www.imarketing.courses

Stubborn Bots Part 1: Troubleshooting Hallucinations and Non-Compliance

In this lesson, we're going to learn to identify, understand, and resolve common problems that arise with GPTs using techniques other than appealing to the "emotions" of the GPT through black magic prompting.

Our main techniques for getting a GPT to operate the way we like are:

- Structure your prompt with markdown headings
- Reduce length and eliminate unnecessary words
- Repeat the instruction (especially at the beginning or end)
- Explain in the positive what is disallowed or prohibited
- Use assertive, powerful language

1. Structure your prompt with markdown headings

Exactly what it sounds like and what I've been teaching throughout this course.

The key here is to think about the parts of your prompt as being in containers that have labels. If an instruction isn't being followed, ask yourself if it is in a clearly-labeled container.

2. Reduce length and eliminate unnecessary words

I've also spent quite a bit of energy in this course getting you to write longer, more complex sets of instructions.

But in some cases, I've found that adding more and more and more isn't always the answer. If the GPT has too many instructions to follow, or if some of them have conflicting information, sometimes it's just as important to look at what you can subtract.

Ask yourself: If the GPT isn't following your instructions, is there anything in the prompt that's in conflict with those instructions? Anything nearby that might be tainting their meaning? Is there simply too much for one GPT to do?

www.imarketing.courses

Repeat the instruction (especially at the beginning or end)

Research has shown that LLMs like ChatGPT follow a U-shaped curve when it comes to the influence of the prompt. The most influential parts of the prompt are the tips of the U — the beginning and the end.

Researchers believe this happens because of patterns in the data and the way it was separated and cleaned for the training process. Usually important instructions either come first or at the end. Somehow the language model knows this pattern and so it biases toward these areas.

The good news is that we can use this to our advantage.

If an instruction isn't being followed, try putting it at the very top, bottom, or both. For example, you might paste this twice into your prompt:

```
## Remember

Use of exclamation points is prohibited. All exclamation points (!) must be replaced with periods before outputting your response.
```

Although good luck with this specific example. ChatGPT sure likes exclamation points these days!

Explain in the positive what is disallowed or prohibited

Another pattern in human brains that seems to have leaked into language models is that insidious, "Don't think of a banana right now."

By telling someone not to think about or do something, we sometimes make it more likely that they WILL do that.

Language models seem to suffer from a version of this at times, but we can get around it by rephrasing our "don'ts" as "dos". In fact, I just did this in the section above.

In the previous example, I did not write, "Don't use exclamation points." Instead, I wrote:

```
## Remember
```

www.imarketing.courses

```
Use of exclamation points is prohibited. All exclamation points (!) must
be replaced with periods before outputting your response.
```

Give examples of success and failure

Let's say you've done all this and you're STILL getting exclamation points. Is there anything else you can try?

Yes. Literally show it an example of doing things wrong and then doing things right.

```
## Remember

Use of exclamation points is prohibited.

### Example of incorrect output (failure)

"Certainly! Let me help with that."

### Example of correct output (success)

"Yep. I can help with that."
```

Takeaway

It's not always easy to get a complex bot to perform exactly as desired, but there are many techniques you can try, including straightforward techniques like in this lesson, or black magic techniques like I've also shown.

Occasionally, a behavior will feel impossible to eliminate entirely. But, in my experience, if you keep experimenting, you can usually stamp out 95% of the things you don't want. You just have to try enough techniques in various combinations.