

Error Analysis: The Highest ROI Technique In AI Engineering

12 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=E2i6JbU2R-S](https://www.youtube.com/watch?v=E2i6JbU2R-S)

<https://www.youtube.com/watch?v=e2i6JbU2R-s>

SUMMARY

Jacob Carter, founder of Nurture Boss, discusses his experience with error analysis (AOR analysis) in AI product development, emphasizing its significance in understanding user interactions and improving product offerings. He highlights the importance of analyzing real data to inform product roadmaps and enhance user experience, particularly in the multifamily housing sector.

- AOR analysis is crucial for building effective AI products, especially for those without a data science background.*
- Nurture Boss creates AI solutions for the apartment industry, including virtual leasing assistants that handle various resident interactions.*
- Analyzing thousands of conversations revealed user assumptions about the product, leading to necessary adjustments and bug fixes.*
- A custom data viewer was developed to streamline the analysis process, allowing for quick categorization of conversation outcomes.*
- Identifying specific issues, such as date handling errors, led to significant improvements in AI response accuracy.*
- The process of error analysis involves detailed note-taking and categorization of errors, which informs product development.*
- Jacob's approach emphasizes the need for tailored metrics and tools rather than relying on generic evaluation methods.*
- The insights gained from data analysis have directly influenced the product roadmap and enhanced the overall customer experience.*

hey everyone in today's video I'm sharing a conversation with Jacob Carter founder of nurture boss and Jacob is going to walk you through his experience with aor analysis which is my favorite technique by far when building AI products because it has such a high impact what's fascinating about aor analysis is that many people I would go even as far to say most people people don't know about it especially if you're

not coming from a machine learning or data science background if you've read my recent blog post a field guide to improving AI products you'll notice error analysis as a key theme while the blog post has several case studies in it I'm excited that you get to hear from Jacob himself quick background of nurture boss and what we do we build AI products for for multif family or the apartment industry uh AI products for us means virtual leasing assistant so IM manage

uh uh imagine the folks that are working on site at apartment communities they're answering phones uh having

conversations with prospective residents helping to get tour scheduled uh answer questions around amenities or pricing and availability uh and then also following up with those prospects uh via text and email to try to drive them uh to convert to become residents and then once they are residents there's all kinds of work around collecting rent on time chasing down over do rent uh boosting online

reviews uh for that property um facilitating maintenance requests and service requests so all these things nurture boss has AI products that do that on behalf of the on-site team I'm so worth calling out the main channels that we communicate with are voice um email and text uh and uh those are all going to operate and work a little bit differently um but they do uh ultimately all run uh on open AI so when we released our products um you know it was

a giant black box for us in terms of what was happening if it was going well if it was not going well in an effort to self-educate uh I took to the internet and stumbled upon uh Hamill's blog which I learned a ton from and ultimately just led me to reach out to him and ask if he could help us kind of um become experts uh in developing iterating on uh and measuring uh AI products and uh

that's when we we got into the conversation and the look at your own data conversation so um for context I uh had no idea or understood at all the value that would come from quite literally just looking at our data uh Hamil I don't know if I ever express this with you but it was kind of frustrating when you kept telling me that like what is the secret here that I'm going to learn and then I got into

it uh and realized uh how impactful insightful and um important it was that we do that and we'll continue to do that that so to this point you know I've looked at thousands of conversations via phone email and text that our AI has had with residents or prospective residents and a few things um have have stood out um that I think are worth worth mentioning really quick uh one of those is getting an understanding of what we

should be building so um my background I've been a software engineer for about 15 years I've spent a handful of those years on the product side of the house versus the engineering side of the house and when you're trying to decide what to build for your customers you do Discovery uh you do um you know beta testing you do Journey Maps you have all these processes being able to actually see what people are asking your product

uh every day um and identifying gaps where they're looking to get problem solved that our product does not know how to solve it builds your product road map for you so that's been really exciting and really helpful to be able to get that out of it equally as important though um you start understanding how people assume your product is going to work which never lines up with how you thought people were going to assume your product would

work so it identifies a lot of bugs adjustments and changes uh that you need to make based off of what users are looking for uh out of your product versus what you've you've actually uh built and those patterns um really just surface and become obvious when you start reading more more conversations you also start to get an idea of how uh the AI model is going to respond so you make assumptions there as well but when you see it respond to

the same

type of questions again and again you really start to be able to kind of put yourself into the mind of what the model is going to do which informs uh how you build features uh in the future and how you work on different things like tools and rag um when uh you know impr prompt engineering when um J would it be would it be helpful to share your thing your data viewer bring that up now yeah

um so when we started getting into looking at this stuff um you know we uh built our own Tool uh like a lot of people do to quickly iterate and go through uh and judge uh different conversations so these are text message conversations um and then we can read through this conversation and see in detail what happened with the AI the human said this the AI said that the AI uh called A Tool uh this is the response that the AI got

back from the tool and then I can come through here and say if this conversation was good or bad and then I can quickly type up notes explaining why it was good or bad and the act of doing this I learn a lot but then also it's really easy to build tools that are also powered by AI saying hey for all of my text messages categorize the negative results that I get and by clicking that I see okay I'm getting tour scheduling

errors at the most uh rescheduling which I would group with tour scheduling uh handoffs were not triggered what that means is we wanted the AI to call in help from a human and it didn't um that's what a handoff would be considered so it's really easy for me to see this uh and get an idea of what's going well and not going well and now that I have that I can just get to work like that's enough

for me to just get to work uh as a product owner I can start creating bug tickets creating feature tickets uh instructing the engineering team on what to focus on and what to build so that alone is incredibly valuable but what we then can do is take this data into a tool like Brain Trust and start running uh actual uh automated tasks you know the variety of evals whether it be unit tests uh or or other um to uh start

understanding how changes we're making or actually implementing or solving those problems so one thing that I guess I'll just bring up because it's specific uh that we use Brain Trust to solve um everybody deals with this problem but I think nobody expects to deal with it which is that AI models are not very good at understanding dates what day it is what the future is what the past is so we would have a lot of customers say

hey we want to schedule a tour for two weeks from now and if they ask that question today the AI would say okay two weeks from now that's February 28th or February 29th uh if we're on a leap year March 1st but definitely not two weeks from now right or uh we'd get somebody say hey we want to schedule a tour for March 1st and the AI would look for tour availability for March 1st of 2022 so

this date handling issue was a big problem for us and we were able to use the playground found in uh Brain Trust to um create a bunch of use cases of or tests of different ways to reference dates and then see the output the model is giving us for what it interpreted those dates as and we went from a 33% success rate to a 100%

success rate through iterating on that problem uh through through Brain Trust so I think

that's another just hyper one thing I want to highlight here is like we're not using so we're not we you know we're not talking about generic stuff like conciseness score hallucination score all that nonsense you know we're talking about real pain points that Jacob identified and now he is addressing and you know we're going to start measuring that soon we're just got off the you know we're just ending this like error analysis stage but he's making real

progress like he is moving forward it's not just some you know exercise academic exercise which is a lot of people go through when they try to do evil he's actually making real progress you can even hear it in his story like he has an understanding yeah yeah absolutely every everything that we've been doing really from the start immediately started exposing action items that our product team uh could operate on immediately and start improving our product creating a

better experience for our customers and also educate us internally on how to build things moving forward based on the learnings that you get by just looking at your data which again while it sounds so simple and silly I can't stress enough how impactful it's actually been to just look at the data and see uh what's going on with the conversations the other thing I just I would mention last thing I'd mention really quick around kind of those generic comments uh

that Hamill's making and the lack of value in the generic evals is that inevitably your system's going to be set up in a way that not only makes it so the generic eval doesn't make sense but even you know a lot of the generic eval tools won't work perfectly uh off the shelf you have to understand the nuances of your setup so that you can use these things in a way that's valuable for you so what I mean by that is I think Hill

you know one of the big realizations that we had is when getting into our data is that we really have Evergreen um uh threads if you will right because these conversations don't just start and then they're over they go on for the lifetime of somebody looking for an apartment to renting the apartment to living in the apartment to renewing the lease so on and so forth so um it's not like I just have conversation I can say

here's a start and stop completed conversation it's Evergreen there may be more there may be less uh you know we don't really know when it's going to end and that's something we had to think about and consider by getting into the data by looking at the data you start to realize why it's not working or making sense what your unique use case is and you can start to either build tooling around that or know how to leverage

existing tooling uh in a way that makes sense for the way that you have things set up okay so we heard from Jacob about his journey with error analysis the there's a couple of things I think we should take away from that one Jacob didn't use off-the-shelf metrics or generic approaches to diagnosing his AI issues he started by looking at his own data and you heard from Jacob himself he used what he found to drive his entire product road map it

was that powerful second Jacob created his own data viewer and I think this is really important because it removed all friction from looking at data and Jacob was able to render the data in a way that made sense for him the process of doing error analysis involves looking at data writing detailed notes about problems you find and then categorizing those notes and Counting which errors are uh occurring

more most often and that's the simplest form of of error analysis and that's what I recommend starting with what you saw is that Jacob even automated a bit of this by using llms to help categorize those notes and then aggregating those categories and Counting them lastly I'd like to take I'd like to thank Jacob for sharing his experience not all companies are that open I hope you learned something about error analysis today are it has piqued your

interest either way I'll drop a link to the blog post that accompanies this video in the show notes thank you