(no title)

1 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=UYFDFFAO4D4
https://www.youtube.com/watch?v=UyFDfFao4d4

SUMMARY

Guardrails in AI products serve as protective measures to prevent undesirable outputs, such as irrelevant responses or inappropriate content. The framework discussed, Guardrails AI, offers a variety of guardrails tailored to different risk categories, emphasizing the importance of ensuring AI responses are relevant and appropriate.

- *Guardrails protect AI from generating harmful or off-topic content.*
- *The Guardrails AI framework includes a Guardrails Hub with various risk categories.*
- *Examples of guardrails include checks for brand risk and question-answer relevance.*
- *A specific guardrail example is a validator that assesses the relevance of AI responses to user prompts.*
- *Implementing guardrails often involves using existing prompts, which may not always be tailored to specific applications.*
- *Off-the-shelf prompts may lack the specificity needed for optimal performance in unique products.*
- *Guardrails are not foolproof and may not catch all errors in AI outputs.*
- *Additional resources for learning about guardrails are available in the comments.*

If you're building AI products, you may have heard of the term guardrails. So, what is a guardrail? Guardrail is something that is meant to protect you from something bad your AI might be doing, such as talking about competitors, or going off-topic, or using profanity. And it's worth knowing what exactly a guardrail is, and how does it work? I'm going to share my screen here, and what we have here is a very popular framework, Guardrails AI, and they have a Guardrails Hub. And this is a very popular way to use guardrails in your application. And so, if we go here, you'll see a whole bunch of guardrails, all kinds of risk categories. I'll click on brand risk, and you'll see all kinds of things guardrails for that, like competitor check, detecting gibberish. I'm going to go ahead and select QA relevance, just because question answer relevance is important in a lot of applications. And so, what is this guardrail? This is a validator that checks whether an answer that your AI is giving is relevant to the question that is asked by the user.

[music] And say, what does this mean? How does it work? So, let's go ahead and look at the code. Click on validator, and then I'm going to click on main. And this is the code for that QA relevance guardrail. If I scroll all the way down here, so this guardrail is this prompt, which says, "Is the above response relevant to the following prompt?" So, this guardrail is just someone else's prompt. That's all it is. You could use something exactly like this by just writing this prompt yourself. In fact, if you were to use this guardrail, this is about five levels of abstraction just to get this one sentence prompt. So, the thing to know about guardrails is all you're doing is using someone else's prompt. And you have to ask yourself, is that really going to do what you want? A lot of times, that off-the-shelf prompt is not specific enough to your product to do a good job. In this case, this is just a very simple sentence that says, "Hey, is the response relevant?" So, it's worth knowing what a guardrail is, and it's not a 100% guaranteed solution to catching errors. If you want to learn more things like this, I left some

resources in the comments.