Mời các bạn đặt câu hỏi. Mỗi câu 1 dòng, có khoảng cách. Mình sẽ nhìn vào đây trả lời.

> Anh cho em xin link Github hoặc Google Scholar của anh ạ
> https://scholar.google.com/citations?user=0RvAVR4AAAAJ&hl=en
> Hieu Pham

2. vậy agent cũng giống như reinforcement learning framework đúng không ạ?

3. Dạ vậy làm sao mình biết kết quả phản hồi có chính xác không ạ?
4. Dạ vậy so với single agent, multi có gì nhược điểm không ạ?

5. Dạ Feedback trong Loop cụ thể có những gì vậy ạ?

6. Vậy trong giai đoạn inference thì nó giống reinforcement learning đúng ko ạ

7. Dạ anh cho em xin slide được không ạ.

9. Các phương pháp học tăng cường đa tác nhân truyền thống (Multi-Agent Reinforcement Learning - MARL) có thể tích hợp vào MAS-LLM để cải thiện khả năng phối hợp không ạ?

10. Dạ vậy làm cách nào mình biết vòng lặp của multi-agent nên dừng, liệu rằng cơ chế này dễ làm model “overthinking”

11. ném ngược cho LLM (feedback) nghĩa là fine-tuning lại cho LLM đúng không ạ?
12. hiện tại đang có rất nhiều framework hỗ trợ build agent ví dụ như của openai agents sdk, langgraph, crew, .. Vậy thì đâu có framework hỗ trợ tốt nhất tại thời điểm hiện tại? cảm ơn anh

13. Dạ thường dự đoán nó dựa vào độ gần của vector, agent có làm thay đổi giá trị biểu diễn của tập từ không ạ?

14.Tại sao một mô hình auto-regressive có thể sinh ra code, image,... và dường như có khả năng suy nghĩ?

15. Dạ nếu model ko dc update tham số vậy thì những kiến thức update mới khi đó LLM agents sẽ ko học được đúng ko ạ ? Liệu rằng có các kĩ thuật knowledge editing có thể dùng để chỉnh ko ạ. Nếu ko mỗi lần hỏi câu đó thì model sẽ phải lại tương tác với môi trường trước khi trả lời.

16. Anh nói lại phần Auto-regressive generation và cách môi trường feedback giúp em được không ạ? :>

17.dạ Có cách nào rút ngắn feedback loop ạ
18.Dạ cách khởi tạo vector embedding cho từng từ của mạng là random hay phụ thuộc vào dictionary như thế nào ạ? Dạ em hỏi vì anh giới thiệu transformer

19. Với các end user sử dụng LLM API để xây dựng agent, hiệu suất của việc thực hiện các LLM computations có phụ thuộc vào phần cứng của end user k ạ? Hay sẽ phụ thuộc phần lớn vào sever của LLM provider ạ?

20. Vậy feedback không phải để sửa sai mà là để dự đoán next-word tốt hơn đúng k ạ?

21. Em chào anh, dạ vậy nó giống rag (bổ sung context để câu trl tốt hơn, vậy mình phải lưu cache hay sao ạ, vì đây k phải learning nên em nghĩ nó sẽ không giữ được những gì nó được cung cấp, vậy mỗi lần hỏi là feedback cũ mất ạ, vậy nếu hỏi lại cùng câu có nội dung cũ thì mình giải quyết như nào vậy ạ.

22. Với không gian từ lớn, dung full connect có ảnh hưởng hiệu suất không ạ? Và hiện tại có x tính khả thi chưa ạ?

23. Ngoài ffn thì LLM có dùng các mạng lưới khác như cnn,.. không ạ

24. đối với multi-agents system, mỗi phần LLM của agent chúng ta nên dùng từ các MLaaS (openai, grok, gemini) hay chúng ta sẽ tự fine-tuning một LLM cho riêng công ty ạ? em cảm ơn anh

25. Anh nghĩ về việc dùng kết hợp giữa Mixture of Experts(MoE) và AI agents như thế nào ạ?

26. Auto-regression xảy ra với mỗi lần agent cho ra một câu trả lời cho mình hay là cứ mỗi từ mà LLM nó generate thì việc auto-regression xảy ra với mỗi từ ạ?

27. Mô hình Mamba có được xem là tiềm năng hơn transformer không ạ?

28. Khi tạo ra một system để train AI thì làm như thế nào là khéo vậy ạ? Anh có bí quyết nào để chia sẻ không ạ?

29.tại sao đối với multi-agent system, bên phía anh Hiếu lại tự build thay vì dùng một framework nào đó ạ? sự khác nhau và giá trị mang lại là gì để khiến anh quyết định tự build multi-agent

30. What are the best practices for designing communication and coordination protocols among agents in a scalable multi-agent system?

31. thay vì sử dụng các LLM từ các provider ví dụ như openai, gemini,...thì chi phí để build một LLM không phải từ Scratch (ví dụ open-source của deepseek) thì chi phí sẽ mất khoảng bao nhiêu để thông minh gần bằng với GPT hoặc gemini ạ? em cảm ơn anh

32. Mỗi từ được chuyển về 1 vector cao chiều, ví dụ [1, 2, 3, 4,...], khi ghép nhiều từ trong một câu lại thì ta có chẳng hạn
[[1,2...],
[3,4...],
...
[9,10...]]
thì trông nó lại giống như một ma trận 2 chiều mặc dù các vector thành phần là cao chiều. Xét về mặt hình dung thì em thấy hơi cấn chỗ này, anh giải thích giúp em được ko ạ.
33. theo anh Hiếu, AI Engineer là gì? và nó có khác gì với ML engineer hay không? Và so với một software engineer thì sự khác nhau là gì ạ? Cảm ơn anh!
34.Mình không thể đưa trực tiếp ma trận vào tensor core được mà nhất thiết phải đi qua global memory hả a?

35. Dạ hiện tại các công cụ giúp tối ưu vấn đề memory bound này có được tích hợp vào package nào trong python chưa ạ hay nó độc quyền của từng công ty.

36. Kernel anh nói khác/giống gì ý tưởng của kernel trick trong svm?
2 cái đó đều gọi là "kernel" nhưng ko liên quan gì đến nhau hết em nhé. Trong computer science hay có những thuật ngữ giống nhau như thế. Một ví dụ khác là "Linux kernel", vốn ko liên quan gì đến cả 2 cái kia.

37: Đến giờ đã có chip để chỉ tính mỗi exp(x) chưa ạ?
Không có. Cái này trong bài giảng, mình đã nói đến rồi. NVIDIA chưa đủ tự tin rằng exp(x) là thật sự cần thiết, ko thể bỏ đi, nên họ chưa dám đầu tư $$$ để phát triển chip tính exp(x).

38. Nếu sau buổi này nhóm em viết paper thì tụi em cite anh như nào ạ? =)))
Ko cần phải cite anh đâu em. Tất cả những điều anh nói đều có thể tìm được trong các paper khác để em dùng làm source.

39. Anh có thể review nhanh môi trường làm việc trong xAi, làm việc cùng Elon Musk cảm giác như thế nào được không ạ, vì Elon Musk là thần tượng của em.
Mình đã trả lời câu hỏi này trong buổi nói chuyện rồi.

40. Anh được gặp trực tiếp Elon Musk chưa ạ? Nếu rồi thì ấn tượng của anh khi gặp Elon Musk là gì ạ?
Mình đã trả lời câu hỏi này trong buổi nói chuyện rồi.

41. Flash Attention chỉ tối ưu quá trình copy, nên không ảnh hưởng độ chính xác đúng không anh?

Có ảnh hưởng. Bởi vì để tối ưu quá trình copy, flash attention phải thay đổi thứ tự của các phép tính. Trong GPUs, thay đổi thứ tự phép tính ảnh hưởng đến kết quả. Ví dụ như a+b+c sẽ chỉ ra kết quả *gần giống* với a+c+b thôi.

42. Anh review buổi phỏng vấn vào xAi được không ạ? Anh có được trực tiếp được phỏng vấn bởi Elon Musk không ạ?
Cái này hơi cá nhân, mình xin phép ko trả lời.

43. Môi trường làm việc trong 1 trong những công ty AI hàng đầu thế giới có áp lực nhiều về mặt đào thải không ạ?
Cực kỳ áp lực.

44. Anh đã từng gặp và làm việc với Andreji Karapathy chưa ạ, nếu có thì trải nghiệm đó thế nào ạ
Đã từng (vẫn còn bằng chứng đây nhé). Andreij là người rất hiền lành, nhã nhặn, khiêm tốn, cũng như là biết cách cư xử, nói chuyện.

45. 2 phương pháp StreamK và Flash Attention có 2 mục tiêu khác nhau để giảm memory bound. Vậy người build LLM có thể tận dụng cả 2 (và cả các pp khác nữa) để tận dụng lợi ích của cả 2 đúng ko ạ?
Có. Không ai muốn tối ưu LLM mà không dùng **cả hai** cái đó cả. Và còn dùng nhiều cái khác nữa.

46. Anh có định về Việt Nam họp fan club offline không ạ? Cụ thể là về trường ĐHKHTN ạ?
Chưa từng nghe mình có fanclub. Chắc ko có đâu.

47. Anh nghĩ rằng sinh viên Việt Nam cần làm gì, học gì để tồn tại trong AI world ngày nay ạ?
Mình đã trả lời câu hỏi này nhiều lần rồi: phải ko ngừng suy nghĩ về giá trị của bản thân, tìm ra giá trị mà bản thân mình có thể mang lại để cống hiến cho tập thể và xã hội. Nếu suy nghĩ theo hướng đó, sẽ thấy AI chỉ là công cụ để nối dài cánh tay của mình thôi.

48. Làm sao để có thể vào xAI?
Bạn có thể email cho mình (hyhieu@gmail.com). ở xAI, các nhân viên tự kiểm tra hồ sơ ứng viên. Tất nhiên là mình chỉ hiểu biết về kernels, chips, và kiến trúc. Nếu bạn làm mảng này, mình sẽ đọc hồ sơ của bạn. Nếu bạn làm mảng khác, mình có thể gửi hồ sơ của bạn cho các bạn làm trong các mảng đó.

49. Anh có suy nghĩ ì về work-life balance khi bản thân là lập trình viên trong công ty AI hàng đầu thế giới ạ?
Chỉ có work. Gần như ko có life.

50. Anh có định viết sách phổ cập AI cho sinh viên Việt Nam không ạ?gì

Mình ko có ý định viết sách, vì lý do gần như tương tự như việc ko dám tính toán quá xa vời khi làm chips. Đó là: (1) viết sách rất tốn thời gian; (2) khi mình viết xong, có khi thế giới AI đã thay đổi quá xa rồi.

50. Vấn đề chính của cải thiện hiệu suất chính chỉ nằm chính ở hàm e^ và tính toán trên ma trận phải không ạ?
Chắc chắn 2 cái đó là quan trọng, nhưng còn rất nhiều cái khác nữa. Tiếc rằng trong giới hạn thời gian của bài nói chuyện, mình chỉ có thể đưa ra 2 ví dụ đó cho các bạn thôi.

51. Anh review quá trình anh qua Mỹ du học được không ạ? Và lời khuyên cho các bạn trẻ muốn du học qua Mỹ là gì ạ?
Cái này quá dài, mình review ngắn gọn thì lại chẳng ra gì, nên mình xin phép ko trả lời.

51. Dạ vậy anh trả lời câu hỏi ở đầu buổi "Tại sao MultiAgent có thể nhanh hơn SingleAgent" với ạ!
Bạn nhớ câu chuyện **m=1** không? Nếu dùng multi-agent, bạn có thể giảm số bước auto-regressive, và trong mỗi bước, bạn có **m>1.** Quan trọng nhất là bạn có **m>1**.

52. Dạ vậy kerẻnl được sử dụng ở đâu ạ, chip hay hệ điều hành….
Chip và hệ điều hành dùng 2 thứ kernels khác nhau. Nó hoàn toàn ko liên quan gì đến nhau hết, chỉ tình cờ cùng được gọi là kernel thôi.

Trong tiếng Việt, chắc nó giống như từ 'năm tháng' trong câu "tình yêu của chúng ta sẽ kéo dài qua năm tháng" đó.

53. Dạ thưa anh, theo như anh nói và em đọc qua nhiều tài liệu thì một con agent đã phải có LLM. nếu em sử dụng knowledge graph , rules, tự lập trình sensors và actions, mục đích là không cần dùng LLM, thì cho em hỏi, cách này có khả thi không ạ?
Có khả thi, nhưng mà knowledge graph và rules của em rất khó có thể đạt tính tổng quát cao như LLM.

54: LMM cơ chế đoán từng từ của câu rồi lặp lại n lần cho tới khi đoán hết câu, vậy thì sao tận dụng được cơ chế song song của GPU ạ?
Thì thế nên nó mới khó!!!! ^^"

Có nhiều thuật toán đã được tạo ra để giúp quá trình autoregressive tận dụng khả năng tính toán song song của GPUs tốt hơn. Một trong những thuật toán nổi tiếng nhất là speculative decoding. Các bạn có thể tìm hiểu. - Dạ em cảm ơn anh!

55: Link slide của anh up lên đâu anh ạ?
https://docs.google.com/presentation/d/1mD7Qv0Y5-KLDgVWblYQCwDFjtdz951rGsVp54ytKZ2M/edit?usp=sharing - Dạ em cảm ơn anh!