AI Engineering Worlds Fair

Jun 27, 2024

- Thomas Dohmke
    - Human centric approach - “co-pilot”
    - Copilot helps devs be in the flow of software
    - Democratizes access to information - onboarding
    - Agent - ai dishwasher
        - (side note: need an uptime display)
        - Security tooling - currently adds to backlog, not a dishwasher
        - New abstraction layers to get control of SDLC
    - AI brings the fun back to software development
- Whitepaper
    - https://www.oreilly.com/radar/what-we-learned-from-a-year-of-building-with-llms-part-i/
    - https://www.oreilly.com/radar/what-we-learned-from-a-year-of-building-with-llms-part-ii/
    - https://www.oreilly.com/radar/what-we-learned-from-a-year-of-building-with-llms-part-iii-strategy/
    - Hidden Technical Debt in Machine Learning Systems
- “Value is only created when the metal gets bent” - Shigeo Shingo
- Character.ai
    - Hierarchy of data needs
        - Clean
        - Evaluate
        - Systems for management
        - Analytics
        - Data set selection
        - Quality scoring and sampling methods
        - synthetics (enrichment, augmentation, translation)
    - Materialize data on demand during training
        - CAP theorem for AI data
    - Lance format
            - Columnar
            - fast lookups
            - Minimize io
            - No row groups + inline blobs
        - Table format
            - Version manifests + data files
            - Time travel
            - Schema evolution
        - Indexing extensions

            - IVF, PQ, HNSW*, SQ*, GQ*
            - Scalar indices
            - Full-text
    - Single table, many workloads
        - https://lancedb.github.io/lancedb/
    - Speed of iteration is critical
- Anthropic
    - Adding steering API - in beta
- Moondream
    - Focused on understanding screen caos
    - Fused with phi-1.5
    - Training data
        - Alt-text?
    - GPT-4 - not great - train it to hallucinate
    - Synthetic Data
        - COCO to detailed captions
        - Localized narratives - filter???
    - Key learnings
        - Engagement helped pivot, Connect with partners, mentors and more
        - Open Source is Critical - devs prefer, critical for engagement, needed because of competition
        - Safety guardrails should be implemented at the application layers
            - Dev-tools are b2b
        - Smol models - efficiency matters, critical use cases like drones, robotics. Privacy, latency cost
            - In prod - big models during dev?
        - Two types of use cases
            - Do new things (caption, question answering)
            - Do old things more easily (prompt for object detection, classification)
    -

Jun 26, 2024

- https://github.com/simonw - Simon W
- https://simonwillison.net/
- https://github.com/Mozilla-Ocho/llamafile - multi-platform wexecutable models
- “Iteration is the compound interest of software development” - Hypermode CEO
- Groq is building racks to keep up with developer demand
    - This implies high occupancy?
- Codium
    - Highest rated
    - Dimensions in embedding space is too small?
    - MTEB - needle in haystack (single needle)
    - Recall-50 - what percentage of relevant documents are in your top 50?
        - Use commits (text to code) for evals

        - Retriever does way better on recall50
        - M-query - parallel calls over lots of documents -> ranking
- Gradient - https://huggingface.co/gradientai - https://gradient.ai/blog
    - Custom language models
        - Finance domain models (Albatross)
            - Domain Knowledge
            - Hallucination
            - Tabular
            - Reasoning
            - Auditablility
            - Efficiency
        - Need 1000s of documents to answer questions on a topic well - https://arxiv.org/abs/2211.08411
        - Automatically detect if the domain-specific knowledge is contained - filter data
        - Alignment - SFT, DPO
    - **ICL is the most direct and sample efficient way to reduce hallucinations**
        - Puts demands on context length
        - Long context helps if there is lots of interrelated information
    - [2405.09798] Many-Shot In-Context Learning in Multimodal Foundation Models
    - Models
        - V-alpha-tross
        - llama3-70B-1048K
- Unsloth - fixing bugs in llama3 - https://unsloth.ai/blog/phi3
    - Tools and detailed guidance for finetuning
- Liquid - finetuning and model merging
    - https://github.com/mlabonne/llm-course
    - When to use fine-tuning?
    - Libs
        - UnSloth
        - LLaMA-Factory
        - Axylotl
    - Preference alignment - hf blog post - also author's post
    - https://github.com/mlabonne/llm-datasets
    - Learning rate - as high as possible until the loss explodes
    - Model merging
        - MergekKit
        - Results in good models (on openllm)
        - SLERP - spherical linear interpolation
            - mlabonne/NeuralBeagle14-7B
        - DARE -
            - Reduces redundancy
        - Passthrough
            - mlabonne/Meta-Llama-3-120B-Instruct

            - FrakenMoE - not as good as other methods
                - mlabonne/Beyonder-4x7B-v3
                - mlabonne/phixtral-4x2_8
- Define success criteria early
    - Iteraiton generates good conversation
- Synthetic data allows fast adaptation
- BotDojo demo
    - Use real support sessions for evals

- https://x.com/_chenglou
    - Predicting second order effects
    - Who is learning?
        - People will use it to learn more quickly by reducing assistants
        - Manipulation of UI is learning for yourself
            - Lifestyle interfaces - you are using it to learn
    - Information Bandwidth
        - Personalize communication to translate between people's nuanced communication styles
        - Help with conflict resolution ?
        - Confusing behavior with implicit ruleset (drawing app with stylus)
    - Raise Order of Magnitude
        - The more agents, the more the aggregate matters
        - The more agents, the less you care about the individuals
        - Generate tons of options and filter
            - Like media-queries
            - User is onboarding
                - Progressively show new layouts
            - → Dynamic UIs
- Spreadsheets are all you need
    - https://www.lesswrong.com/tag/transformers
    - Sparse autoencoders for open source - neuronopedia https://www.neuronpedia.org/
        - Representation engineering
        - Activation steering
- Jerry Liu (llamaindex) - future of llm assistants
    - Contex-augmented research assistant
        - Advance data and retrieval modules
        - Advanced single-agent query flows (tools)
        - General multi-agent task solver (orchestration)
    - Advanced data and retrieval
        - Parsing - extract data well (tables for example)
        - chunking and indexing
    - Advanced single agent flows
        - Routing

        - Function calling / tool use
        - Query planning
        - Conversation memory
    - Agentic RAG
        - Every data interfaces is a tool
        - Use agent reading loops
    - Remaining gaps
        - Use specialist agents
        - Agents may interface with other agents
    - Multi-agent task solver
        - Specialization
        - Parallelization
        - cost/latency???
        - Llama Agents
            - Agents as microservices
            - Communicate through central API (e.g. slono's task queue)
            - Orchestration happens via a control plane
        - https://github.com/run-llama/llama-agents

Jun 25, 2024

- Manual stories
    - Example book
    - Questions needed to qualify candidates and get the good data
    - https://www.linkedin.com/in/jourdan-smith-7298545a/

Slono's presentation

- 2024-06-24 - Workshop AI Programmer Handout (1).pdf
- KBall's link: https://sourcegraph.com/blog/the-death-of-the-junior-developer

DeepGram

- The ESP-BOX is a new generation AIoT development platform released by Espressif Systems.
- Friend: Open Source AI Wearable Recording Device by Nik Shevchenko — Kickstarter

Code INterpreter

- https://e2b.dev/docs/getting-started/api-key
- https://github.com/e2b-dev/e2b-cookbook code interpreter
- https://github.com/e2b-dev/e2b sandbox
- https://sdk.vercel.ai/docs/introduction
- https://firecracker-microvm.github.io/

Neo4j

- https://colab.research.google.com/drive/1ucnpA-biyng_1dUFr3wuPkiA6_MKct5Z#scrollTo=67Tm1p3LdyXe
    - you can use gradio inside of colab!
    - Uses graph embeddings, text embeddings, examples in the UI