribution shifts in incoming data streams. Although our environment is academic, the principles should transfer directly to production services.

Looking ahead, you might consider exploring:

- Multimodal & Retrieval-Augmented Models: Integrating vision and language for richer user interactions, or tapping into retrieval-augmented generation to keep model outputs grounded in your own documentation.

- Privacy-Preserving Training: With GDPR and CCPA considerations, techniques like federated learning or differential privacy could be critical for large-scale deployments.

- Ethical & Robust AI: As systems grow more autonomous, methods for adversarial robustness and fairness auditing become ever more important—something both industry and academia are prioritizing.

Please let me know if you’d like any of our code repositories or if you’d like to schedule a deeper technical discussion. I’m also curious to learn how your prompt-engineering experiments fare in a high-throughput environment like Amazon’s.

Wishing you continued success, and I look forward to our next conversation.

Warm regards,

Associate Professor, Computer Science Division

University of California, Berkeley