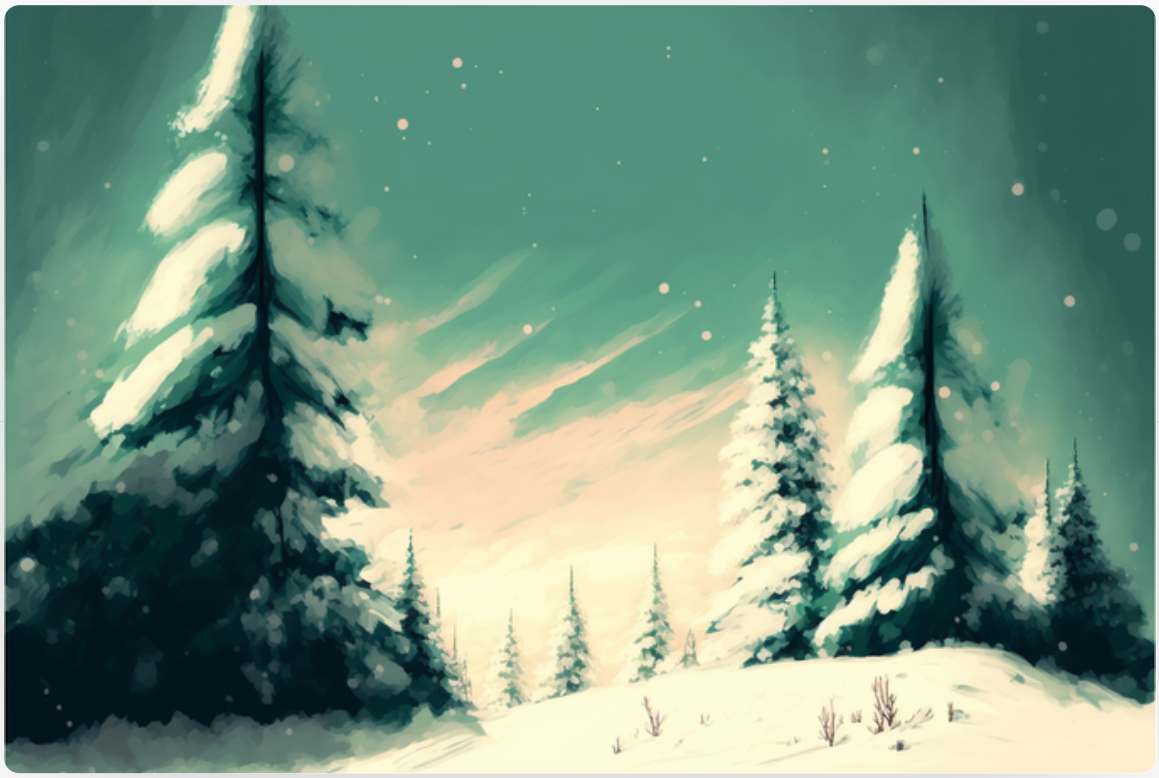
Output from AI LLMs is Non-Deterministic. What that means and why you should care.



By Steve Engelbrecht

May 12, 2023

Discover the challenges of non-deterministic language models and strategies to get more consistent results

Picture this: You're using a cutting-edge large language model (LLM) like ChatGPT to generate a product description for your online store. You input the same prompt twice, expecting to receive identical results each time. But, to your surprise, the outputs are entirely different! In some cases, you

may see artifacts that don't even make sense – as if the LLM is making things up.

This unpredictable behavior is due to the non-deterministic nature of LLMs, which can be both a blessing and a curse in content creation. In this article, I'll explore what non-deterministic means, why it can be a concern for businesses, and how to overcome these challenges to achieve more consistent results.

Understanding Non-Determinism in LLMs

Non-determinism, in the context of LLMs, means that the model can produce different outputs even when given the same input. This behavior is a byproduct of the complex neural networks and vast amounts of data used to train these models. While non-determinism can lead to creative and diverse outputs, it can also cause inconsistency, which may be undesirable in certain business applications.

The Risks of Non-Determinism

For businesses like Sitation, which rely on LLMs to generate product descriptions and other content, non-determinism can be a double-edged sword. On the one hand, it allows for a broad range of responses and can inject creativity into the content generation process. On the other hand, inconsistency can lead to confusion and frustration for both the business and its customers.

Imagine the confusion when a customer notices two different product descriptions for the same item, or when a marketing team member needs to approve content, but they receive varying outputs each time they run the same prompt. These scenarios can harm a company's brand image and

credibility, making non-determinism a serious concern for businesses. Too many creative liberties will allow the model to "hallucinate" – to make things up – that were not part of the original prompts.

Non-Determinism is Also a Good Thing!

There's also a major benefit to non-determinism… a degree of randomness adds a creative spark to the output, and allows you to quickly iterate if you don't like the initial draft. Sitation's AI product, RoughDraftPro, makes extensive use of this feature by allowing for a rapid fire "rewrite" of AI-generated product content, including titles, short descriptions, feature bullets, or entire PDPs. Dialing in the "temperature" parameter in the GPT APIs is key to understanding the degree of randomness in output. Perhaps a better way to think of it is this – when we dial that up, we let the model take more risks. In doing so, sometimes it will come up with something better. And indeed, turning temperature all the way down does in fact make the output far more mechanical and deterministic.

Mitigating Non-Determinism in LLMs

While it's impossible to eliminate the risks of non-determinism entirely, there are some strategies that can help you achieve more consistent results when using LLMs like ChatGPT:

01. **Craft Specific Prompts**: Provide clear and detailed instructions to the LLM, specifying the format and content you expect. This can help narrow down the range of potential outputs.
02. **Use Templates**: Create a template with placeholders for the variable parts of your content. This can ensure that the structure remains

consistent while allowing the LLM to fill in the specific details.

03. **Iterative Refinement**: Instead of relying on a single pass, iteratively refine the output by providing feedback and making adjustments to the input prompt. This can help guide the model towards the desired output.

04. **Human-in-the-Loop**: Combine the power of LLMs with human expertise. Have a person review and edit the generated content to ensure consistency and quality.

All of these strategies are used extensively in Sitation's AI toolset, RoughDraftPro, with the intention of creating consistency and scalability such that the technology can be leveraged in a production environment for high-volume content creation. Learn more about RoughDraftPro here.

Conclusion

Large language models like ChatGPT and their underlying APIs have revolutionized the way businesses generate content. However, their non-deterministic nature can pose challenges when consistency is crucial. By understanding the implications of non-determinism and adopting strategies to mitigate its effects, you can harness the full potential of LLMs and create high-quality, consistent content for your business.