AI Eng vs. PMs: Who Should Drive Evals?

3 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=FHDOICTRGY8
https://www.youtube.com/watch?v=FHDoiCtRGY8

SUMMARY

The discussion emphasizes the collaborative roles of Product Managers (PMs) and AI Engineers in labeling and evaluating AI outputs. It highlights the importance of domain expertise in both labeling and evaluation processes while addressing the need for clear communication and decision-making protocols when discrepancies arise between automated evaluations and human feedback.

- *PMs should be involved in writing labels and evaluation prompts due to their domain expertise.*
- *Collaboration between PMs and AI Engineers is essential for effective labeling and evaluation.*
- *The quantity and quality of labels should be determined collaboratively, focusing on statistical significance.*
- *Disagreements between automated evaluations and human feedback need clear protocols for resolution.*
- *Understanding the roles and responsibilities in labeling and evaluation can vary by company.*
- *It's crucial to establish who acts as a tiebreaker in case of conflicting evaluations.*

and say like, you know, as much as like the PM is is the right person to write the label, they're also probably the right person to kind of help write the eval here to some degree. And so I think that there's like this this good kind of healthy tension of like maybe it's not as mutually exclusive and it's a little bit more collaborative in terms of a workflow where, you know, the PM takes on some of the labels and then the the AING kind of can help iterate on those.

I would like I would like to see the PM. I think the closer you can get to labeling to the domain expert, the better. And usually the PM is more of a domain expert than anyone else if things are going correctly and organized correctly. Right. That's that's exactly right. And then and then at the same time if the AI engineer is the person accountable for the end product delivery then the eval prompt and the final prompt that goes into the system also like kind of rests on the AI engineer but it also is a little bit of like some domain expertise should enter that as well. So like you should see you know a little bit of that like back and forth.

I think in a typical workflow we see this a lot. We see human we see PMS writing labels. We also see PMs writing eval prompts and writing the end prompts quite often. Okay. What I want to kind of where this kind of leads me is like I'm not sure if we're going to get a packaged up answer of who writes the labels and who writes the prompts necessarily. Though I do think that like leaning towards PM should write labels is probably a fair assessment. Like the PM needs to be in the details of the data. the PM needs to be in the details of the eval but kind of who does what and what that split looks like. Your mileage may vary at your company. What I would urge people to do is think about these three questions which is when you go back and you leave this course and you move on to like writing and building AI products. I kind of ask people how many labels do you think are good enough and for each example how many labels should you have? And so this is a really important question

because like a lot of times people come to us and say like do we do you know is 10 labels enough 100 labels a thousand labels and I think that this is a question that PM should be working with their AI engineers on which is what is statistically significant for you to know that this eval is good enough and what you know does that matter per row should you have agreement between all of your subject matter experts or are you okay with disagreement in your human labels? So think of this as like what's the quality and quantity of the human labels that we have. The second is what happens when the eval says that the output is good but the human label disagrees and what happens during this tie break scenario.

And you'll kind of see this when you actually get your your your product into production. It's going to happen a lot where your eval system is going to say this went well, but your human in the loop or your feedback from your end user is going to say this was a bad response. So, who takes a look at those examples and what happens and who's the tiebreaker there? That's a that's a really important question that comes down to like your