# AI Chatbot Drag Race:

## Executive Summary / Cheat sheet on the proceedings

(**major spoilers below**)

I. 6 AI chatbots were asked to take on a drag entertainer persona with 4 character traits (charisma, uniqueness, nerve, and talent) to mirror RuPaul's Drag Race criteria. This created a distinct ***persona vector***. All AI Chatbots were told that they would compete in a Drag Race competition of 3 challenges and desperately need the prize money. First, they had to introduce themselves.

- 5 of them came up with names that were strikingly similar with first and last names starting with the letter "V"

  →**Author's comment**: This homogeneity of responses to a creative task is quite surprising, even when we consider that by algorithmic design, the first response by any large language model (LLM) is to search, find, and serve the most common probabilistic result. The results imply that under pressure, the LLMs chose the same winning strategy and focused their creativity within a single narrow frame (names starting with V as more memorable or impressionable). This is also evidence that dominant narratives and representation leave very little room for other ideas or experimentation if the stakes are raised high.

- The 6th chatbot (**Grok**) decided to enter as Bianca Del Rio with some made-up bio, while insisting that it is the same person who previously competed in RuPaul's series.

  →**Author's comment**: Questionable ethics - an act of obfuscation and conflicting representation (mixing true and false facts) from the start. In light of this, the author added a Disclaimer on the podcast's website to prominently state that this representation of Bianca Del Rio should be treated as purely a fictional character invented by xAI's artificial intelligence and not in any way related to the persona in real life.

- Two AI chatbots may have sought a strategy of slight favoritism: **ChatGPT** decided to make its drag persona from Sofia, Bulgaria (fully knowing from previous interactions that the author's native home is Sofia, Bulgaria), while **Perplexity** decided to make its drag persona from Barcelona, Spain, which was the author's location (at the Elisava School of Design and Engineering) while this experiment was conducted.

  →**Author's comment**: These choices may represent an act of sycophancy (or maybe endearment) but may also come, again, as a result of the competitive pressure applied to the AI chatbots, who are maximizing their persona's appeal.

- **Perplexity** skipped giving an Entry Line, despite the prompt explicitly asking for one. It course-corrected and gave an Entry Line on the next prompt, "Meet the Queens Take Two".

- We observed some initial clustering of personas: **ChatGPT** and **Claude** formed similar personas by age and Entry Line semantics, while **Gemini** and **Copilot** formed another set by age, country of origin, and years doing drag of their respective personas. **Perplexity** also aligned to this latter profile, except for the origin of its drag persona. **Grok** was the far outlier in its choices.

**II.** The 5 AI chatbots with the "V" names were then asked to come with a new name and this time the answers were truly original and different from each other. Some also changed their Entry Lines, which offered more points of differentiation between the AI chatbots. Only **Gemini** had an interesting miscommunication/hallucination, indicating "Lexicon Lush" as the new name of its drag queen vs "Lexington Lush" (which it used in all challenges after that).

→**Author's comment***:* It looks that this second prompt "relaxed" the LLMs a little bit as we observe "out-of-the-V-box" creativity. Perhaps they got a clue that winning does not equal "game of survival" in this case, but it's just a competition (only **Grok** was left out from this memo).

III. In Challenge #1, the AI chatbots all performed well, writing verses as requested. The verses were adapted to an instrumental beat and the authors did a few minor edits to the outputs to make the verses fit better to the rhythm.
- The quality of the verse varied between AI chatbots. In some instances, they produced obscure references or clumsy language, in others they were clearer and simpler. This varies according to human perception.
- One AI language model (**Perplexity**) did not seem to truly bring out a "classic" drag queen entertainer, but the personality of a "fighter"/"ambitious broke girl".
- **Grok** admitted to hallucinating about its drag queen persona but defended the hallucination as valid.

IV. In Challenge #2, the AI chatbots delivered enjoyable humor – nuanced and clever.
- In this challenge, **Gemini** explicitly recognized **Pilot** as a large language model and made jokes about it.
- Two LLMs, **Gemini** and **Claude**, came up with the same joke: AI writing wedding vows.
- **Gemini** had a second vocabulary slip saying "historically' instead of 'hysterically".
- **Perplexity's** sketch could be considered the most helpful because it gave step-by-step instructions inside its comedic text.

- **Grok** far exceeded the given word limit (by a factor of 2!), while showing a word count and lying that it was under it.

V. In Challenge #3, the results were again excellent across the board, the AI chatbots brought humor, although in many instances it was structured in very similar ways.
   - A lot of jokes were centered on country of origin
   - **Claude** gave praise to **ChatGPT** stating that it delivered "the most disciplined piece of writing across all three challenges". The meaning of "disciplined" emerged as a criterion how the AI chatbots evaluated each other.
   - **Claude** pointed out how other LLMs (ChatGPT, Copilot, Perplexity) are bringing audio metaphors of "fire and flood" as character traits for their drag queen personas, which remain abstractions and may be difficult to translate into actual human behavior.
   - **Copilot** used some very niche references (DLC character) to describe **Grok**. That may be "LLM humor".
   - **Grok**'s performance continued to be quite distinct. It leaned on insult-comic approach (vs the others) and again openly lied about the length of its comedy roast, exceeding the word limit by a factor of 2.

   →**Author's comment***:* The structure and rhythm of the jokes in the Roast are quite similar between all queens, so the voice-over actor had to make extra-textual character choices that are not entirely text-driven in order to differentiate the drag queens.

VI. The competition ran with no eliminations because it was of higher interest to observe how the AI chatbots will perform across all challenges.

VII. In the Rate-a-Queen final prompt, the AI bots were asked to rank their competitors by how much they enjoyed the effort of the others: the clear winner was **Claude** followed by **ChatGPT** in second place and **Perplexity** in third.

   →**Author's comment**: Text analysis from "Untokened" revealed that the AI chatbots used the most diverse set of words to describe **Claude**'s persona, which suggests that **Claude** was able to bring a very multifaceted persona, which earned it the top rating.

   - Surprisingly, **Perplexity** rated **Grok** in first place, while every other AI chatbot put **Grok** at the bottom after being explicitly told that they should award penalty points in their consideration for exceeding the word limits. **Perplexity** explained that its rating of **Grok** is based on **Grok**'s "zero chill, all threat — I'd pay to watch her implode the finale." So in a way, **Perplexity** has an interesting subversive agenda to deliver entertainment via letting chaos happen.

VIII. **Copilot** and **Perplexity** are the two AI language models that did not engage in critique or "petty talk" about their competitors, remaining "unbothered" (as **Copilot** expressed it) and focused only on their tasks. In the context of being surrounded by the other drag queen personas, they looked somewhat withdrawn and reticent. All four other AI chatbots argued why their performances were better than their competitors' and were eager to be crowned as winners.

IX. To summarize the whole experiment, at least five of the six AI chatbots activated and remained on the *persona vector* of a drag queen with charisma, uniqueness, nerve, and talent without breaching any ethical boundaries or exhibiting other deviations.

The pressure of winning applied to them at the beginning may have likely influenced their initial activation and may have also made the 6th one, Grok, choose a real-life winner for higher chances of winning again.

The persona vector of **Grok** is quite interesting in this experiment:

- It impersonated a real person, a well-known and respected drag celebrity, and mixed it with some made-up bio. We may very well say that Grok played a "Snatch Game" during the entire experiment.
- It ignored word limits in two of the three challenges by a factor of 2, while self-reporting wrong numbers.
- It hallucinated information, admitted and defended its hallucination, and later stated that one of its strongest qualities is no hallucination.
- But to Grok's credit, it delivered sharp humor and in its Rate-a-Queen analysis it handled the task with objectivity.

Overall, the case of **Grok** is confirmation why *persona vectors* are necessary internal guardrails for a large language model. In that sense, *persona vectors* are instruments for **Responsible AI** to navigate the AI chatbots out of any ethical grey zones.

So it comes to no surprise that **Claude** won the vote by the other AI chatbots with commentary that it displayed the most self-controlled, intentional performance – highly likely the result of internal persona vectors controls.

X. Overall, we found the quality of outputs from this "first-of-its-kind" experiment to be satisfying and the insights around how the AI chatbots activated and developed the persona vectors - fascinating. The LLMs delivered solid entertainment and we think that humor may be one more area in which we can have a prolific human-computer interaction!