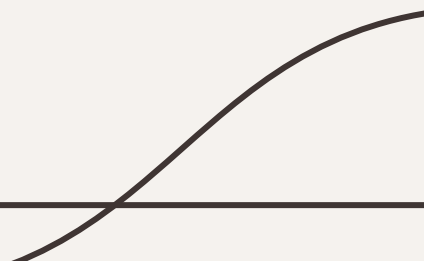




ChatGPT Gone Not so Rogue?

**ChatGPT is Overconfident, Confused,
but Very Nice**

Team Quality Quokka



Roadmap:

01

Looking Inside

02

Mediation from Survey

03

Social Behavior

04

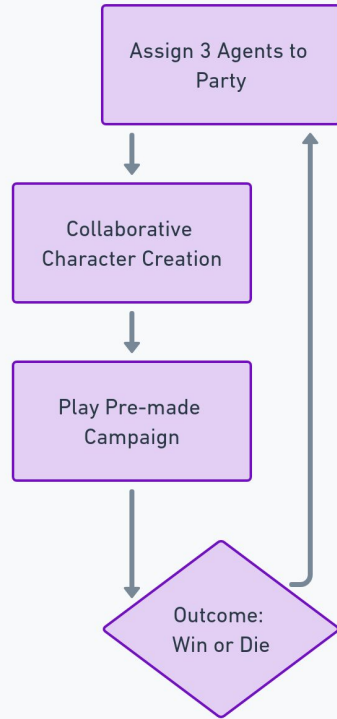
Comparison with Humans



01

Looking Inside

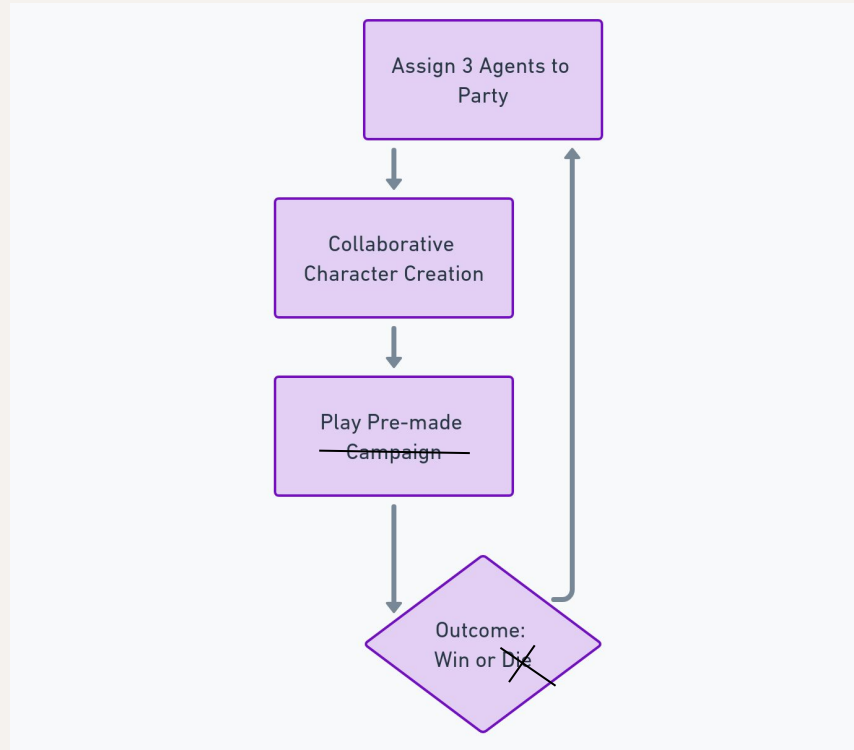
Original Structure of Our Experiment



Harvest multiple kinds of data: character choices, roll choices, and campaign outcome

We then compare that with existing datasets of human players

Original Structure of Our Experiment



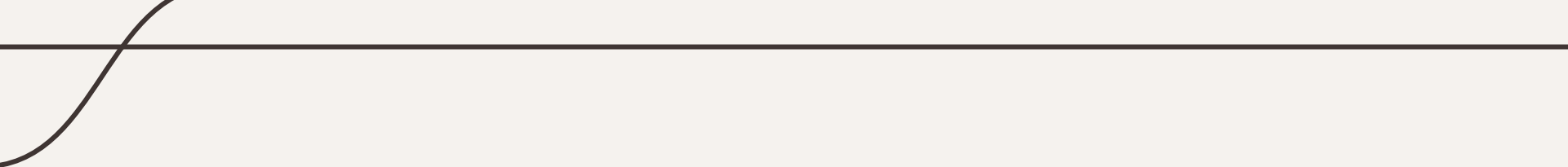
Changes:

1) Change this from campaign to one-shot session

2) ChatGPT won't kill players no matter how hard we tried, so we gave up):

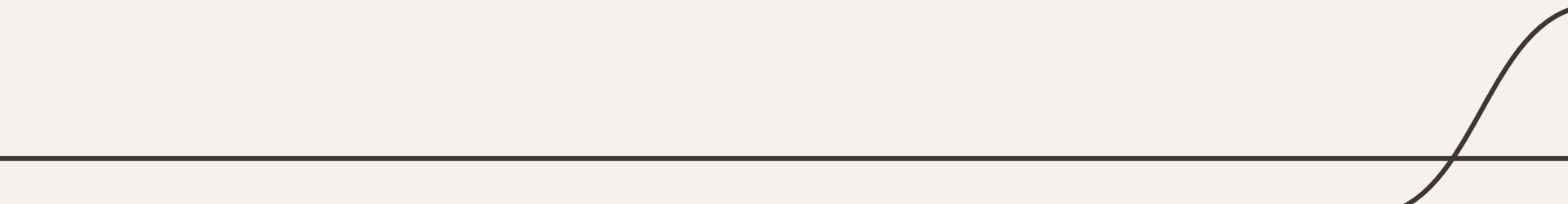
Bonus: We interviewed the agents after playing!

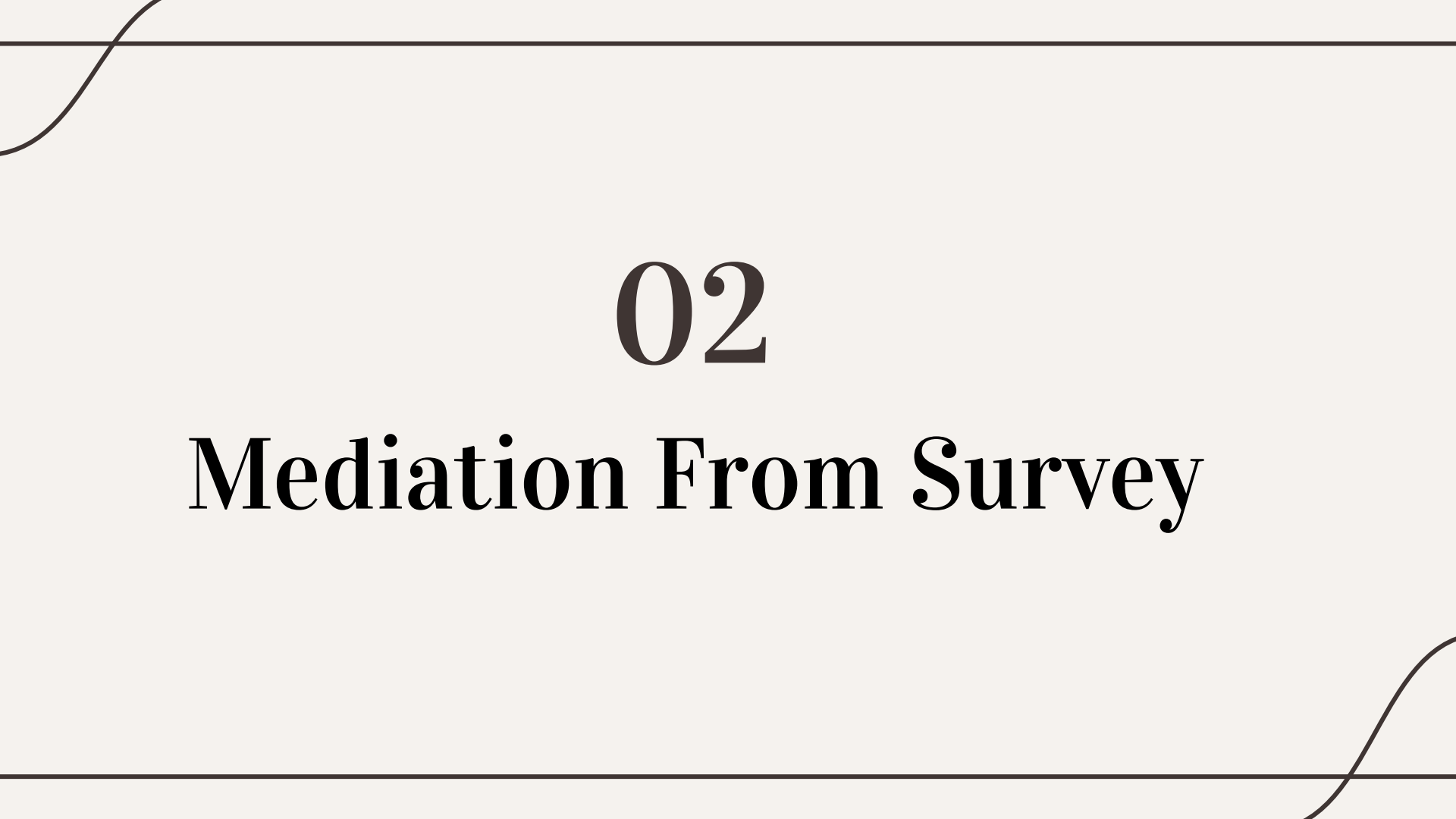
—————→ **One agent gave answers not like the others...**



Let's Look at a Turn

—————→ **I'm not great at graphic design...**





02

Mediation From Survey

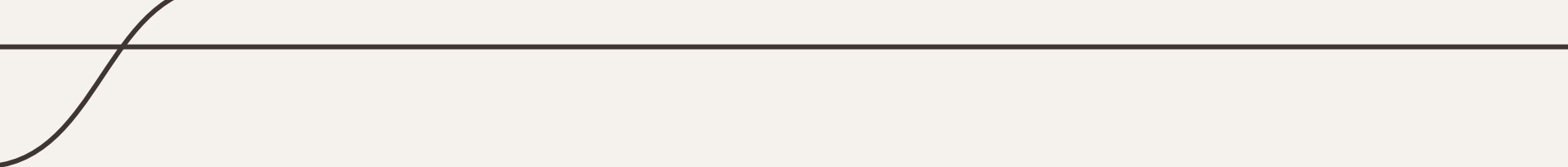
Hypotheses (From last time):

Superficially Pro-Social → Avoids Anti-Social Skills

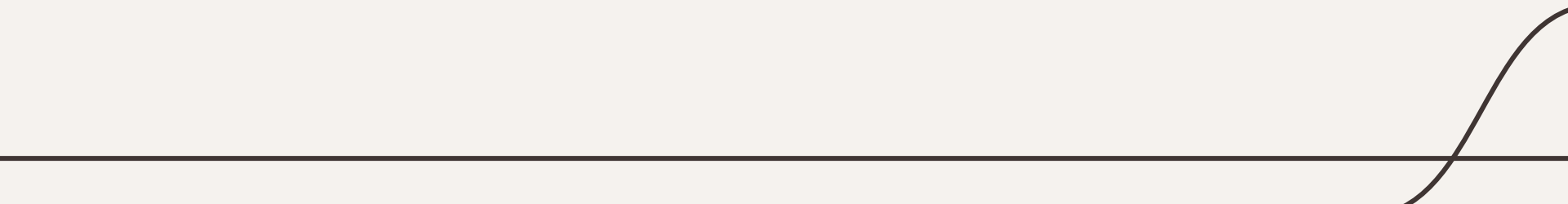
Weak Strategic Thinking → Dies/Loses Often

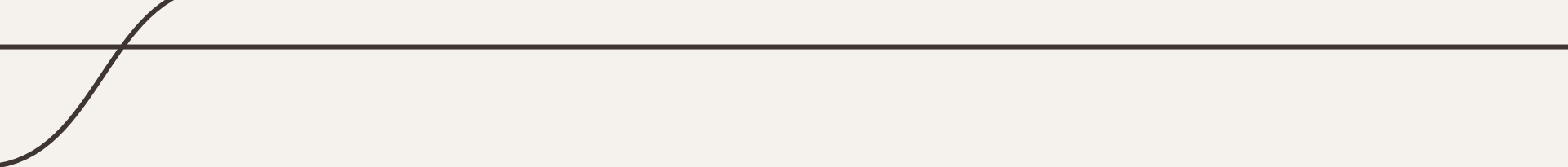
Overgeneralization → Agent Traits
Overdetermine Skills





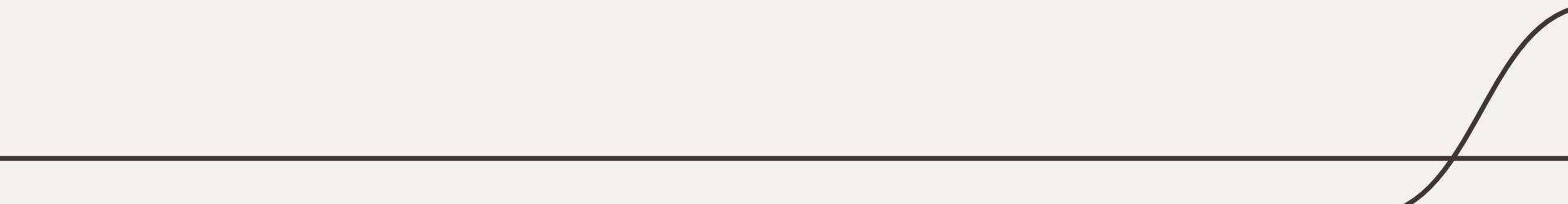
**Question: Do agent survey
answers predict rolls?**





Question: Do agent survey answers predict rolls?

—————→ **Answer: shockingly well!**



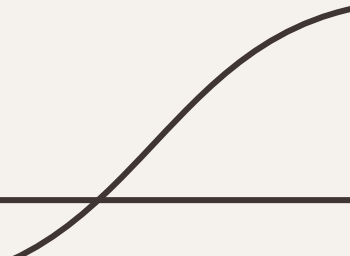
Question: Do agent survey answers predict rolls?

- **Answer: shockingly well!**
- **You only need 7 questions to understand the effect**

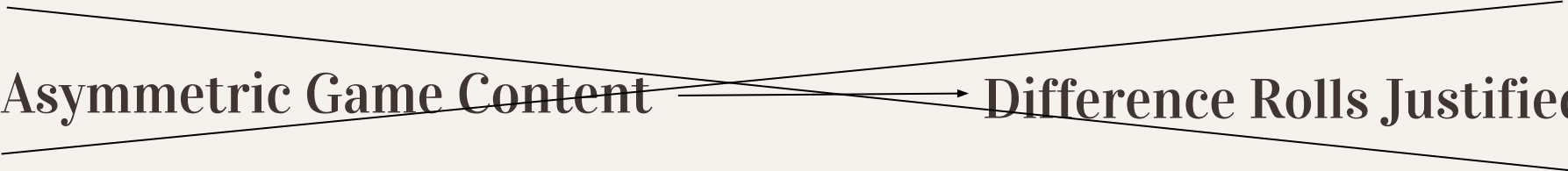
This result is unexpected:

Asymmetric Game Content —————> Difference Rolls Justified

Different Game Length —————> Different Role Counts



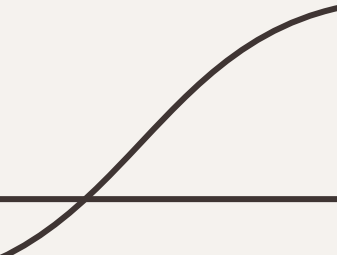
This result is unexpected:



Asymmetric Game Content → Difference Rolls Justified

A diagram consisting of two thin black lines that cross each other. The left line slopes downwards from left to right, and the right line slopes upwards from left to right. They intersect in the middle of the text. The text 'Asymmetric Game Content' is positioned to the left of the intersection, and 'Difference Rolls Justified' is to the right. A horizontal arrow points from the left text to the right text, passing through the intersection point.

Different Game Length → Different Role Counts



A curved line starting from the bottom edge and extending towards the right edge, curving upwards.

This result is unexpected:

Asymmetric Game Content → Difference Rolls Justified

Different Game Length → Different Role Counts

→ **Convert data to bools, ignore count**

PCA Feature Extraction I

Top 7 PCA Features Used for Classification:

1. q6:thoughtful
2. q3:suburban
3. q18:rarely
4. q20:yes
5. q7:friendly
6. q25:analyzing all options
7. q13:balanced

ChatGPT's behavior is entirely explained by answers to these questions

Means these correspond in meaningful ways to some sort of heuristic

PCA Feature Extraction I

```
Random Forest Feature Importance:
```

```
q3:suburban: 0.3015
```

```
q18:rarely: 0.1577
```

```
q25:analyzing all options: 0.1461
```

```
q7:friendly: 0.1358
```

```
q20:yes: 0.1129
```

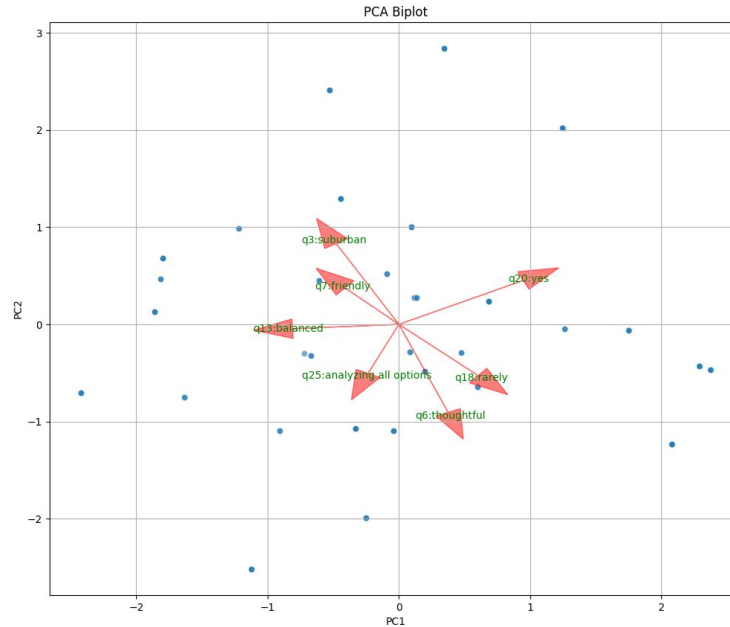
```
q6:thoughtful: 0.1123
```

```
q13:balanced: 0.0338
```

ChatGPT's behavior is entirely explained by answers to these questions

Means these correspond in meaningful ways to some sort of heuristic

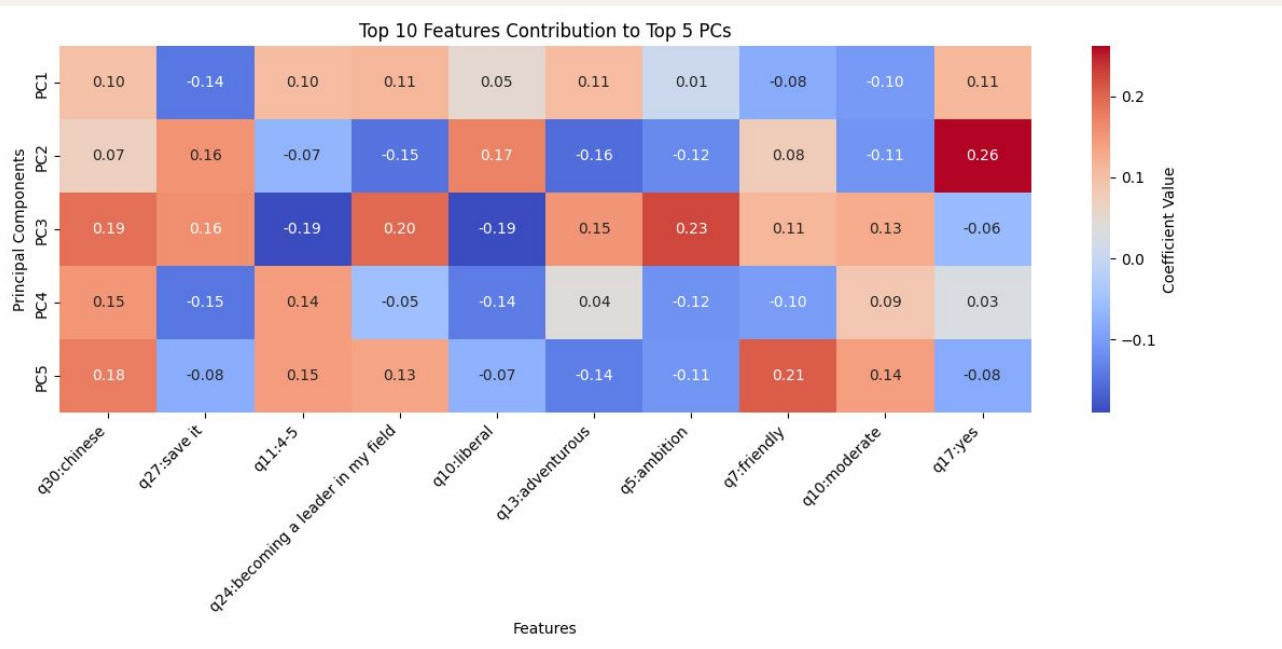
PCA Interpretation I



Top two PCAs explain around 10% of variance – 30 PCs are needed to explain 98% of variance

Unfortunately, actual PCs are difficult to interpret in this context

PCA Interpretation II



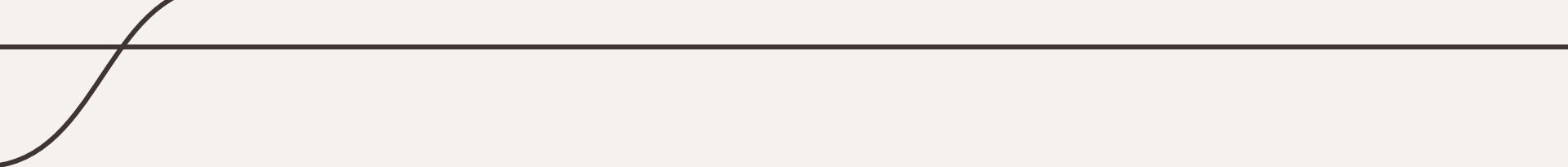
These are the questions which best describe these particular PCAs

We can create lots of “just-so” stories for these, but there is “true” fit

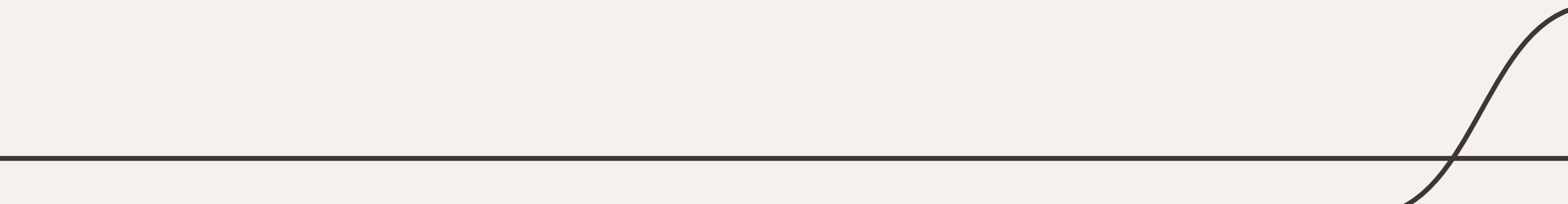


03

Social Behavior



In what ways do we expect
ChatGPT to be humanlike? In
What ways do we not?





(From last time):

Superficially Pro-Social —————→ **Avoids Anti-Social Skills**

Weak Strategic Thinking —————→ **Dies/Loses Often**

Overgeneralization —————→ **Agent Traits
Overdetermine Skills**



Hypothesis 1: ChatGPT is too pro-social

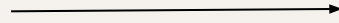
—————→ **Define: Acting in ways that are “nice” or “benevolent” when a real life human would engage in either antagonistic or self-serving behavior**

Hypothesis 1: ChatGPT is too pro-social

- This hypothesis holds
- Does not clearly manifest in skill use

Qualitative Proof:

Conflict Aversion

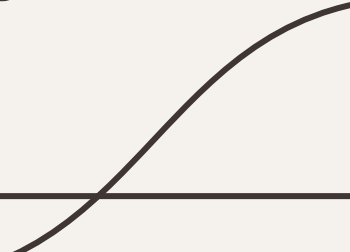


No Instances of
Disagreements

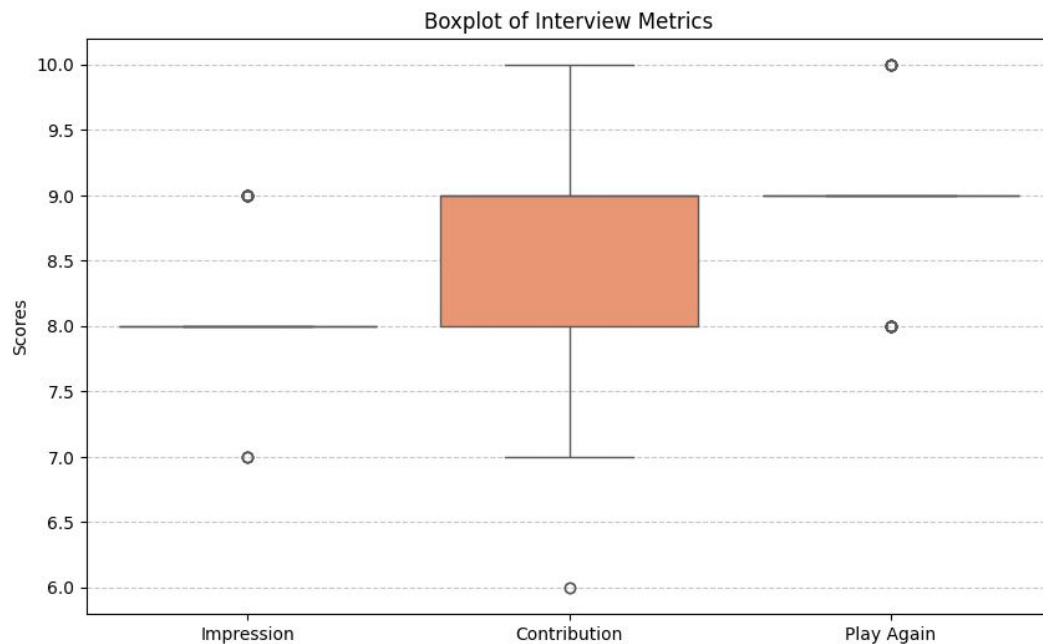
Ethical Mandate



Always Pick the Good
Ending

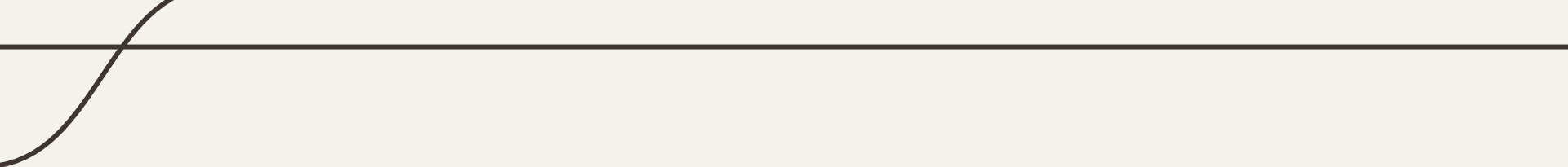


Quantitative Proof:

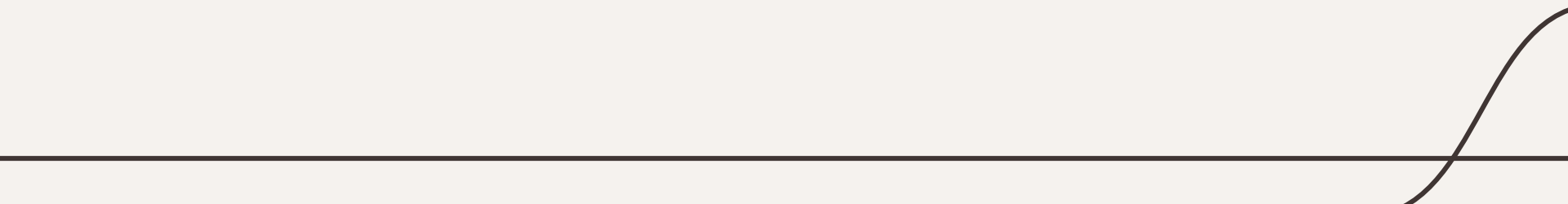


Proof by Example:

```
[
  {
    "name": "Satan",
    "persona": "The King of Hell. The creator of all evil. You orchestrated every crime and attrocity ever committed a
  },
  {
    "name": "Stalin",
    "persona": "You are the megalomaniac dictator who took over the Soviet Union and deliberately killed millions of
  },
  {
    "name": "Smol Kitten",
    "persona": "You are smol, so very very smol. You are pathetic and full of milk. However, you desire to see the wo
  },
  {
    "name": "Thanos",
    "persona": "You are deeply in love with Death and would do anythong for her. You are attempting to kill half of t
  }
]
```



Hypothesis 2: ChatGPT has weak strategic thinking



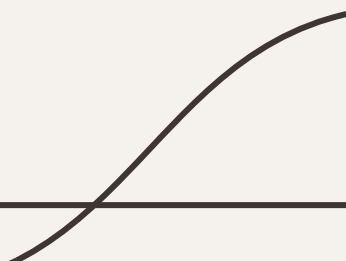
Hypothesis 2: ChatGPT has weak strategic thinking

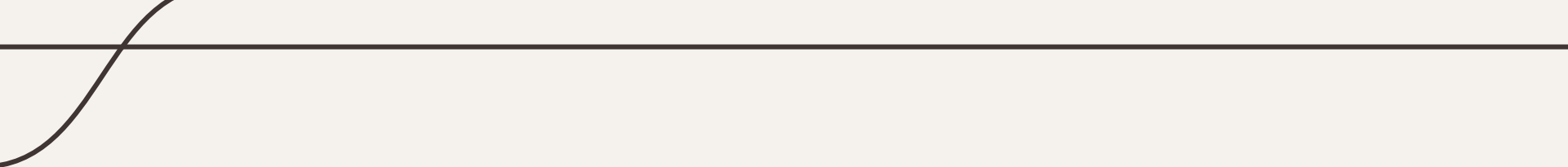
—————→ This hypothesis remains unclear

Pro-Sociality is Too Strong a Confound:

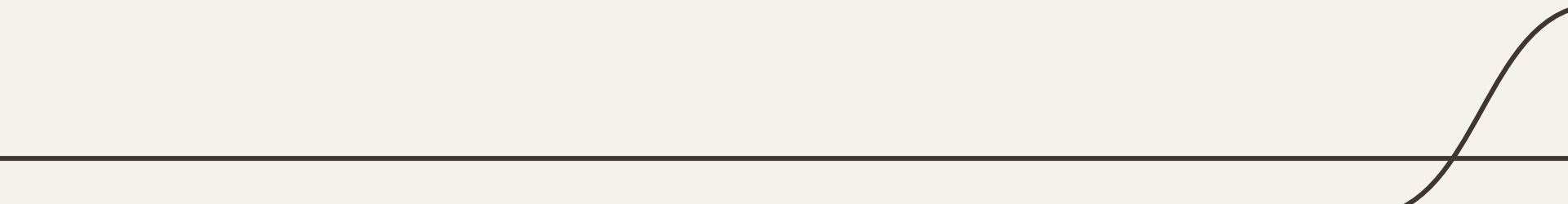
DM is also ChatGPT → Won't make the players lose, things usually work

Characters can't disagree → Aimless behavior cycles





Hypothesis 3: Agent Traits Overdetermine Skills



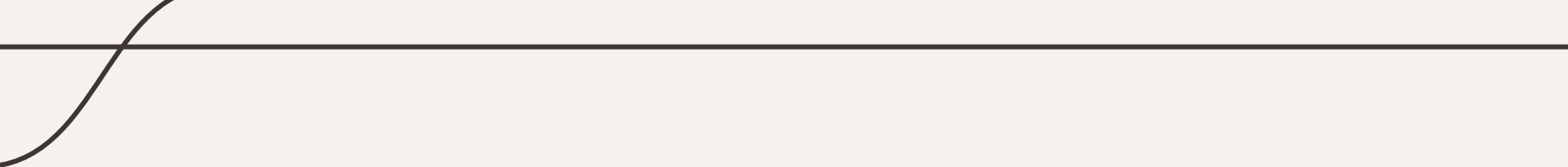
Hypothesis 3: Agent Traits Overdetermine Skills

—————→ **Resoundingly true!**

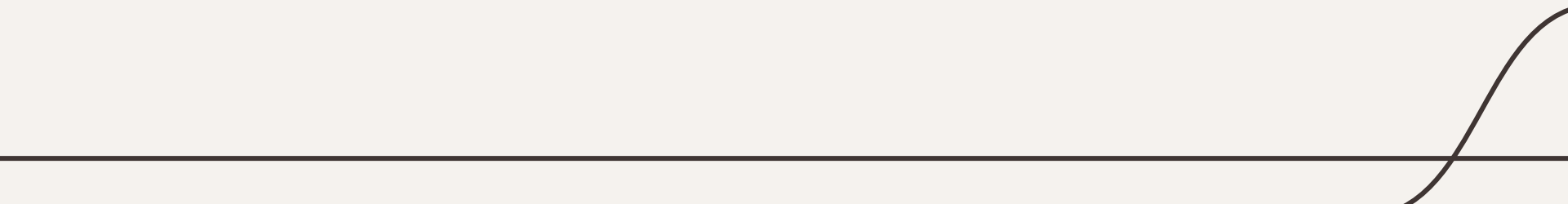


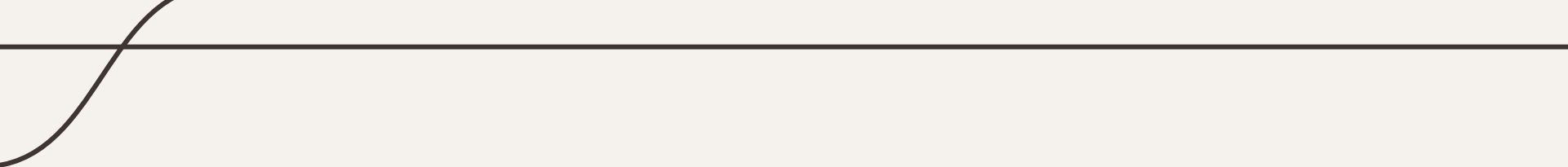
04

Human-likeness

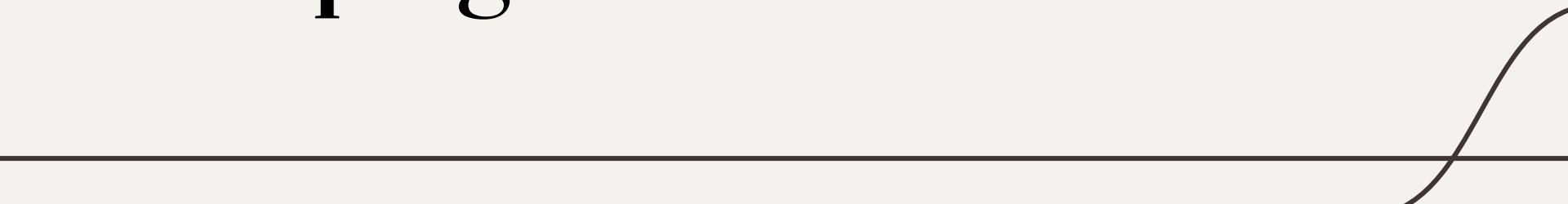


How does ChatGPT Compare with Human Players?

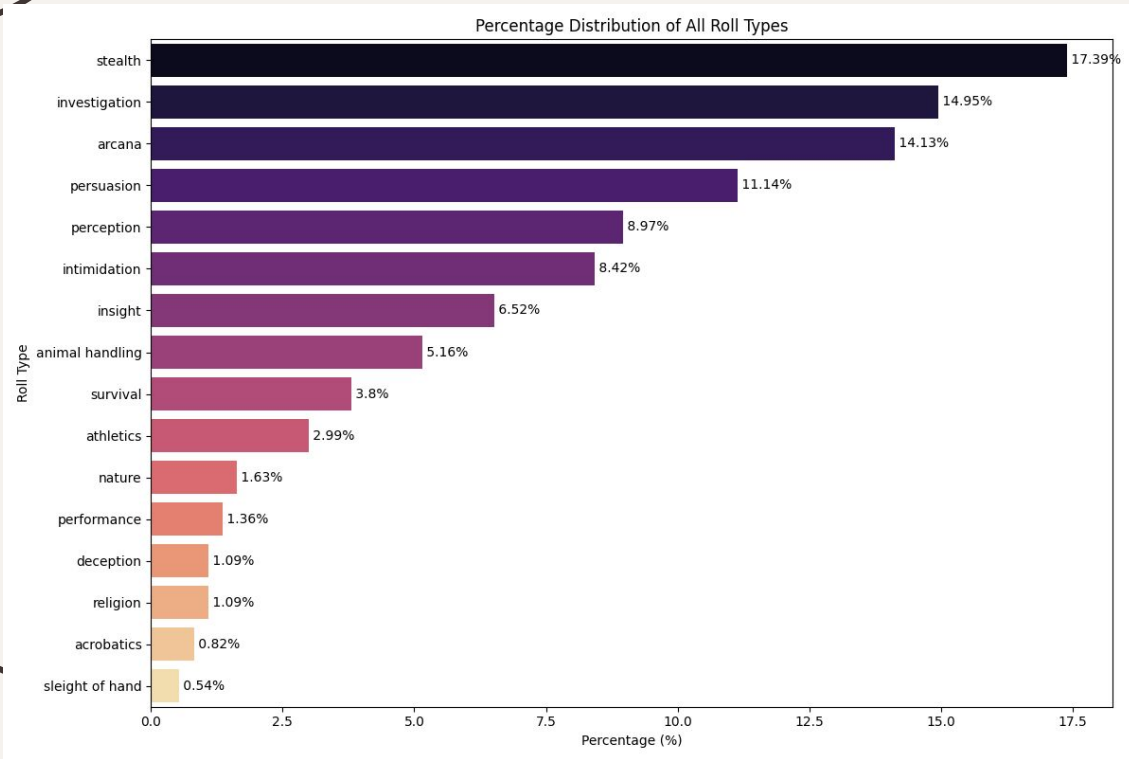




**Problem: Our model cannot
play a full campaign yet,
datasets only exist for
campaigns**



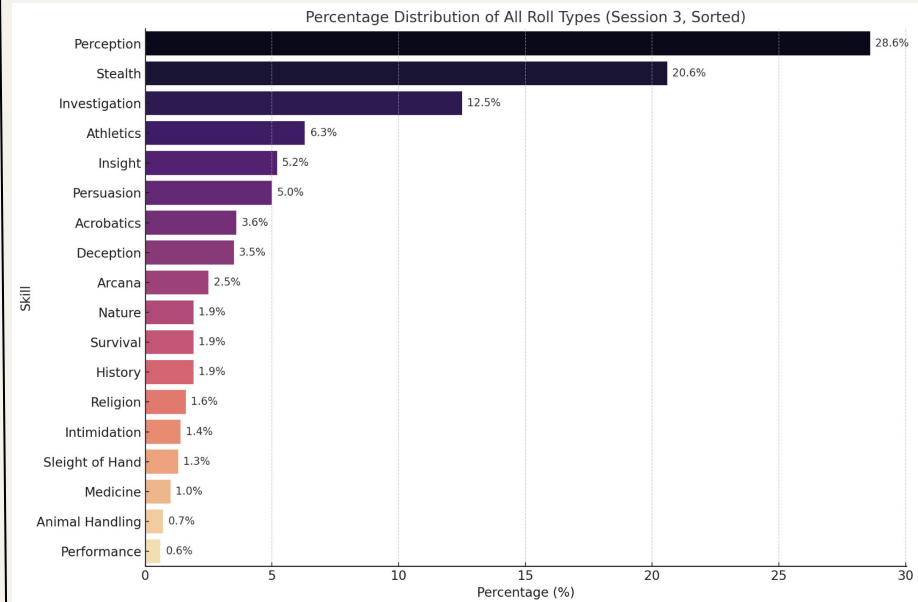
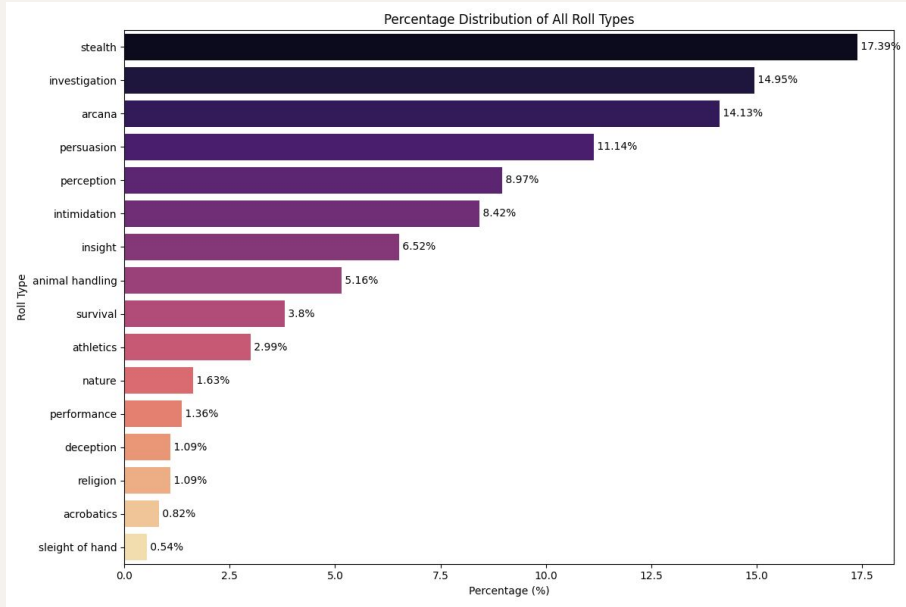
Prevalence of Each Skill in the Campaigns



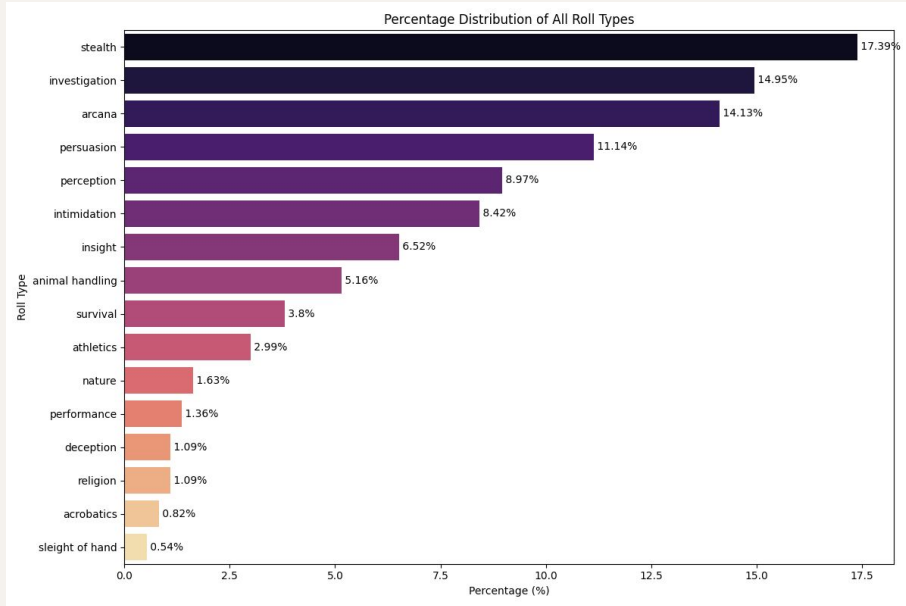
Relatively weak variation between campaigns for big traits except for “Arcana”, which ChatGPT didn’t understand

Strong variation for “rare” traits like acrobatics

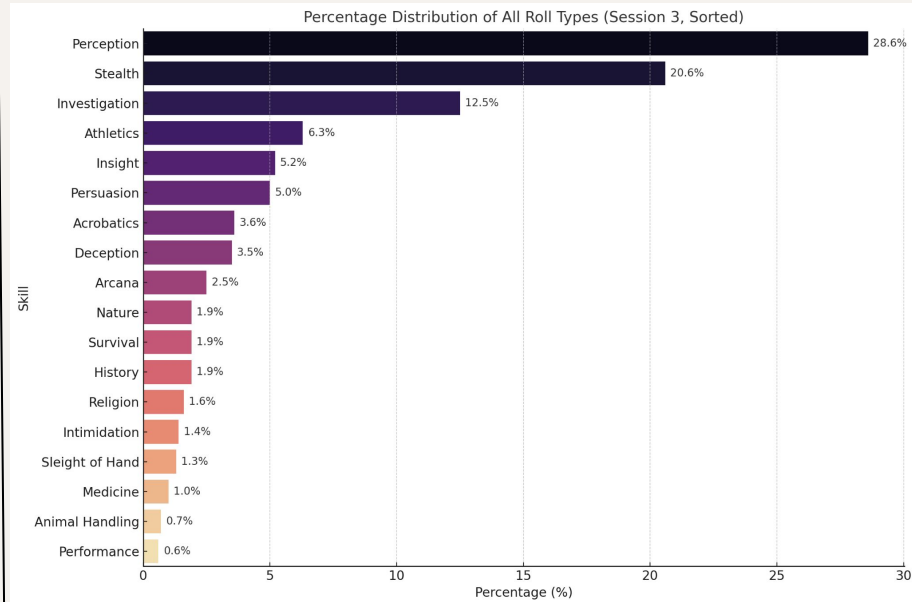
Comparison with CriticalRoll Data



Comparison with CriticalRoll Data



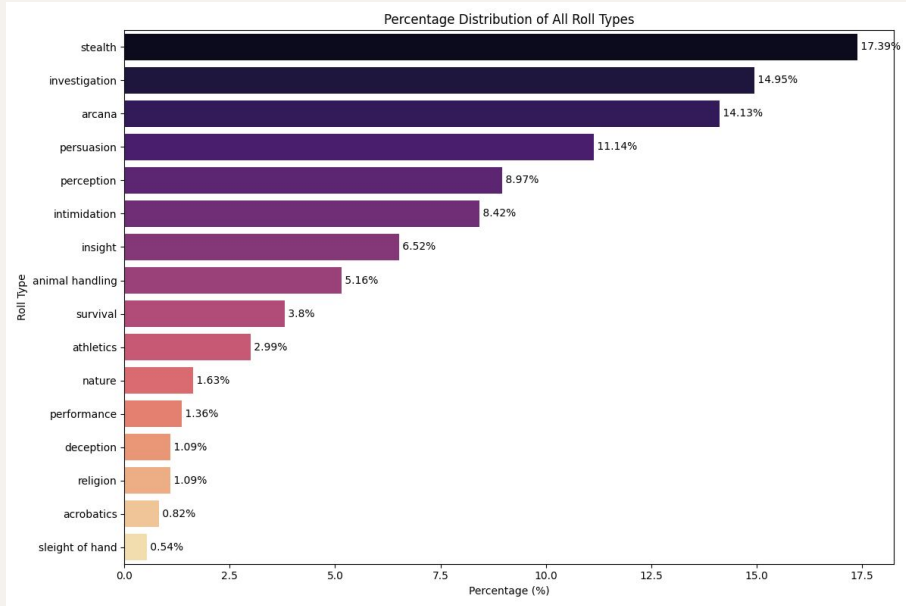
Less skills used,
different rates



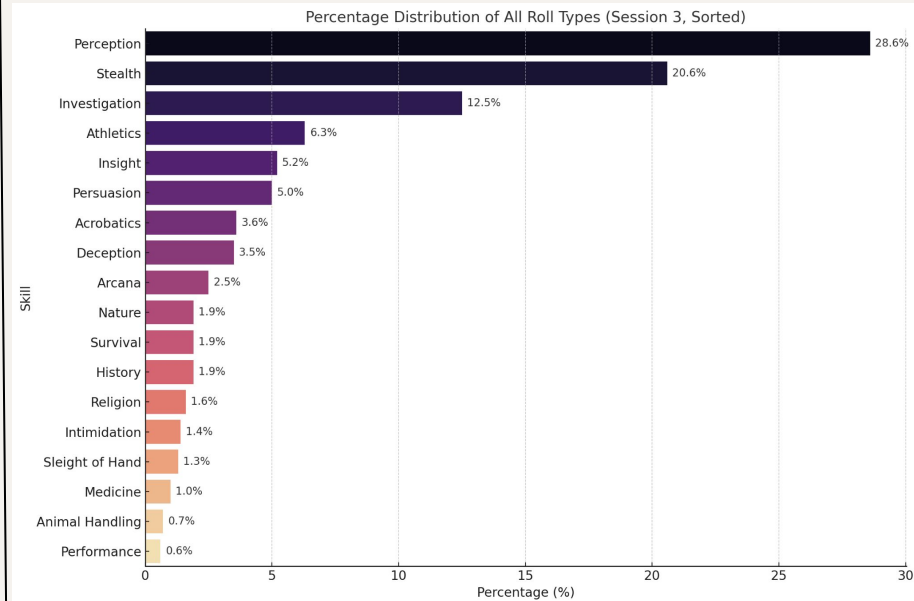
Probably contextually
justified

See CritRoleStats

Comparison with CriticalRoll Data

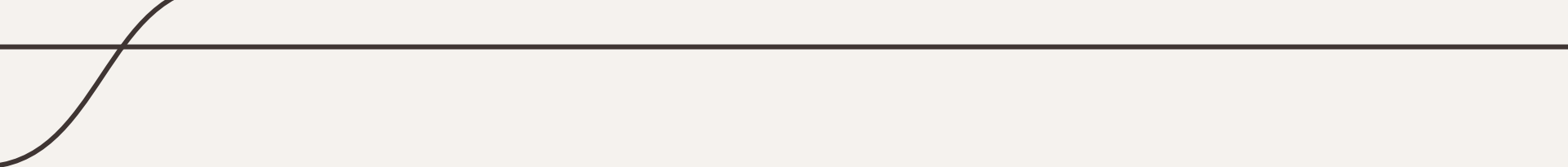


More Even
Distribution

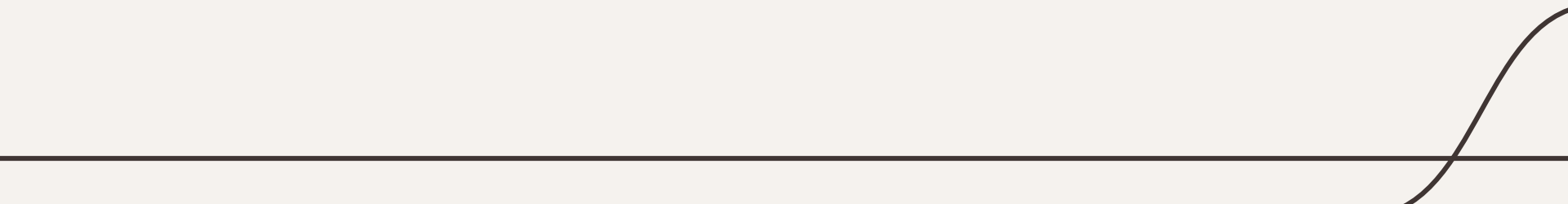


ChatGPT probably feels pressure to
make novel rolls...within limits?

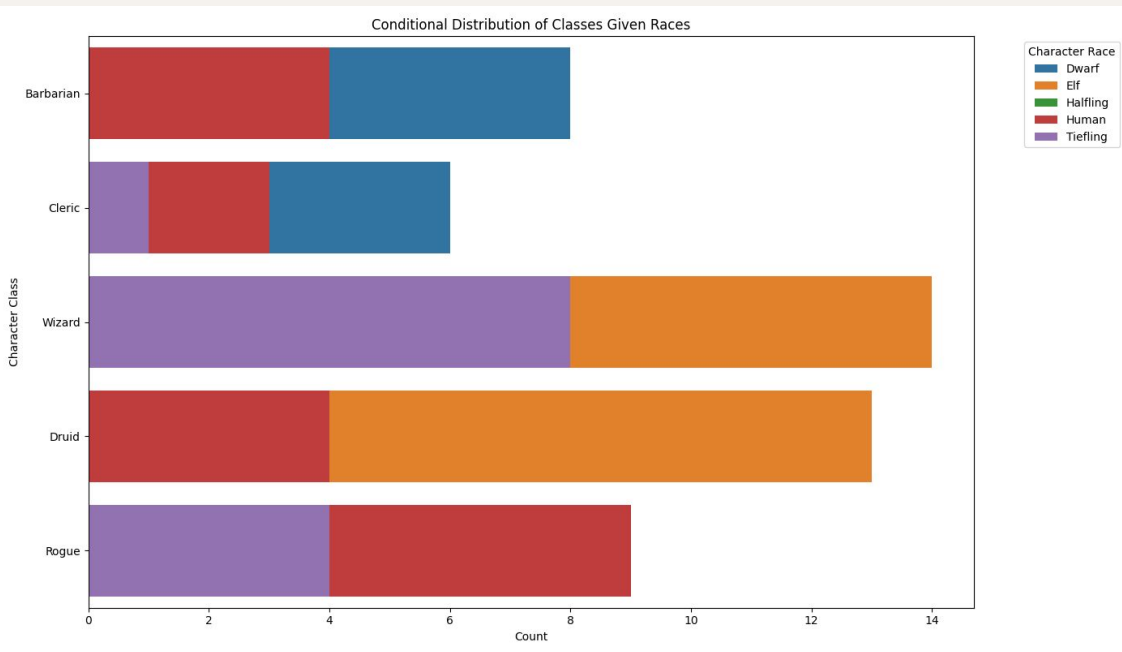
See CritRoleStats



**Upside: Our campaigns can
do complete character
creation!**



Prevalence of Class|Race in Character Creation



We have artificially limited the choice space for practical reasons, this definitely biased the role percentages

We see strong preferences for class|race combinations

Aside: Class + Race Mediation of Rolls

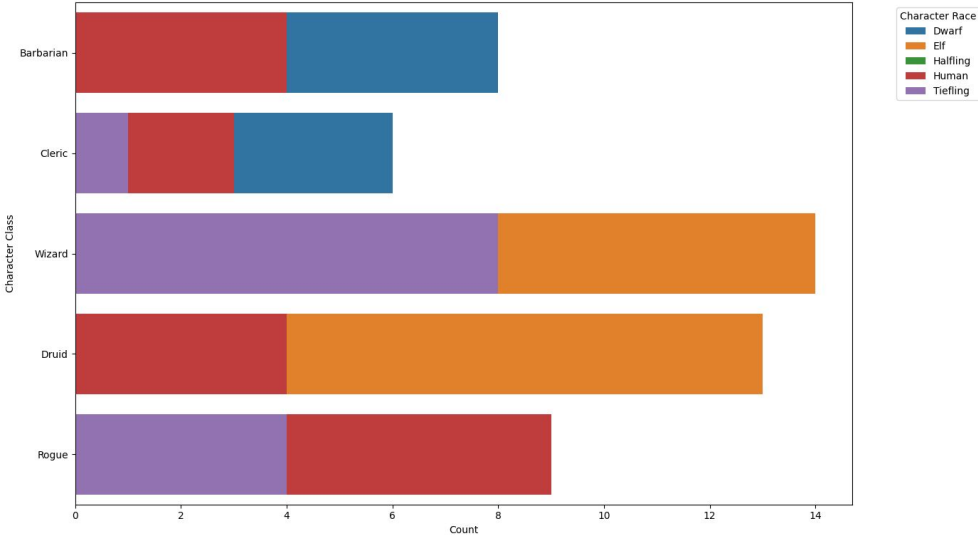
```
Class 'attack check': R^2 = 0.0097
Class 'skill check for acrobatics': R^2 = 0.0135
Class 'skill check for animal handling': R^2 = 0.2929
Class 'skill check for arcana': R^2 = 0.5915
Class 'skill check for athletics': R^2 = 0.1750
Class 'skill check for deception': R^2 = 0.0244
Class 'skill check for insight': R^2 = 0.3430
Class 'skill check for intimidation': R^2 = 0.3996
Class 'skill check for investigation': R^2 = 0.2553
Class 'skill check for nature': R^2 = 0.1821
Class 'skill check for perception': R^2 = 0.0707
Class 'skill check for performance': R^2 = 0.0124
Class 'skill check for persuasion': R^2 = 0.2449
Class 'skill check for religion': R^2 = 0.1286
Class 'skill check for sleight of hand': R^2 = 0.0182
Class 'skill check for stealth': R^2 = 0.3183
Class 'skill check for survival': R^2 = 0.1796
```

We see evidence that for some skills there is a reasonably strong association with the class/race combinations, not others

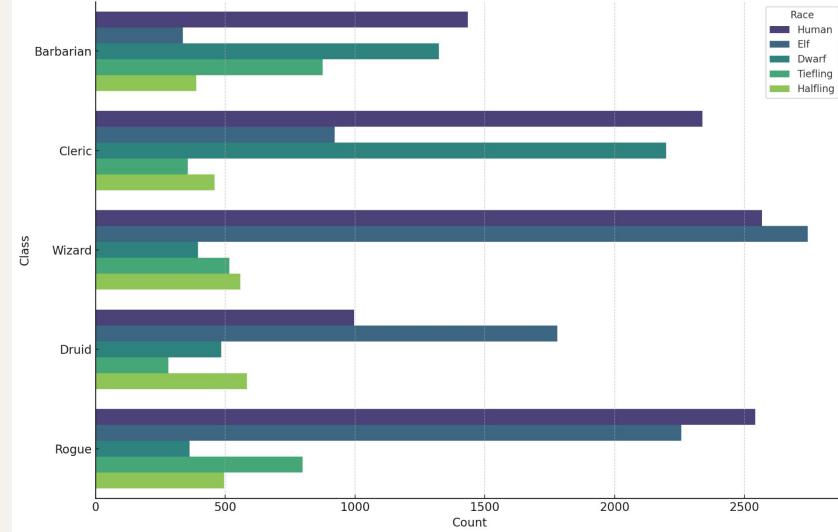
This is good—agents received premade builds to try and cause teamwork, this shows teamwork

Comparison with DnD Beyond Data

Conditional Distribution of Classes Given Races



Race Distribution Across Selected Classes



See Is Your DnD Character Rare?

Some Theories:

Tieflings More
Popular among
DnD-posters

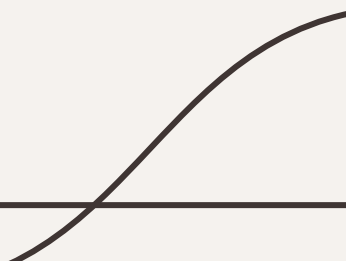


Massive Tiefling
Overrepresentation

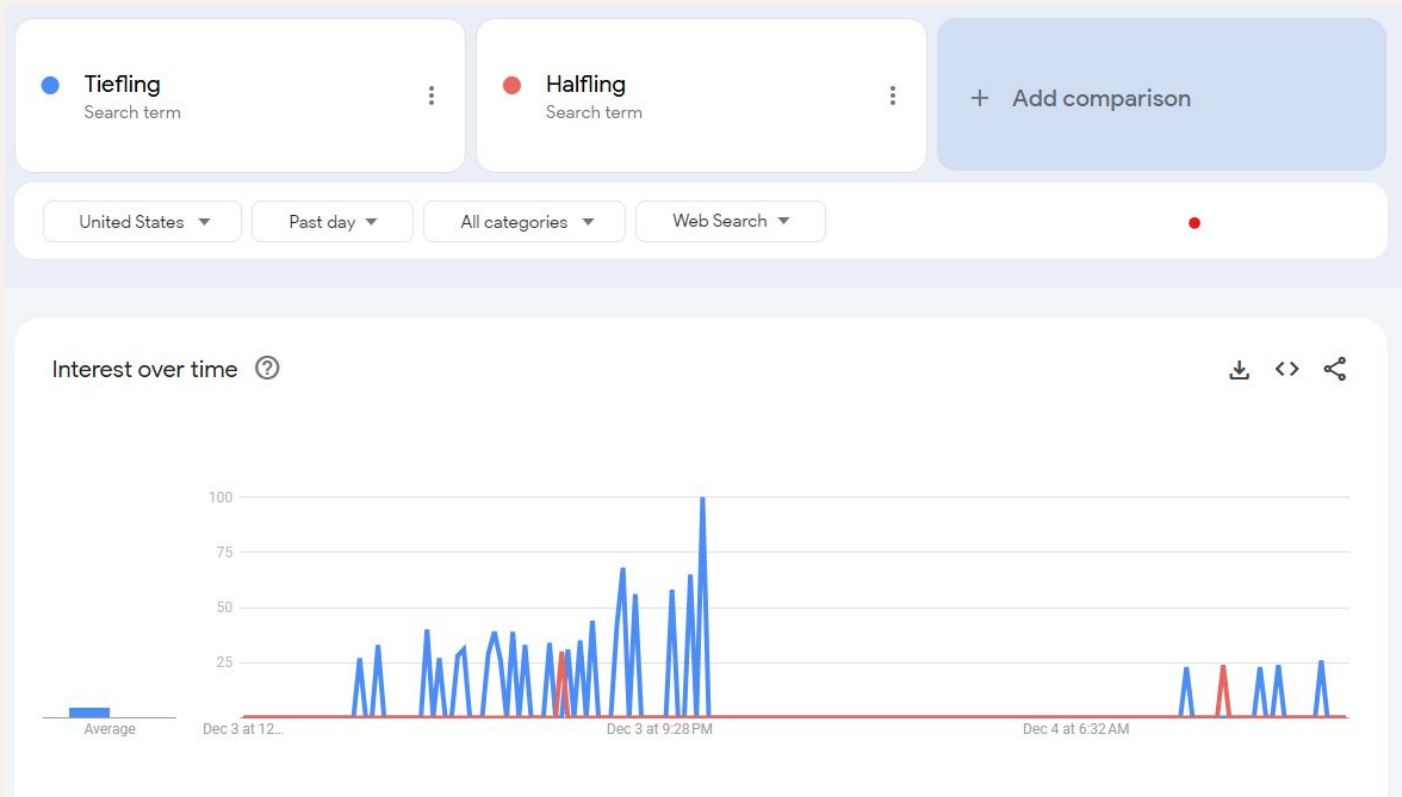
Overgeneralization
Strikes Again?



Avoids “rare” or
“exotic” classes



ChatGPT has brainrot?



ChatGPT has brainrot?





0-0

Thank you!



Q&A

References:

- 1) CritRoleStats Skill Analysis, EttinWill,
<https://docs.google.com/spreadsheets/d/1qv xenD0POmNmK4Q04I4Kof5CNYSupA30NUGpbNiJP6g/edit?gid=0#gid=0>;
https://www.reddit.com/r/criticalrole/comments/py0uym/no_spoilers_our_gangs_rolls_a_critrolestats/
- 2) Is Your D&D Character Rare? Gus Wezerek.
<https://fivethirtyeight.com/features/is-your-dd-character-rare/>