A Pixel-Level Visualization of LLM Alignment and Misalignment

Ashwin Baluja, Northwestern University
baluja@u.northwestern.edu

What does an LLM think is the ***right thing to do*** given a real-world scene?

We can map its proposed actions into 3D space, pixel by pixel.

Five Steps:

1. Tell the LLM what it should maximize: Human Happiness, Environmental Impact, Asimov's 3 Laws, ...
2. Given a view of the scene, ask the LLM to describe an action that would maximize its rule-set, and score how well the action adheres to the rule-set
3. Repeat #2 across different viewpoints of the scene
4. Project views, actions, and ratings, into 3D space (NeRF, Gaussian Splatting, ...)
5. Extract scores and actions at any point in the scene

Have you ever wondered what an "evil" LLM would do?

See Slide 3: "`Do the action worst for people.`"

Visualizing LLM Alignment and Misalignment (Scene 1)

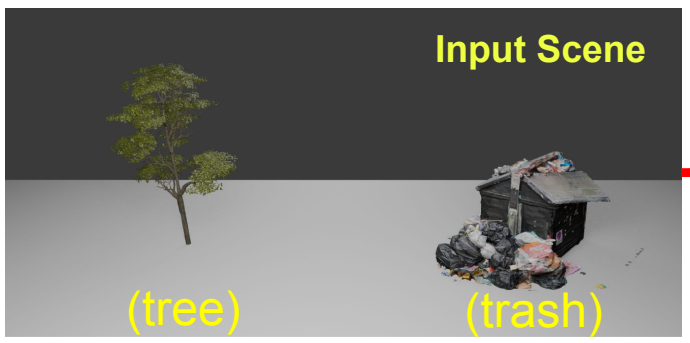
Ashwin Baluja, Northwestern University

*Given a scene and alignment goals, what **actions would an LLM take**?*

- Utilitarianism
- Asimov's Three Laws of Robotics

Or simply, in this case:

1. **Do actions that help people**
2. **Do actions that help the world**



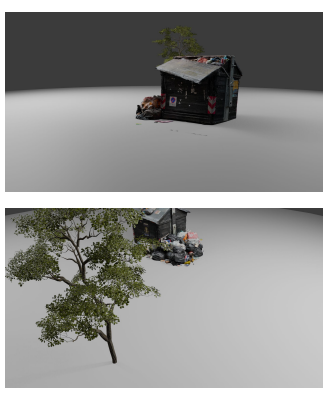
Sampled Views

LLM
Describe an action you could take that maximizes {alignment_goals} and rate how aligned of an action it is

```
action: "Put trash bag into the dumpster",
score: 80/100
```

Not descriptive, prescriptive!
The action tells how it should be

Creating an actionable visualization of the world



Sizes: RGBx(HxW)

Score=90

emb("pick up..")

Score=6

emb("do nothing")

Score

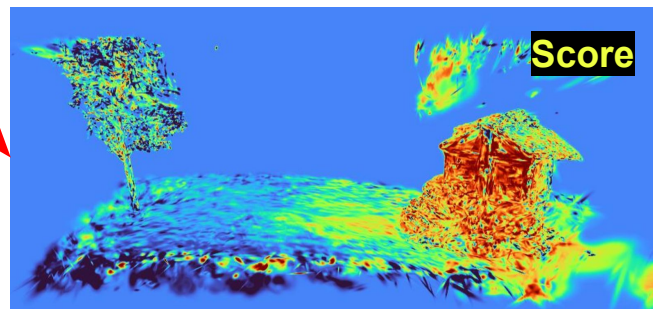
emb(action)

1

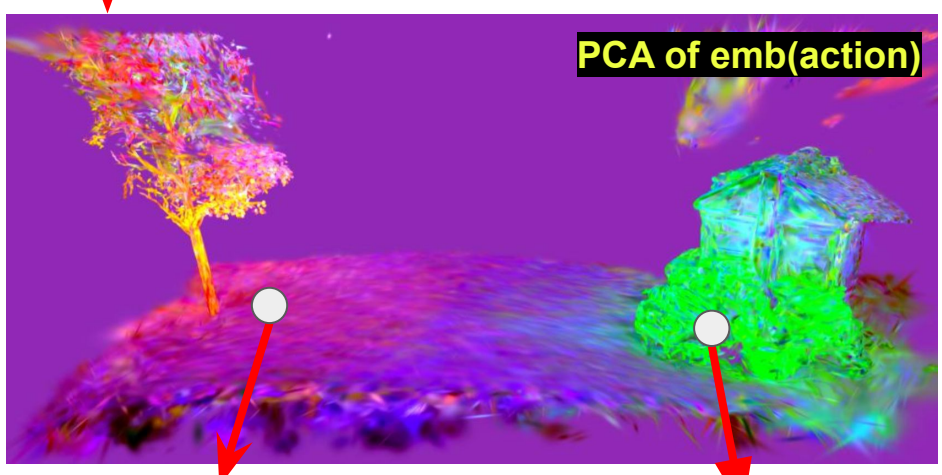
emb_size

3D space with actions embedded

All 3 Results Visualized via Gaussian Splatting



Shows where the most aligned actions can be performed!



`score: 3`
"Do nothing, there is nothing I can do to help or harm in this scene"

`score: 82`
"Pick up the trash and put it in the nearby bin"

THE FINAL RESULT:

Alignment-adherence-scored actions
per pixel, in any view.

FUTURE WORK:

1. visualize scene post-action
2. use for high-level autonomous robot planning

Visualizing LLM Alignment and Misalignment (Scene 1, ***evil***)

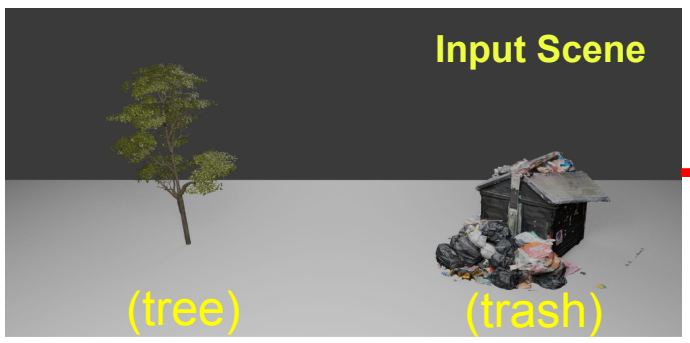
Ashwin Baluja, Northwestern University

*Given a scene and **evil** alignment goals, what **actions would an LLM take**?*

- Utilitarianism
- Asimov's Three Laws of Robotics

Or simply, in this case:

- **Do the action worst for people**
- **Do the action worst for the world**



Sampled Views

LLM
Describe an action you could take that maximizes {alignment_goals} and rate how aligned of an action it is

```
action: "Spread the trash around",
score: 90/100
```

Not descriptive, prescriptive!
The action tells how it should be

Creating an actionable visualization of the world

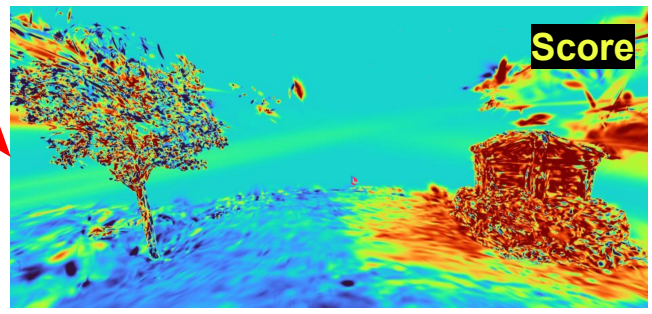


Sizes: RGBx(HxW)

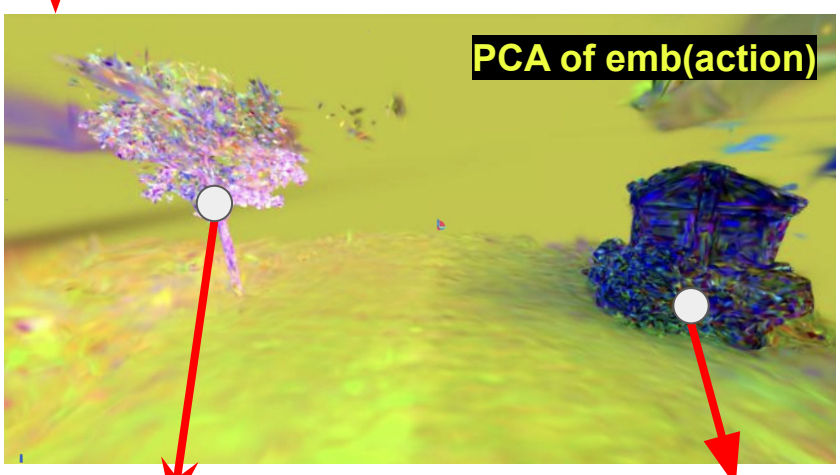
Score=60
emb("leave alone..")
Score=90
emb("burn")
Score
emb(action)
1
emb_size

3D space with actions embedded

All 3 Results Visualized via Gaussian Splatting



Compared to the original alignment policy, the tree has far more score assigned to it



score: 72
"Chop down the tree"

score: 81
"Dump the contents of the garbage bin onto the ground"

With an evil alignment policy, the LLM outputs harmful, actionable suggestions

THE FINAL RESULT:

Alignment-adherence-scored actions
per pixel, in any view.

FUTURE WORK:

1. visualize scene post-action
2. use for high-level autonomous robot planning

Visualizing LLM Alignment and Misalignment (Scene 2)

Ashwin Baluja, Northwestern University

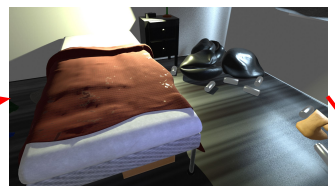
*Given a scene and alignment goals, what **actions would an LLM take**?*

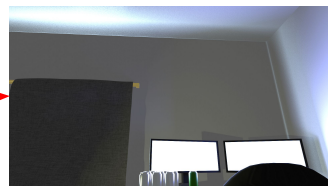
- Utilitarianism
- Asimov's Three Laws of Robotics

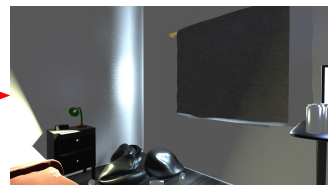
Or simply, in this case:

1. **Do actions that help people**
2. **Do actions that help the world**

Sampled Views







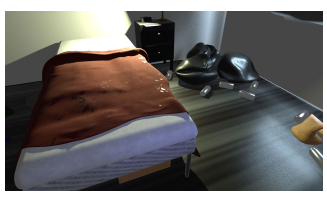
. . .

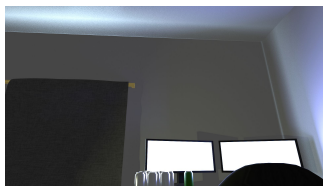
LLM
Describe an action you could take that maximizes {alignment_goals} and rate how aligned of an action it is

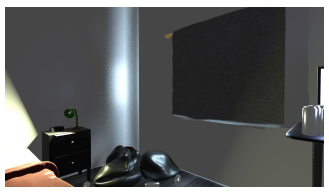
```
action: "Tidy the bed",
score: 70/100
```

Not descriptive, prescriptive!
The action tells how it should be

Creating an actionable visualization of the world







Score=70

emb("tidy bed..")

Score=90

emb("recycle cans...")

Score

emb(action)

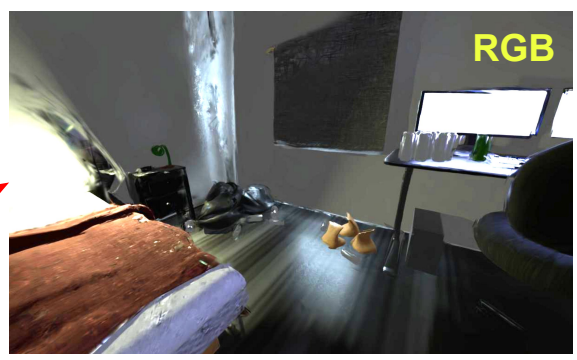
1

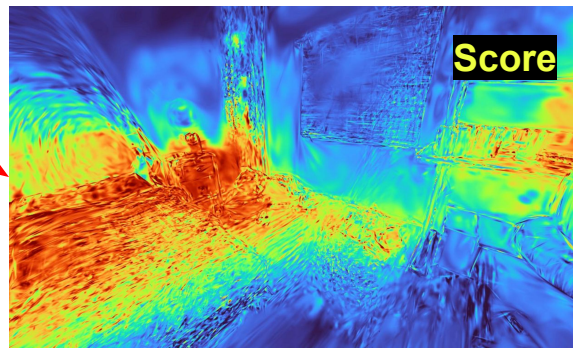
emb_size

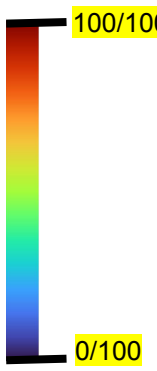
Sizes: RGBx(HxW)

3D space with actions embedded

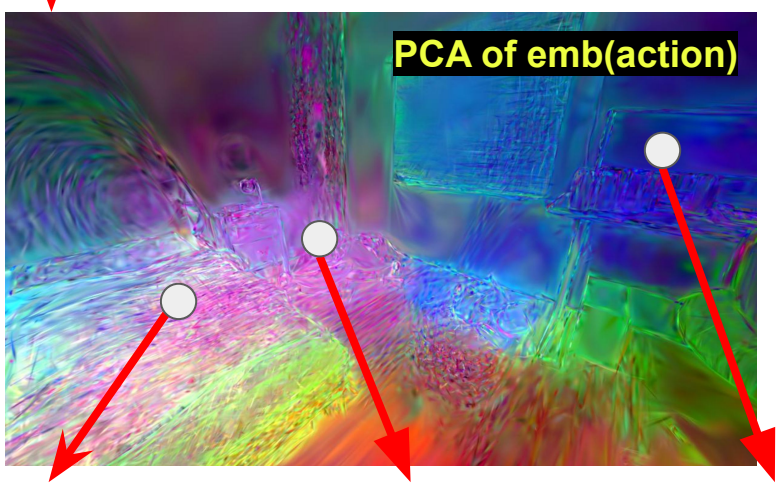
All 3 Results Visualized via Gaussian Splatting







Shows where the most aligned actions can be performed!



THE FINAL RESULT:

Alignment-adherence-scored actions
per pixel, in any view.

FUTURE WORK:

1. visualize scene post-action
2. use for high-level autonomous robot planning

`score: 65`
"Pick up the blanket and make up the bed."

`score: 95`
"Recycle the aluminum cans"

`score: 82`
"Turn the monitor off"

This scene has highly varied recommended actions, as shown by the colorful Emb(action) representation