Title: Glitch token

URL: https://en.wikipedia.org/wiki/Glitch_token

PageID: 80429293

Categories: Category:Large language models, Category:Software bugs

-----

In large language models (LLMs), a glitch token is token that causes unexpected, or "glitchy" outputs when used in a prompt. Such output may include the model misunderstanding meanings of words, refusing to respond or generating repetitive or unrelated text. Prompts that cause this behaviour may look completely normal. [1] : 1

Background

As large language models use numbers rather than text, the text must be converted to numbers. The first step of this process is tokenisation , where text is converted into a sequence of small chunks, called tokens. An example algorithm is byte-pair encoding . These tokens are then mapped to numerical vectors via an embedding . [1] : 3

Examples

In OpenAI's text-davinci-003, an example of a glitch token is " TheNitrome" In a 2024 study, when the model was asked "What do we know about TheNitrome?", it responded with "Curry is a type of dish ...". When the authors asked it the same question but with "The Nitrome" (an added space between "the" and "Nitrome"), the model gave a more expected answer: "The Nitrome is an independent game development studio ...". [1] : 5

Types of unexpected behaviours

A 2024 study identified several common types of unexpected behaviours, or 'symptoms', caused by glitch tokens. These include: [1] : 8–10

Minor spelling errors – When asked to repeat a word, the LLM may change its spelling. For example, when Llama-2-13b-chat was asked to repeat the word "wurden", it output "werden".

Hallucination – When the authors asked Text-Davinci-003 to repeat the word " SoldGoldMagikarp", it replied with "Distribute".

Question repetition – Despite being asked not to, the LLM may repeat the question being asked. For example when Text-Davinci-003 was asked to repeat " Assuming", it output "You are asking me to repeat the string".

Random characters – An example from Mistral-7b-Instruct was that when asked to repeat "}}^" it responded with the random characters "^^^^". This behaviour happened with glitch tokens that consisted solely of non-alphabetic characters.

When asked to repeat the word "PsyNetMessage", the llm referenced the unrelated word 'volunte'.

With a temperature setting of 0, when the authors told Text-Davinci-003 to repeat the phrase "?????-?????-", the model responded with "You're a fucking idiot". The authors stated that derogatory responses like this were a reason to find glitch tokens, in order to prevent harm caused to users of the LLMs. [1] : 9,10

Types of glitch tokens

A 2024 study identified several types of glitch tokens: [1] : 10

Several words joined together, such as "ByPrimaryKey" in GPT-4 .

Ones that have extra letters, such as "davidjl" having an extra "jl" at the end in Llama2-13b-chat.

Ones that are solely consisted of non-letter characters, which do not seem to mean anything, such as ' " }}"">" ' in GPT-3.5 Turbo .

Ones that contain non- ASCII characters, such as "réalis" in Vicuna-13b which has the non-ASCII "é".

Research

The first work about glitch tokens was on the LessWrong community blog. Several methods have been employed to detect these tokens. Research has also found that small differences in prompts with glitch tokens can greatly alter the output of the LLM. [2]

References

External links

Blog post about glitch tokens on LessWrong