LLM Eval Office Hours #4: Taming Complexity by Scoping LLM Evals

26 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=QFY_HRJJN-S
https://www.youtube.com/watch?v=qfY_HRjjN-s

SUMMARY

The discussion revolves around the challenges and strategies involved in improving the evaluation process for a lawn and garden startup's chatbot, which provides tailored advice to customers. The focus is on using a language model (LM) as a judge to assess the quality of responses, while navigating the complexities of open-ended questions and varying customer needs.

- *The startup offers a subscription service for lawn and garden products, utilizing technology for personalized recommendations.*
- *The chatbot, Sunny, assists users by providing tailored advice based on their lawn's specific conditions.*
- *The evaluation process currently achieves about 80% alignment between human critiques and LM assessments, an improvement from 50-60%.*
- *Challenges include handling open-ended questions and ensuring consistent quality across various topics, especially during peak seasons.*
- *The conversation highlights the importance of focusing on a smaller set of high-traffic topics to improve the chatbot's performance and user experience.*
- *Suggestions include using synthetic data generation to anticipate seasonal question variations and refining prompts for the LM to enhance evaluation accuracy.*
- *The need for ongoing human oversight in the evaluation process is emphasized, as complete automation may not be feasible.*
- *The discussion concludes with a plan to articulate a focus on specific topics to users while acknowledging that other areas may not be as polished initially.*

so I work at Sunday which is a lawn and garden startup we sell lawn and garden products and plans that are on a subscription service and it uses Tech and e-commerce to customize these to the unique needs of your lawn and I'll show I want to share my screen here I'm a customer by the way thanks for saying that yeah I actually I I took a look at your some of your account and I have some questions for you about your lawn

but I'll start yeah I don't even know what I'm doing I just buy stuff and it can I ask you question yeah I saw that you bought Bermuda grass seed why did you buy that grass seed I don't have a good answer I think like I have some weird places in my house there's like in I have a backyard that is North facing that's shaded like 100% And then like a my front lawn is south facing so maybe it's some one of

those I think there are better grass seeds for that that set yeah you should buy our seed called shade select I'll follow up with the email after this oh cool okay yeah yeah for your location but can you see this is our website

yeah I can see it yeah on our website we have a yard assistant it's a chatbot situation it looks like this so meet your yard expert Sunny and there's some prompts to get started here what can you tell me about

my yard send that and then it has some it has some information about this user's yard location your cool season grass is in your lawn size of 2700 ft plant hardiness Zone and a little bit more about your lawn and your soil test so this has been out for this fall like three months now and we are just putting together a very holistic evals I guess approach so we did topic modeling and there's probably

there's a lot of surface area here I think there's probably 40 distinct topics and it's yard guidance subscript questions or how to use Sunday product type questions and we're trying to use real user data as our eval questions or data set and I've gotten to the point I went through your blog post I had that open for three days you know in a tab on how to do the evals write CR the critiques and then what we're trying to do now is

use this to set up the llm as a judge and allow us to scale these evals and go a lot faster and this is where I'm having trouble this is what I've been doing so I I really AP follow blog so it made like a python shiny app and then this is input the output and I can write if this is a pass or fail the critique and then some any other issues or the era analysis

stuff and then what I get to is I will have I have something like this so I have my this is like a just a the topic area aerification discount shipping moving this was the input query this is the output from our llm app score pass or fail here's my written critique and then some of the the notes here are optional but I can't get these critiques to I guess work I can only get up to the

80% or so alignment between my critiques and using LM as a judge and I think my problem might be the prompting The Prompt that I'm using for this let me stop you there for one second yeah before you started this process was because 80% sounds like it kind of good okay that's nice to hear but I think okay if one and five are wrong that's wrongness about I worry about just letting that run in an automated way or

and okay what was it before you started what was the B what's the Baseline of oh probably like 50 or 60% okay this is a huge Improvement oh okay well that makes you feel better but it's also very inconsistent too sometimes it will pass and sometimes it will fail based on my prompt and my critiques that I have listed and I think that's because our questions are very open-ended it'll be like what what can I do for dandelion there's many possible

answers that would acceptable there and so it's hard to have them all articulated in the critique and I just have to keep watching it and keep adding and adding to the critique and is if that's what it is then that's what it takes no I mean yeah okay so you've already made like significant progress it sounds like you you have something like LM as a judge is not perfect and it's hard to sometimes get it to be really high

alignment especially if your service area is large yeah and if you have a lot of different open-ended questions it's very difficult and proxy in a way of as judge it's going to be really hard to in this situation it might be really hard

to get it you know the alignment like super high because he has like just insane surface area you can ask anything about your lawn yes and you're expecting it to give you a coherent answer about any

someone throws at it which is and you're doing like you're doing everything very you're doing it correctly you're even like segmenting it by topic I can see that and you're trying to analyze what's happening have you do you notice is there like errors in certain topics more than others or anything yes there is for example we do really good at answering like air questions about aerification because we have information about that in our knowledge based so it's retrieved and it

does really good job answering anything about that but we are having a hard time with like next shipment questions right now because we're in 2024 and the the subscriptions are mostly done so the next shipment is until spring 2025 so it's the way our shipping data is formatted is not easy to articulate to the llm so that's an example of yeah there are differences okay one thing you could do is do is your prompt so are you using us

is okay I think from talking to you a little bit before like you have some you're doing some rag in here with the examples that you're putting in the prompt are those static or are those like Dynamic based on the topic or what right now they are static that's another area of work we're undergoing right now is how to basically do topic classification and have different routes or workflows for different types of topics where we could

C be a little more custom with the prompts and break down the tasks to be more specific but right now it's all one way you could try it like on the things that like next shipment for example since it's like the highest has the highest error you could try having a classification that routes that or like that changes the prompt if detects that okay we're talking about next shipment and brings like that information that kind of information into the prompt that

okay so that I think is right that's just logically the next move that you would take on this I'm I'm trying to get the eval setup so I could run them quickly and not have me have to read every question and score it so before we make that change I wanted to be able to establish here's our Baseline we can run it quickly then go and and start to make some of those changes to our system and

be able to rerun EV vals and monitor progress does that sound right does that or does that not seem possible no it's a really good question It's tricky because you okay so the question is like how good is enough in eval and sometimes it it sounds like you're doing a lot already and you're hitting a plateau and I don't see you're doing more than 99% of people are doing like most times I have these office hours and

no one's even looked at the data yet so you're like Way Beyond like most people and so like you have to balance it because you can't just put you can't just put all your effort into evals either even though that's all I talk about all day that's not really what you want to do you have to see sometimes you already know from all this work what's wrong yeah it's likely like that yeah you can go into evils but like

it's supposed to be a proxy of and even though it has one in five or completely off and I know it's like inconsistent you can iterate it on over time but you should like go back and forth a little bit and make your product better go back TW your EV vals make your product but it's like a process you don't want to just stuck too much in evals all the time and that's a little bit counterintuitive just

because and see okay is it a good proxy or not if you try to improve next shipment does it actually can you see a difference in your evals maybe you do even though it's no maybe you see some signal from there okay that's what I would I can when I as I've gone through this I can see oh this is very obvious like things that we need to to change and make it better and I'm pretty

confident it will work but that's a hypothesis and it will be an experiment and I feel like in order to get the the engineering work done for some of these experiments I do have to articulate this is what the value of this would be so that's something we're going through with the group now too is just the idea of this is applied research iterative experimental and not just build and release features and expect a certain level of performance

yeah you could have special specialist judges as well but you have a lot of to topics yeah how many topics do you say there were well about 40 and can I ask to so I was using I was basically using a template like this for my L as a judge so this is from the Brain Trust Auto evals Library so I was using something like this and I also thought too oh should I be using more customized

prompts for the LM as a judge not just using the S like the same prompt for every question or every topic can you say that again so okay okay so your critiques should be shaping the like what the what is in the prompt like not exactly to some extent you want to see if so one trick I use and that I put in the blog posts I get my judge to critique before having a

judgment but then also if you want to use those critiques in other ways you want to see okay can you use that to come up with some kind of guideline that you can give the judge as well if there's something that you uncover in those critiques okayy you should be doing XYZ you should be telling the model okay you need to consider XYZ as a general rule when evaluating this output does that make sense yes and

I think that when I use this type of a prompt it's here's the input here's the output here's the criteria to submission meet the Criterion that this might be too it's called a close QA so I can say here's acceptable responses but if it happens to have a new response to the question it will grade it wrong or or fail when it could be acceptable you know what I mean what do you mean a new response to

the question when I run the evals I'll be using the same query but it might generate a new response that I haven't written a critique for yet like if it's like how to control dandelion might come up with some new way to say how to control dandel lines that is an acceptable response but I haven't seen it yet and I don't have a specific critique for this is good or this is bad okay so correct me if I'm misunderstanding but when you

WR you shouldn't have a critique for every possible thing the language model might say so that might be my problem because that's what that's what I'm doing yeah okay okay though because if you do that you can't like this imposs first of all I think it's impossible in a way because like how are you going to do that that's it's going to be hard and then you don't want to overfit the it's going to yeah I don't know it

sounds difficult to try to come up with every possible thing you could say you want to try to generalize it a little bit somehow like figure out like what are the general principles you want the judge to follow rather than trying to figure out like what are the every single like specific scenario that I can encounter does that make sense otherwise it might yeah if you go with the other approach it becomes a little bit more brittle I

think yeah that's and that's what's happening where it's like I said it's inconsistent or run after run I think the word that use over fifth might be that makes sense here now I think I might be doing that but then if I write it generally like here's some here's generally good or or what you need to say or what you need to not say then I'm like that's basically the main prompt of the actual product no for sure okay it

does share a lot like it can share a lot of the same same thing and what you want the judge to do is you want to have really good guidelines that you can you feel that you can apply in any situation it's hard because first of all it's really hard when you have something that can just do anything it's really hard to evaluate it that's just the overall it's it's very difficult but then you want to show examples of oh how

do how you applied those guidelines in the prompt you want to show okay because of XYZ this is a good example and why or because of ABC this is a example and why you have enough of those examples that are drawn from your critiques that you feel comfortable that it's showing it's giving the language model enough information to like approximate whatever is in your head yeah like it's like something you could follow or that

you could follow before you joined Sunday yeah yeah okay and I haven't been using like examples in the way of sort of question answer pair and I could probably do that a lot better I've just been putting here's critique here's all the critiques dumping it in yeah it's closely related like I think in my blog I showed like something like dumping the critique in but yeah really what you want to do is you want to show the model the

reasoning of why you want to get to a particular kind of judgment in yeah you want just want to teach the model model like just show like their General reasoning patterns and the critiques you want to try to see if you can do the more tricky ones okay this is how you apply this guideline these are like different ways to apply this guideline or whatever you should find that you are like editing your guidelines as well while you're doing

this because yeah usually you'll find something that is not working okay good related to something you said earlier of if this if we're doing okay if we're having like 80% alignment is that I guess I'm trying to get away from having to read question not look at the data but like having to read and score questions I have been spending daily time doing this and that's just not scalable yeah I guess how good could it

get with LM as a judge maybe I have an unrealistic expectation that we could oh it will just be automated and we can just hit a button run vows and we can know if we're making progress or not yeah I don't think you can ever get completely away from looking at data because at the end of the day you have to trust it and there is no other way to trust it other than checking it yeah

there's just no there's no way out of that doesn't mean you have to look at every single thing right you know you can sample but you still have like someone has to look at it now the things that now the thing that you're doing about segmenting into topics that's where it's going to be more in most interesting because you will be able to see you know in this area there's like high alignment in this other area there's not that much

agreement between me and the judge it's like kind of I can't really rely on it here so like you might find it's very uneven and you have to account for that potentially in your evals and also in your review process you might have to say okay you know what in this area where we can't quite automate myself I might have to look at it more okay and you probably will figure out how to automate yourself more in that situation it's

hard like a lot of times it might make sense to try to focus these different judges but it's a little bit of cat in the mouse game you don't want to over engineer your evals either because there a lot of things you're going to find when you're doing this the real magic of this entire process is that you're looking at it and you're thinking about what's wrong yeah and then if you find something obvious that's wrong just go fix it that's what

I would do because it's it's a little bit circular in a way okay then I'm I'm thinking about like scalability if we have thousands of conversations happening it like in the springtime we get a lot of of traffic that's a very important time of the year for us we're trying to get ready for that so if there's thousands of conversations happening you can't possibly read them all how do you ensure quality what do you what how do you

define quality like getting the questions right like I'm doing with the evals now yeah there's different ways you can think about it one is you can so you can okay one thing you can do is for things that have high alignment have you can rely on that a little bit more you can sample in areas of low alignment heavier to like spot check like what is happening there's other Advanced methods

you can use like you can try to you can once you figure out like what is going wrong you can decide hey you want specialized tests that go beyond just like General judges that like evaluate things like rag or evaluate factual correctness or hallucinations you want to do that kind of based on some error not just like for the hell of it okay because it can become

distracting but what you want to do is you also want to know if you can trust that too so it has to be grounded in some real error so like when you roll this out and you have all these thousands of cars like whatever in the spring and the summer have some topic areas like or many topic areas that have high enough alignment where you can use it as a proxy but then other ones like yeah you have to look at it you can

also see if you can another thing is you can see if get a user feedback yeah I don't know how much like what the dynamic is like some people don't trust their users we do we have this the thumbs up thumbs down thing that's often recommended what how often are people providing feedback is it I want to say 8 to 10% of messages something like that yeah what is the

actual sort of is there a business metric that's being tracked do you know if people that use this these features what the impact is on their purchasing and stuff it's still quite new but I would say the goal is engagement get people to have conversations and tell us more about them and then it will make product recommendations but that's one off product at a time it will probably be a long a longer term metric for retention and reducing

churn yeah the real advice that I would say here is you might not want to hear this but it's going to tell you is like this product it's like a little bit the surface area is like way too large if it was okay if you were my client and I would like I was like Consulting with you I would say look it's going to be really hard to evaluate this it's going to be just like all over the place let's

pick like a few topics and have something that specializes on that and really understand that and get that to work well before trying to solve everything okay this is great no because it's actually very is very difficult like it's very difficult to just have something that can you can ask anything and evaluate that like it's it's quite hard like it's you're trying to essentially have yeah is is is usually

not the best idea everyone tries to do it but it's you want to start somewhere else usually okay I didn't think like I said there's 40 40 things people ask I didn't think that was too much necessarily no it's good to be I don't want to dissuade you but I guess what I'm saying is like even the 40 topics like 40 is a lot of different it would be good to focus on something and say hey we nailed it like

this is like hey what seed like the problem I have for example what seed do you need I don't know I'm just like reading descriptions and I'm just like buying different ones I have no idea I'm just like let me just put it on my grass and see what happens I don't know whatever like some like target areas like it doesn't have to be 40 maybe it's five things that's my intuition anyways okay is to like start like scoped and like

really make it good okay this is great okay yeah so another kind of related to this the large surface area one of the questions I I had for you was the topics change over the time of the year the questions people are asking in Fall are a lot about is it too late to do this there's Frost can I do this or that that's going to be very different from what they ask in the springtime when the weather is different and I in

the past few weeks I've figured out I was concerned that oh I we don't know what people are going to ask at different times of the year because we haven't experienced it yet or had this out I actually think I can probably do some type of synthetic data generation about that it's they're going to ask different questions in the fall about seating timing it being too late or what about Frost but they're going to ask the

same they're still going to ask about seating timing in the spring it might just be is it too early or is it too hot but it's all just seating timing and I think in the last few weeks I think I've realized I can probably address that with some synthetic data generation what's the distribution of of like topics is it like take a specific season let's say whatever spring yeah are the is it like two or three questions that

are driving like 80% of the volume you think or I would say five topics five or six topics driving most of the conversation but they're very different it's like this fall it was seating timing when is the renal when does the plan start or what am I getting next year those are like the top two some more subscription questions and maybe a little bit about weeds but yeah maybe you could focus on those and say and

have a compromise in a way where you're like hey most of traffic is here we're going to focus on this make it really good make the element as a judge as good as it can get make it work as good as it can be yes there's the other things but in order to drive progress like we're focusing on the things that people ask most about so that we can learn and then as we go along we'll refine this tail the tail

that way like it'll be more tractable cuz 40 is a lot okay and it just yeah it'll just help also to make progress on makes something really good yeah okay great I don't know if you have the ability to do that or not if you can say hey we're focusing on these things we can or I I feel like I could but how do we articulate that to the user it would just be like this only

answers questions about seating now and if you ask about this we'll just say oh I can't answer that is that what you do you could do that you could ALS also just leave it and say and just know hey this is something that you tolerating as a business to say hey these other questions they're not going to be as polished as these ones that people ask about a lot that's okay we're learning and that's okay it's just that just

internally you're like focusing on those buckets so that you can refine those that's all okay this is great I can't wait to send this conversation to my colleagues yeah that it was really nice to connect with you I don't buy the products really anymore because I have a I have a landscaper now so I'm not even doing it myself okay okay if you ever have any lawn garden questions please let me

know we'll do yeah okay cool I'll send you the I I'll send you the recording it's recorded and then I'll just email to you yeah you want to send it to other folks okay great all right sounds all right thank you yep see you all right thanks