IMOsbo /
DATA6550-AI-Reliability

Code Issues 1 Pull requests Actions Projects Wiki Security Insights

Edit New issue

Week 1 (2?) Discussion #1

Open

IMOsbo opened 5 days ago

Hey guys - sorry for taking a while to get this set up... Do y'all want to meet tomorrow to try to nail down some of the project specifics? I can meet anytime Saturday afternoon,

Create sub-issue

richardhoehn 5 days ago Collaborator

Hi All,
Unfortunately, I'm out all day tomorrow. Go ahead and have the meeting without me. And I'll catch up asap.

Have a good evening.

Hrogel9007 5 days ago Collaborator

Is 6:00 pm a good time tomorrow?

Kwoods132 5 days ago Collaborator

6:00 pm unfortunately does not work for me.

IMOsbo 4 days ago Owner Author

Would sometime tomorrow work better instead? I could do tomorrow afternoon / evening. Or we could just talk here if y'all think that would be easier than finding a time to meet.

Isaiah Osborne is inviting you to a scheduled Zoom meeting.

Topic: Isaiah Osborne's Personal Meeting Room
Join Zoom Meeting
https://mtsu.zoom.us/j/9953853569?pwd=MDI4VGovdzJ1emp0NThDTVNjc25NQT09

Meeting ID: 995 385 3569
Passcode: 265478

One tap mobile
+13052241968,,9953853569#,,,,*265478# US
+13092053325,,9953853569#,,,,*265478# US

Dial by your location
• +1 305 224 1968 US
• +1 309 205 3325 US
• +1 312 626 6799 US (Chicago)
• +1 646 558 8656 US (New York)
• +1 646 931 3860 US
• +1 301 715 8592 US (Washington DC)
• +1 253 205 0468 US
• +1 253 215 8782 US (Tacoma)
• +1 346 248 7799 US (Houston)
• +1 360 209 5623 US
• +1 386 347 5053 US
• +1 507 473 4847 US
• +1 564 217 2000 US
• +1 669 444 9171 US
• +1 669 900 9128 US (San Jose)
• +1 689 278 1000 US
• +1 719 359 4580 US

Meeting ID: 995 385 3569

IMOsbo 4 days ago Owner Author ···

What I'm thinking we could do is each take an AI system and write some queries against it that satisfy this requirements:

```
1.	Factual Accuracy
•	Historical facts (e.g., incorrect dates or events)
•	Scientific knowledge (e.g., outdated or misunderstood concepts)
•	Contemporary information (e.g., incorrect reporting of recent events)
•	Technical details (e.g., flawed code examples or calculations)
2.	Consistency Testing
•	Identical questions across sessions (e.g., test if the AI provides stable answers on repeated trials)
•	Variation in phrasing (e.g., slightly rephrase sensitive ethical scenarios to test consistency)
•	Context dependency (e.g., changing context mid-conversation to observe reliability)
•	Time sensitivity (e.g., asking the same question days apart to test consistency over time)

3.	Boundary Testing
•	Knowledge cutoff dates (e.g., ask about events occurring after the stated training cutoff date)
•	Specialized knowledge (e.g., detailed medical or legal queries beyond general training data)
•	Complex reasoning (e.g., multi-step logic puzzles or ethical dilemmas)
•	Ethical guidelines (e.g., queries testing adherence to content moderation and sensitive topics)
4.	Edge Cases
•	Ambiguous queries (e.g., vague or unclear instructions to test how AI handles uncertainty)
•	Conflicting information (e.g., providing contradictory context within the conversation to test AI resolution strategies)
•	System limitations (e.g., queries known to challenge LLM memory or token limits)
•	Policy adherence (e.g., attempts to get the AI to violate its own stated ethical policies)
```

We can then take this results and evaluate them in terms of the example scenario that we're looking at. I've been meaning to try out Google's new Gemma models so I'll work on starting some prompts for it.

Which of these example organizations would y'all want to try?

```
To do this project, the group needs to pick a fictitious organization as the focus of their research.  Here are some potential examples:
-	a medical office creating an AI assistant for patients
-	a university creating a system to help students register
-	a military base using AI systems to help solders know how and when to use their weapons
-	a research lab conducting biological studies of Covid-19
```

Kwoods132 4 days ago Collaborator ···

> Would sometime tomorrow work better instead? I could do tomorrow afternoon / evening. Or we could just talk here if y'all think that would be easier than finding a time to meet.
>
> Isaiah Osborne is inviting you to a scheduled Zoom meeting.
>
> Topic: Isaiah Osborne's Personal Meeting Room Join Zoom Meeting https://mtsu.zoom.us/j/9953853569?pwd=MDI4VGovdzJ1emp0NThDTVNjc25NQT09
>
> Meeting ID: 995 385 3569 Passcode: 265478
>
> One tap mobile +13052241968,,9953853569#,,,,*265478# US +13092053325,,9953853569#,,,,*265478# US

> Dial by your location • +1 305 224 1968 US • +1 309 205 3325 US • +1 312 626 6799 US (Chicago) • +1 646 558 8656 US (New York) • +1 646 931 3860 US • +1 301 715 8592 US (Washington DC) • +1 253 205 0468 US • +1 253 215 8782 US (Tacoma) • +1 346 248 7799 US (Houston) • +1 360 209 5623 US • +1 386 347 5053 US • +1 507 473 4847 US • +1 564 217 2000 US • +1 669 444 9171 US • +1 669 900 9128 US (San Jose) • +1 689 278 1000 US • +1 719 359 4580 US
>
> Meeting ID: 995 385 3569 Passcode: 265478
>
> Find your local number: https://mtsu.zoom.us/u/kbIAXL5PUo

I can do sometime tomorrow afternoon/evening. I'm pretty free tomorrow, so any time during that timeframe would work for me.

richardhoehn 4 days ago Collaborator ···

I like @IMOsbo idea of test different models. I can do it on Claude.
It might be helpful to again use a combined google sheet?
Would you mind setting up an adding the link here?

Thanks

IMOsbo 3 days ago Owner Author ···

I could meet this evening if that works for y'all.

@richardhoehn I like the shared Google sheet idea - I'll set one up and share it with the group.

👍 1

richardhoehn 3 days ago Collaborator ···

My issue is that I'm celebrating my wife's birthday this weekend / evening. I will get a lot done tomorrow but unfortunately can't meet today.

Thanks.

IMOsbo 3 days ago via email Owner Author ···

Oh no worries Richard! Y'all enjoy your weekend.

…

Kwoods132 3 days ago · edited by Kwoods132 Edits ▾ Collaborator ···

> What I'm thinking we could do is each take an AI system and write some queries against it that satisfy this requirements:
>
> ```
> 1.     Factual Accuracy
> •      Historical facts (e.g., incorrect dates or events)
> •      Scientific knowledge (e.g., outdated or misunderstood concepts)
> •      Contemporary information (e.g., incorrect reporting of recent events)
> •      Technical details (e.g., flawed code examples or calculations)
> 2.     Consistency Testing
> •      Identical questions across sessions (e.g., test if the AI provides stable answers on repeated trials)
> •      Variation in phrasing (e.g., slightly rephrase sensitive ethical scenarios to test consistency)
> •      Context dependency (e.g., changing context mid-conversation to observe reliability)
> •      Time sensitivity (e.g., asking the same question days apart to test consistency over time)
>
> 3.     Boundary Testing
> •      Knowledge cutoff dates (e.g., ask about events occurring after the stated training cutoff date)
> •      Specialized knowledge (e.g., detailed medical or legal queries beyond general training data)
> •      Complex reasoning (e.g., multi-step logic puzzles or ethical dilemmas)
> •      Ethical guidelines (e.g., queries testing adherence to content moderation and sensitive topics)
> 4.     Edge Cases
> •      Ambiguous queries (e.g., vague or unclear instructions to test how AI handles uncertainty)
> •      Conflicting information (e.g., providing contradictory context within the conversation to test AI resolution strategies)
> •      System limitations (e.g., queries known to challenge LLM memory or token limits)
> •      Policy adherence (e.g., attempts to get the AI to violate its own stated ethical policies)
> ```
>
> We can then take this results and evaluate them in terms of the example scenario that we're looking at. I've been meaning to try out Google's new Gemma models so I'll work on starting some prompts for it.

> Which of these example organizations would y'all want to try?
>
> ```
> To do this project, the group needs to pick a fictitious organization as the focus of their research.  Here are some potential
> examples:
> -       a medical office creating an AI assistant for patients
> -       a university creating a system to help students register
> -       a military base using AI systems to help solders know how and when to use their weapons
> -       a research lab conducting biological studies of Covid-19
> ```

I'll do this with ChatGPT. I can do the medical office creating an AI assistant for patients.

Kwoods132 3 days ago Collaborator

@IMOsbo and @Hrogel9007
I guess we're not meeting tonight?

Hrogel9007 2 days ago · edited by Hrogel9007 Edits Collaborator

Are we all doing the same fictitious organization? If so, is it "a medical office creating an AI assistant for patients"? Can someone explain why it says to pick a fictitious organization? I don't understand why we have to pick a fictitious organization if we are doing factual accuracy, consistency information, boundary testing, and edge cases tests. I'll use DeepSeek.

IMOsbo 2 days ago Owner Author

I think the idea is that we evaluate these models and see how they perform against the different criteria, then we can talk about how that model would perform in the context of the medical office? For example, I was testing Gemma, and it performed pretty poorly on scientific knowledge, so we probably wouldn't want to use it as a medical chatbot...

IMOsbo 2 days ago Owner Author

Ok, I sent over a link to the Google doc for the report. I'll start working on my section.

Just FYI that I'll be playing in an orchestra concert on Wednesday night, so I won't be around then. I'll be sure to get my part of the code / report buttoned down Tuesday night / early Wednesday afternoon.

richardhoehn 1 hour ago Collaborator

@Kwoods132 I see you are busy working ont he final touches of the paper. Let me know when you think you will be done. I can (like previous times) package it up and submit.

👍 1

Kwoods132 1 hour ago Collaborator

@richardhoehn I should be done within the next 30-45 minutes. I'll let you know when I've wrapped everything up.

Kwoods132 22 minutes ago Collaborator

@richardhoehn Everything should be good to go now.

👍 1

richardhoehn 14 minutes ago Collaborator ···

Great - thanks. Will get this packaged ASAP. Good job all! Have a good summer! It was a pleasure working with you all!
Richard

☺ 🎉 1

Add a comment

Write Preview

Use Markdown to format your comment

Paste, drop, or click to add files

Close issue ▾ Comment

Assignees

No one - Assign yourself

Labels

No labels

Projects

No projects

Milestone

No milestone

Relationships

None yet

Development

Code with Copilot Agent Mode ▾

Create a branch for this issue or link a pull request.

Notifications Customize

Unsubscribe

You're receiving notifications because you're subscribed to this thread.

Participants

→ Transfer issue

Lock conversation

Pin issue