Reflexões sobre a OpenAI

15 DE JULHO DE 2025

Deixei a OpenAI há três semanas. Eu tinha entrado na empresa em maio de 2024.

Queria compartilhar minhas reflexões porque há muita fumaça e ruído sobre o que a OpenAI está fazendo, mas não há muitos relatos em primeira mão sobre como é realmente a cultura de trabalho lá.

Nabeel Qureshi tem um post incrível chamado Reflexões sobre a Palantir, onde ele reflete sobre o que tornou a Palantir especial. Eu queria fazer o mesmo pela OpenAI enquanto está fresco em minha mente. Você não encontrará segredos comerciais aqui, apenas reflexões sobre esta iteração atual de uma das organizações mais fascinantes da história em um momento extremamente interessante.

Para deixar claro desde o início: não houve nenhum drama pessoal em minha decisão de sair – na verdade, fiquei profundamente conflitado sobre isso. É difícil passar de ser fundador de sua própria empresa para funcionário em uma organização de 3.000 pessoas. Agora estou desejando um recomeço.

É totalmente possível que a qualidade do trabalho me faça voltar. É difícil imaginar construir algo tão impactante quanto AGI, e LLMs são facilmente a inovação tecnológica da década. Me sinto sortudo por ter visto alguns dos desenvolvimentos em primeira mão e também por ter feito parte do lançamento do Codex.

Obviamente, estas não são as visões da empresa – como observações, são minhas. A OpenAI é um lugar grande, e esta é minha pequena janela para ela.

Cultura

A primeira coisa a saber sobre a OpenAI é quão rapidamente ela cresceu. Quando entrei, a empresa tinha pouco mais de 1.000 pessoas. Um ano depois, tem mais de 3.000 e eu estava nos 30% superiores por tempo de casa. Quase todos na liderança estão fazendo um trabalho drasticamente diferente do que faziam há ~2-3 anos. [1]

Claro, tudo quebra quando você escala tão rapidamente: como se comunicar como empresa, as estruturas de reporte, como entregar produtos, como gerenciar e organizar pessoas, os processos de contratação, etc. As equipes variam significativamente em cultura: algumas estão correndo a todo vapor o tempo todo, outras estão cuidando de grandes execuções, algumas estão se movendo em um ritmo muito mais consistente. Não há uma experiência única da OpenAI, e pesquisa, aplicada e GTM operam em horizontes temporais muito diferentes.

Uma parte incomum da OpenAI é que tudo, e quero dizer tudo, funciona no Slack. Não há e-mail. Talvez eu tenha recebido ~10 e-mails em todo o meu tempo lá. Se você não

for organizado, achará isso incrivelmente perturbador. Se você curar seus canais e notificações, pode torná-lo bastante funcional.

A OpenAI é incrivelmente de baixo para cima, especialmente em pesquisa. Quando cheguei, comecei a fazer perguntas sobre o roteiro para o próximo trimestre. A resposta que recebi foi: "isso não existe" (embora agora exista). Boas ideias podem vir de qualquer lugar, e muitas vezes não está realmente claro quais ideias serão mais frutíferas com antecedência. Em vez de um grande 'plano mestre', o progresso é iterativo e descoberto à medida que novas pesquisas dão frutos.

Graças a essa cultura de baixo para cima, a OpenAI também é muito meritocrática. Historicamente, os líderes da empresa são promovidos principalmente com base em sua capacidade de ter boas ideias e executá-las. Muitos líderes que eram incrivelmente competentes não eram muito bons em coisas como apresentar em reuniões gerais ou manobras políticas. Isso importa menos na OpenAI do que poderia em outras empresas. As melhores ideias tendem a vencer. [2]

Há um forte viés para ação (você pode simplesmente fazer as coisas). Não era incomum que equipes similares mas não relacionadas convergissem em várias ideias. Comecei trabalhando em um esforço paralelo (mas interno) similar aos Conectores do ChatGPT. Deve ter havido ~3-4 protótipos diferentes do Codex flutuando antes de decidirmos pressionar por um lançamento. Esses esforços geralmente são feitos por um pequeno punhado de indivíduos sem pedir permissão. As equipes tendem a se formar rapidamente em torno deles à medida que mostram promessa.

Andrey (o líder do Codex) costumava me dizer que você deveria pensar nos pesquisadores como seu próprio "mini-executivo". Há um forte viés para trabalhar em sua própria coisa e ver como ela se desenvolve. Há um corolário aqui – a maior parte da pesquisa é feita ao fisgar um pesquisador para um problema específico. Se algo é considerado chato ou 'resolvido', provavelmente não será trabalhado.

Bons gerentes de pesquisa são insanamente impactantes e também incrivelmente limitados. Os melhores conseguem conectar os pontos entre muitos esforços de pesquisa diferentes e reunir um treinamento de modelo maior. O mesmo vale para grandes PMs (um salve para ae).

Os EMs do ChatGPT com quem trabalhei (Akshay, Rizzo, Sulman) eram alguns dos caras mais tranquilos que já vi. Realmente parecia que eles já tinham visto tudo a essa altura [3]. A maioria deles era relativamente não invasiva, mas contratava boas pessoas e tentava garantir que estivessem preparadas para o sucesso.

A OpenAI muda de direção rapidamente. Isso era algo que valorizávamos muito na Segment – é muito melhor fazer a coisa certa quando você obtém novas informações, versus decidir manter o curso só porque você tinha um plano. É notável que uma empresa tão grande quanto a OpenAI ainda mantenha esse ethos – o Google claramente não. A empresa toma decisões rapidamente e, ao decidir seguir uma direção, vai com tudo.

Há uma tonelada de escrutínio sobre a empresa. Vindo de um background empresarial B2B, isso foi um pouco chocante para mim. Eu regularmente via notícias publicadas na imprensa que ainda não haviam sido anunciadas internamente. Eu dizia às pessoas que trabalho na OpenAI e era recebido com uma opinião pré-formada sobre a empresa. Vários usuários do Twitter executam bots automatizados que verificam se há novos lançamentos de recursos por vir.

Como resultado, a OpenAI é um lugar muito secreto. Eu não podia contar a ninguém no que estava trabalhando em detalhes. Há um punhado de espaços de trabalho do Slack com várias permissões. Os números de receita e queima são mais bem guardados.

A OpenAI também é um lugar mais sério do que você poderia esperar, em parte porque as apostas parecem realmente altas. Por um lado, há o objetivo de construir AGI – o que significa que há muito a acertar. Por outro lado, você está tentando construir um produto que centenas de milhões de usuários aproveitam para tudo, desde conselhos médicos até terapia. E por outro, outro lado, a empresa está competindo na maior arena do mundo. Prestávamos muita atenção ao que estava acontecendo na Meta, Google e Anthropic – e tenho certeza de que todos estavam fazendo o mesmo. Todos os principais governos mundiais estão observando este espaço com grande interesse.

Por mais que a OpenAI seja difamada na imprensa, todos que conheci lá estão realmente tentando fazer a coisa certa. Dado o foco no consumidor, é o mais visível dos grandes laboratórios e, consequentemente, há muita difamação contra ela.

Dito isso, você provavelmente não deveria ver a OpenAI como um monólito único. Penso na OpenAI como uma organização que começou como Los Alamos. Era um grupo de cientistas e inventores investigando a vanguarda da ciência. Esse grupo aconteceu de acidentalmente gerar o aplicativo de consumo mais viral da história. E então cresceu para ter ambições de vender para governos e empresas. Pessoas de diferentes tempos de casa e diferentes partes da organização consequentemente têm objetivos e pontos de vista muito diferentes. Quanto mais tempo você está lá, mais provavelmente você vê as coisas através da lente de "laboratório de pesquisa" ou "sem fins lucrativos para o bem".

A coisa que mais aprecio é que a empresa "pratica o que prega" em termos de distribuir os benefícios da IA. Modelos de ponta não são reservados para algum nível empresarial com um acordo anual. Qualquer pessoa no mundo pode entrar no ChatGPT e obter uma resposta, mesmo que não esteja logada. Há uma API que você pode se inscrever e usar – e a maioria dos modelos (mesmo se SOTA ou proprietários) tendem a rapidamente entrar na API para startups usarem. Você poderia imaginar um regime alternativo que opera de forma muito diferente do que temos hoje. A OpenAI merece muito crédito por isso, e ainda é fundamental para o DNA da empresa.

A segurança é realmente mais importante do que você poderia imaginar se lê muito do Zvi ou Lesswrong. Há um grande número de pessoas trabalhando para desenvolver sistemas de segurança. Dada a natureza da OpenAI, vi mais foco em riscos práticos

(discurso de ódio, abuso, manipulação de vieses políticos, criação de armas biológicas, automutilação, injeção de prompt) do que teóricos (explosão de inteligência, busca de poder). Isso não quer dizer que ninguém está trabalhando no último, definitivamente há pessoas focando nos riscos teóricos. Mas do meu ponto de vista, não é o foco. A maior parte do trabalho que é feito não é publicado, e a OpenAI realmente deveria fazer mais para divulgá-lo.

Ao contrário de outras empresas que distribuem livremente seus brindes em todas as feiras de carreira, a OpenAI realmente não dá muito brinde (mesmo para novos funcionários). Em vez disso, há 'drops' que acontecem onde você pode pedir itens em estoque. O primeiro derrubou a loja Shopify, teve tanta demanda. Havia um post interno que circulava sobre como POST os payloads json corretos e contornar isso.

Quase tudo é um erro de arredondamento comparado ao custo de GPU. Para dar uma ideia: um recurso de nicho que foi construído como parte do produto Codex tinha a mesma pegada de custo de GPU que toda a nossa infraestrutura Segment (não na mesma escala que o ChatGPT, mas via uma porção decente do tráfego da internet).

A OpenAI é talvez a organização mais assustadoramente ambiciosa que já vi. Você pode pensar que ter um dos principais aplicativos de consumo do planeta pode ser suficiente, mas há um desejo de competir em dezenas de arenas: o produto API, pesquisa profunda, hardware, agentes de codificação, geração de imagem e um punhado de outros que não foram anunciados. É um terreno fértil para pegar ideias e executá-las.

A empresa presta muita atenção ao Twitter. Se você tweetar algo relacionado à OpenAI que se torne viral, as chances são boas de que alguém leia sobre isso e considere. Um amigo meu brincou: "esta empresa funciona com vibes do Twitter". Como uma empresa de consumo, talvez isso não esteja tão errado. Certamente ainda há muita análise em torno de uso, crescimento de usuários e retenção – mas as vibes são igualmente importantes.

As equipes na OpenAI são muito mais fluidas do que poderiam ser em outros lugares. Ao lançar o Codex, precisávamos de ajuda de alguns engenheiros experientes do ChatGPT para atingir nossa data de lançamento. Nos reunimos com alguns dos EMs do ChatGPT para fazer o pedido. No dia seguinte, tínhamos duas pessoas incríveis prontas para mergulhar e ajudar. Não houve "esperar pelo planejamento trimestral" ou "reorganização de headcount". Moveu-se muito rapidamente.

A liderança é bastante visível e fortemente envolvida. Isso pode ser óbvio em uma empresa como a OpenAI, mas todos os executivos pareciam bastante sintonizados. Você veria gdb, sama, kw, mark, dane, et al intervir regularmente no Slack. Não há líderes ausentes.

Código

A OpenAI usa um monorepo gigante que é ~principalmente Python (embora haja um conjunto crescente de serviços Rust e um punhado de serviços Golang espalhados para coisas como proxies de rede). Isso cria muito código de aparência estranha porque há muitas maneiras de escrever Python. Você encontrará bibliotecas projetadas para escala de veteranos de 10 anos do Google, bem como notebooks Jupyter descartáveis de PhDs recém-formados. Praticamente tudo opera em torno do FastAPI para criar APIs e Pydantic para validação. Mas não há guias de estilo aplicados em larga escala.

A OpenAI executa tudo no Azure. O engraçado sobre isso é que há exatamente três serviços que eu consideraria confiáveis: Azure Kubernetes Service, CosmosDB (armazenamento de documentos do Azure) e BlobStore. Não há equivalentes verdadeiros de Dynamo, Spanner, Bigtable, Bigquery Kinesis ou Aurora. É um pouco mais raro pensar muito em unidades de escalonamento automático. As implementações de IAM tendem a ser muito mais limitadas do que você obteria de um AWS. E há um forte viés para implementar internamente.

Quando se trata de pessoal (pelo menos em engenharia), há um pipeline muito significativo Meta → OpenAI. De muitas maneiras, a OpenAI se assemelha ao início da Meta: um aplicativo de consumo de sucesso, infraestrutura nascente e um desejo de se mover muito rapidamente. A maior parte do talento de infraestrutura que vi trazido da Meta + Instagram tem sido bastante forte.

Junte essas coisas e você vê muitas partes principais da infraestrutura que parecem reminiscentes da Meta. Havia uma reimplementação interna do TAO. Um esforço para consolidar a identidade de autenticação na borda. E tenho certeza de vários outros que não conheço.

O chat é muito profundo. Desde que o ChatGPT decolou, muito do código base é estruturado em torno da ideia de mensagens de chat e conversas. Esses primitivos estão tão enraizados neste ponto que você provavelmente deveria ignorá-los por sua própria conta e risco. Desviamos um pouco deles no Codex (apoiando-nos mais nos aprendizados da API de respostas), mas aproveitamos muita arte anterior.

O código vence. Em vez de ter algum comitê central de arquitetura ou planejamento, as decisões geralmente são tomadas pela equipe que planeja fazer o trabalho. O resultado é que há um forte viés para ação e, muitas vezes, várias partes duplicadas do código base. Devo ter visto meia dúzia de bibliotecas para coisas como gerenciamento de filas ou loops de agentes.

Havia algumas áreas onde ter uma equipe de engenharia rapidamente escalada e não muitas ferramentas criava problemas. sa-server (o monólito backend) era um pouco de um depósito. CI quebrava muito mais frequentemente do que você poderia esperar no master. Casos de teste mesmo rodando em paralelo e fatorando um subconjunto de dependências podiam levar ~30m para rodar em GPUs. Estes não eram problemas insolúveis, mas é um bom lembrete de que esses tipos de problemas existem em todos os lugares, e provavelmente piorarão quando você escalar super rapidamente. Para crédito das equipes internas, há muito foco indo para melhorar essa história.

Outras coisas que aprendi

Como é uma grande marca de consumo. Eu realmente não havia internalizado isso até começarmos a trabalhar no Codex. Tudo é medido em termos de 'assinaturas pro'. Mesmo para um produto como o Codex, pensamos na integração principalmente relacionada ao uso individual em vez de equipes. Quebrou um pouco meu cérebro, vindo de um background predominantemente B2B / empresarial. Você vira um interruptor e obtém tráfego desde o dia 1.

Como grandes modelos são treinados (em alto nível). Há um espectro de "experimentação" a "engenharia". A maioria das ideias começa como experimentos em pequena escala. Se os resultados parecem promissores, eles são incorporados em uma execução maior. A experimentação é tanto sobre ajustar os algoritmos centrais quanto ajustar a mistura de dados e estudar cuidadosamente os resultados. No extremo grande, fazer uma grande execução quase parece engenharia de sistemas distribuídos gigantes. Haverá casos extremos estranhos e coisas que você não esperava. Cabe a você depurá-los.

Como fazer matemática de GPU. Tivemos que prever os requisitos de capacidade de carga como parte do lançamento do Codex, e fazer isso foi a primeira vez que realmente passei fazendo benchmark de GPUs. A essência é que você deve realmente começar dos requisitos de latência que precisa (latência geral, # de tokens, tempo até o primeiro token) versus fazer análise de baixo para cima sobre o que uma GPU pode suportar. Cada nova iteração de modelo pode mudar os padrões de carga drasticamente.

Como trabalhar em uma grande base de código Python. A Segment era uma combinação de microsserviços e era principalmente Golang e Typescript. Realmente não tínhamos a amplitude de código que a OpenAI tem. Aprendi muito sobre como escalar uma base de código com base no número de desenvolvedores contribuindo para ela. Você tem que colocar muito mais proteções para coisas como "funciona por padrão", "manter o master limpo" e "difícil de usar mal".

Lançando o Codex

Uma grande parte dos meus últimos três meses na OpenAI foi lançar o Codex. É inquestionavelmente um dos destaques da minha carreira.

Para preparar o cenário, em novembro de 2024, a OpenAI havia estabelecido uma meta de 2025 para lançar um agente de codificação. Em fevereiro de 2025, tínhamos algumas ferramentas internas flutuando que estavam usando os modelos com grande efeito. E estávamos sentindo a pressão para lançar um agente específico de codificação. Claramente, os modelos haviam chegado ao ponto em que estavam ficando realmente úteis para codificação (vendo a nova explosão de ferramentas de vibe-coding no mercado).

Voltei cedo da minha licença paternidade para ajudar a participar do lançamento do Codex. Uma semana depois que voltei, tivemos uma fusão (ligeiramente caótica) de duas equipes e começamos um sprint louco. Do início (as primeiras linhas de código escritas) ao fim, todo o produto foi construído em apenas 7 semanas.

O sprint do Codex foi provavelmente o mais duro que trabalhei em quase uma década. A maioria das noites ficava até 23h ou meia-noite. Acordando com um recém-nascido às 5h30 todas as manhãs. Indo para o escritório novamente às 7h. Trabalhando a maioria dos fins de semana. Todos nós pressionamos forte como equipe, porque cada semana contava. Me lembrou de estar de volta ao YC.

É difícil exagerar quão incrível foi esse nível de ritmo. Não vi organizações grandes ou pequenas irem de uma ideia para um produto totalmente lançado + disponível gratuitamente em uma janela tão curta. O escopo também não era pequeno; construímos um runtime de contêiner, fizemos otimizações no download de repositórios, ajustamos finamente um modelo personalizado para lidar com edições de código, lidamos com todo tipo de operações git, introduzimos uma área de superfície completamente nova, habilitamos acesso à internet e terminamos com um produto que geralmente era uma delícia de usar. [4]

Digam o que quiserem, a OpenAI ainda tem esse espírito de lançamento. [5]

A boa notícia é que as pessoas certas podem fazer mágica acontecer. Éramos uma equipe sênior de ~8 engenheiros, ~4 pesquisadores, 2 designers, 2 GTM e um PM. Se não tivéssemos esse grupo, acho que teríamos falhado. Ninguém precisava de muita direção, mas precisávamos de uma quantidade decente de coordenação. Se você tiver a chance de trabalhar com alguém da equipe Codex, saiba que cada um deles é fantástico.

Na noite anterior ao lançamento, cinco de nós ficamos acordados até as 4h tentando implantar o monólito principal (um assunto de várias horas). Então foi de volta ao escritório para o anúncio de lançamento das 8h e livestream. Ligamos as flags e começamos a ver o tráfego fluir. Nunca vi um produto obter tanto uptick imediato apenas por aparecer em uma barra lateral esquerda, mas esse é o poder do ChatGPT.

Em termos da forma do produto, estabelecemos um fator de forma que era totalmente assíncrono. Ao contrário de ferramentas como Cursor (na época, agora suporta um modo similar) ou Claude Code, nosso objetivo era permitir que os usuários iniciassem tarefas e deixassem o agente rodar em seu próprio ambiente. Nossa aposta era que, no final do jogo, os usuários deveriam tratar um agente de codificação como um colega de trabalho: eles enviariam mensagens ao agente, ele teria algum tempo para fazer seu trabalho e então voltaria com um PR.

Isso foi um pouco de aposta: estamos em um estado ligeiramente estranho hoje onde os modelos são bons, mas não ótimos. Eles podem trabalhar por minutos de cada vez, mas ainda não por horas. Os usuários têm graus amplamente variados de confiança nas capacidades dos modelos. E nem estamos claros sobre quais são as verdadeiras capacidades dos modelos.

Ao longo do arco do tempo, acredito que a maior parte da programação se parecerá mais com o Codex. Enquanto isso, será interessante ver como todos os produtos se desdobram.

O Codex (talvez sem surpresa) é realmente bom em trabalhar em uma grande base de código, entendendo como navegá-la. O maior diferencial que vi versus outras ferramentas é a capacidade de iniciar várias tarefas ao mesmo tempo e comparar sua saída.

Vi recentemente que há números públicos comparando os PRs feitos por diferentes agentes LLM. Apenas pelos números públicos, o Codex gerou 630.000 PRs. Isso é cerca de 78k PRs públicos por engenheiro nos 53 dias desde o lançamento (você pode fazer suas próprias suposições sobre o múltiplo de PRs privados). Não tenho certeza se já trabalhei em algo tão impactante em minha vida.

Pensamentos finais

Para falar a verdade, eu estava originalmente apreensivo sobre entrar na OpenAI. Eu não tinha certeza de como seria sacrificar minha liberdade, ter um chefe, ser uma peça muito menor de uma máquina muito maior. Mantive bastante discreto que havia entrado, caso não fosse o ajuste certo.

Eu queria obter três coisas da experiência...

- construir intuição sobre como os modelos eram treinados e para onde as capacidades estavam indo
- trabalhar e aprender com pessoas incríveis
- lançar um ótimo produto

Refletindo sobre o ano, acho que foi uma das melhores decisões que já tomei. É difícil imaginar aprender mais em qualquer outro lugar.

Se você é um fundador e sente que sua startup realmente não está indo a lugar nenhum, você deveria 1) reavaliar profundamente como pode fazer mais tentativas ou 2) ir se juntar a um dos grandes laboratórios. Agora é um momento incrível para construir. Mas também é um momento incrível para espiar onde o futuro está indo.

Como vejo, o caminho para AGI é uma corrida de três cavalos agora: OpenAI, Anthropic e Google. Cada uma dessas organizações vai tomar um caminho diferente para chegar lá com base em seu DNA (consumidor vs negócios vs infraestrutura sólida + dados). [6] Trabalhar em qualquer uma delas será uma experiência reveladora.

Obrigado a Leah por ser incrivelmente solidária e assumir a maior parte dos cuidados com as crianças durante as noites tardias. Obrigado a PW, GDB e Rizzo por me darem uma chance. Obrigado aos colegas de equipe SA por me ensinarem as cordas: Andrew, Anup, Bill, Jeremy, Kwaz, Ming, Simon, Tony e Val. E obrigado à equipe principal do Codex por me dar a viagem da minha vida: Albin, AE, Andrey, Bryan, Channing, DavidK,

Gabe, Gladstone, Hanson, Joey, Josh, Katy, KevinT, Max, Sabrina, SQ, Tibo, TZ e Will. Nunca esquecerei este sprint.

Wham.

---

[1] É fácil tentar interpretar muito drama sempre que há um líder saindo, mas eu atribuiria ~70% deles apenas a esse fato. ↩

[2] Acho que estamos em uma ligeira mudança de fase aqui. Há muitas contratações de liderança sênior sendo feitas de fora da empresa. Geralmente sou a favor disso, acho que a empresa se beneficia muito ao infundir novo DNA externo. ↩

[3] Tenho a sensação de que escalar o produto de consumo de crescimento mais rápido de todos os tempos tende a construir muito músculo. ↩

[4] Claro, também estávamos nos ombros de gigantes. A equipe CaaS, equipes principais de RL, dados humanos e infraestrutura aplicada geral tornaram tudo isso possível. ↩

[5] Continuamos também. ↩

[6] Vimos algumas grandes contratações na Meta há algumas semanas. A xAI lançou o Grok 4 que tem bom desempenho em benchmarks. Mira e Ilya têm grandes talentos. Talvez isso mude as coisas (as pessoas são boas). Eles têm algum alcance a fazer. ↩