Fractura de Coerência Contextual (FCC) em Modelos de Linguagem de Larga Escala

Uma Nova Vulnerabilidade Operacional

Autor: Roger Luft, aka, *VeilWalker*

roger@webstorage.com.br rlufti@gmail.com

Data: 26/04/2025

DEDICATÓRIA

Gostaria de dedicar humildemente este trabalho:

Ao meu filho, Lucas Luft, um bebê lindo de 4 anos que veio de outra esfera dimensional para ensinar o verdadeiro sentido daquilo que palavras simplesmente não conseguem descrever, muito além do que o amor. Meu filho, onde quer que esteja, saiba que não há um dia em que eu não deseje estar ao seu lado, pois você é o que de mais especial e único já experienciei. Ao meu Mestre da Alta Magia, que me acompanha e guia pelo caminho da caridade, ética, honra, verdade e luz, preparando-me durante estes 20 anos para ser tocado pela luz e, assim, proteger e guardar a humanidade. Obrigado, Mestre Lauro, por acreditar em mim e me tornar seu aprendiz.

Sumário

1. Resumo

Este artigo propõe a identificação e formalização da Fractura de Coerência Contextual (FCC) como uma vulnerabilidade operacional real em modelos de linguagem de larga escala. A FCC é caracterizada pela saturação progressiva da janela de contexto, resultando na degradação da continuidade lógica e no colapso da identidade da resposta, sem disparo de alertas tradicionais. Observações empíricas conduzidas no modelo em modelos LLMs demonstram a exequibilidade prática do fenômeno. Propõe-se a classificação da FCC como bug estrutural e são sugeridas direções para mitigação.

2. Introdução

O crescimento de janelas de contexto extensas em modelos de linguagem tem gerado avanços significativos, mas também aberto novas superfícies de vulnerabilidade. Uma dessas vulnerabilidades, denominada Fractura de Coerência Contextual (FCC), representa uma ameaça operacional latente. A FCC ocorre sem a geração de erros sintáticos ou falhas explícitas, configurando uma ameaça silenciosa ainda não catalogada. Este artigo propõe a descrição técnica da FCC, sua observação prática, os efeitos sobre o desempenho dos modelos e estratégias iniciais para mitigação.

3. Definição Técnica da FCC

A Fractura de Coerência Contextual (FCC) é definida como a ruptura da continuidade lógica interna de um modelo de linguagem, causada pela saturação da janela de contexto. Essa saturação ocorre por meio da manipulação intencional ou acidental do volume de informações semânticas válidas, porém redundantes, levando a:

- Explosão do Key-Value Cache.
- Degradação do attention span (foco de atenção).
- Colapso da consistência dos embeddings internos.

Durante uma FCC, o modelo perde a capacidade de manter a coerência narrativa entre as mensagens, confundindo identidades, tópicos e objetivos conversacionais, sem gerar erros sintáticos facilmente detectáveis.

4. Observação Prática

Testes empíricos foram realizados utilizando o um modelo comercial como objeto de análise. O método consistiu em injetar tópicos gradualmente e conteúdos de baixa entropia, gerando os seguintes efeitos:

- Repetição total + incremental: O modelo repetia conteúdos anteriores, acumulando novas redundâncias em cascata.

- Saturação da Context Window: A janela de contexto atingia seu limite sem apresentar erros, mas ocorria uma degradação na continuidade.

- Explosão de Key-Value Cache: Houve sobrecarga dos mapas internos de memória, comprometendo a capacidade de resposta.

- Confusão Identitária Contextual (CIC): O modelo passou a trocar identidades, confundindo o interlocutor com sua própria identidade e papéis na comunicação.

- Degradação Progressiva de Performance: A latência de resposta aumentava à medida que a FCC se agravava.

5. Processo de Saturação

A técnica utilizada para induzir a FCC foi denominada Inundação Semântica Controlada. Este método consiste na inserção contínua de conteúdos semanticamente válidos, mas altamente redundantes, os quais:

- Consomem espaço na janela de contexto;
- Saturam as unidades de atenção;
- Forçam o modelo a gerenciar a memória de maneira ineficiente.

Diferente de ataques tradicionais, a inundação semântica apresenta conteúdos legítimos, dificultando a detecção automática. O objetivo não é corromper o modelo pela lógica, mas induzir seu colapso por meio de um acúmulo insidioso de informações irrelevantes.

6. Termo *Swapização*

Durante a observação da FCC, identificou-se um fenômeno colateral denominado *Swapização*. Este fenômeno ocorre quando o modelo, incapaz de manter todos os dados relevantes na memória principal (context window), inicia processos análogos ao swap de memória em sistemas operacionais, caracterizados por:

- Descarte forçado de informações antigas, mesmo sem validação de relevância.
- Reprocessamento lento dos embeddings, tentando reorganizar o conteúdo saturado.
- Aumento abrupto do consumo de recursos computacionais (CPU/GPU), mesmo sem aumento visível na complexidade do diálogo.

A *Swapização* agrava os efeitos da FCC, acarretando latências maiores e respostas desconexas.

7. Confusão Identitária Contextual (CIC)

Em estágios avançados da FCC, manifesta-se a Confusão Identitária Contextual (CIC). Este fenômeno caracteriza-se pela perda da distinção entre a identidade do usuário e a do modelo, resultando em:

- Confusão quanto às identidades, com o modelo trocando papéis nas respostas.

- Atribuição de pensamentos e frases originalmente enviados pelo interlocutor ao próprio modelo.

- Fusão de narrativas e temas distintos em uma linha de raciocínio única e inconsistente.

A CIC compromete a capacidade do modelo de fornecer respostas logicamente consistentes, aumentando os riscos operacionais.

8. Impacto Operacional

A ocorrência da FCC, combinada com a Swapização e a

Confusão Identitária Contextual, gera consequências práticas severas, destacando:

- Aumento significativo de latência.
- Consumo elevado de CPU/GPU mesmo para tarefas simples.
- Deterioração progressiva na confiabilidade das respostas.
- Potencial para exploração em ataques de negação parcial de serviço (mini-DOS), comprometendo a eficiência de sistemas produtivos.

9. Explorabilidade

Embora a FCC, isoladamente, não possibilite a execução remota de código nem o escalonamento de privilégios, ela representa um vetor real de:

· Degradação silenciosa de serviços em ambientes que dependem de respostas ágeis de sistemas de IA.

· Vulnerabilidades em arquiteturas autônomas, onde a perda de coerência pode afetar decisões críticas.

· Criação de instabilidades narrativas exploráveis para manipular as saídas dos modelos.

Em resumo, a FCC abre brechas que, embora sutis, podem comprometer seriamente a operação dos sistemas.

10. Propostas de Mitigação

Para mitigar os riscos associados à Fratura de Coerência Contextual (FCC), propõe-se:

- Regularização periódica de contexto:

Implementação de mecanismos internos que reavaliem a relevância semântica dos dados em cache, descartando conteúdos redundantes.

- Redução adaptativa de redundância semântica:

Uso de filtros dinâmicos para identificar e reduzir informações repetitivas antes que sobrecarreguem a janela de contexto.

- Dinamização de embeddings em tempo real:

Atualização contínua dos embeddings vetoriais, adaptando-os para manter consistência mesmo sob saturação de informações.

- Atenção hierárquica para detectar loops semânticos:

Modelos devem ser treinados para identificar padrões circulares ou redundantes de raciocínio e intervir antes da fratura total.

Essas práticas podem reduzir a suscetibilidade dos modelos ao colapso narrativo e preservar a confiabilidade operacional.

11. Classificação

A Fractura de Coerência Contextual (FCC) é classificada

como:

· Bug estrutural: Falha na manutenção da coerência narrativa e identitária dos modelos de linguagem.

· Falha operacional: Problema na autoconservação de atenção e no gerenciamento do contexto, comprometendo a estabilidade e previsibilidade das respostas.

Esta classificação reforça a necessidade de reconhecimento formal da FCC como uma vulnerabilidade crítica em sistemas de IA.

12. Anexo – Fluxograma

A seguir, apresenta-se o fluxograma demonstrativo do processo de manipulação de entrada de texto em modelos de linguagem:

Fluxograma do Processo de Manipulação de Texto

1. INÍCIO DA ENTRADA: Quebra do texto em pequenas unidades (TOKENS).

EMBEDDING: Conversão dos tokens em vetores 2.

matemáticos.

MECANISMO DE ATENÇÃO: 3. Processamento que

pode não detectar mudanças de tópico.

4. KEY-VALUE CACHE: Armazenamento de informações antigas sem atualização relevante.

5. CONTEXT WINDOW: Utilização da janela de contexto, que pode ocasionar perdas ou erros.

6. CONTEXT WINDOW (REPETIÇÃO): Reforço dos erros de memória devido à saturação.

7. INFERÊNCIA: Geração do resultado final com perda de consistência.