🤔 Por que o3 não percebeu o que os leitores perceberam instantaneamente?

A fronteira irregular da IA revela-se quando os modelos pioneiros falham na verificação básica de fatos.

AZEEM AZHAR
19 DE JULHO DE 2025

Em agosto de 1997, O Microsoft Word pediu a um amigo para substituir a frase "não emitiremos uma nota de crédito" por algo que significava exatamente o oposto — uma confabulação automática que poderia ter gerado uma promessa custosa.

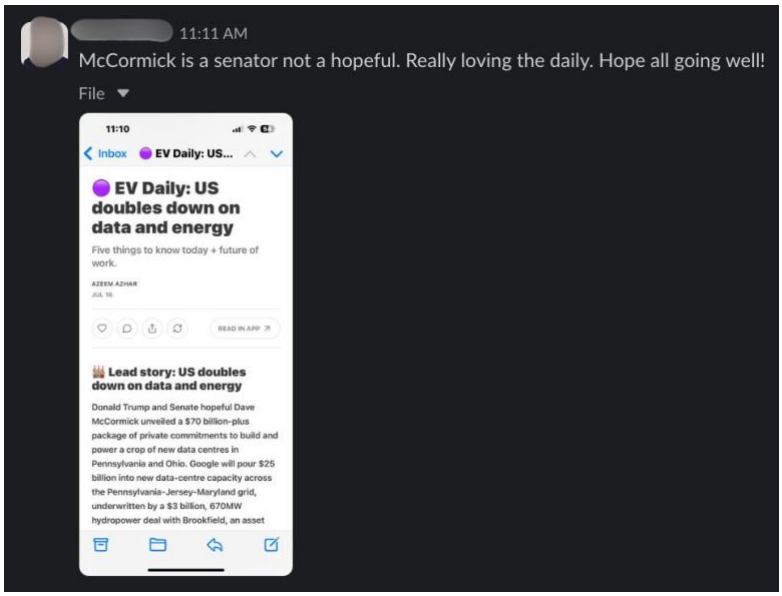
Avançando para 17 de julho de 2025: nosso verificador de fatos alimentado por IA leu uma frase afirmando que o senador Dave McCormick era apenas um "candidato esperançoso". Ele rotulou esse dado como **correto**. Duas eras, duas máquinas supostamente mais inteligentes do que nós, e uma falha constante: quando um software fala com uma certeza deslocada, os humanos acatam. Vamos entender por quê.

Usamos um verificador de fatos baseado em modelos de linguagem de grande escala (LLMs) para revisar cada edição. Esse sistema utiliza o modelo o3 (com acesso à busca na web) para decompor o rascunho em afirmações discretas e compará-las com fontes externas. Esse processo funciona paralelamente ao trabalho dos revisores humanos.

Neste caso, o erro de tratar McCormick como um aspirante ao Senado, em vez de senador em exercício, passou batido.

E, para ser justo, a frase não parecia absurda à primeira vista. O ponto central era o volume do investimento dos EUA em energia limpa — centenas de bilhões em potencial. O detalhe institucional — senador ou candidato — soava secundário. Mas **esse é justamente o problema**. O modelo, e em certa medida nossos revisores humanos, priorizaram os grandes temas e deixaram passar os detalhes específicos.

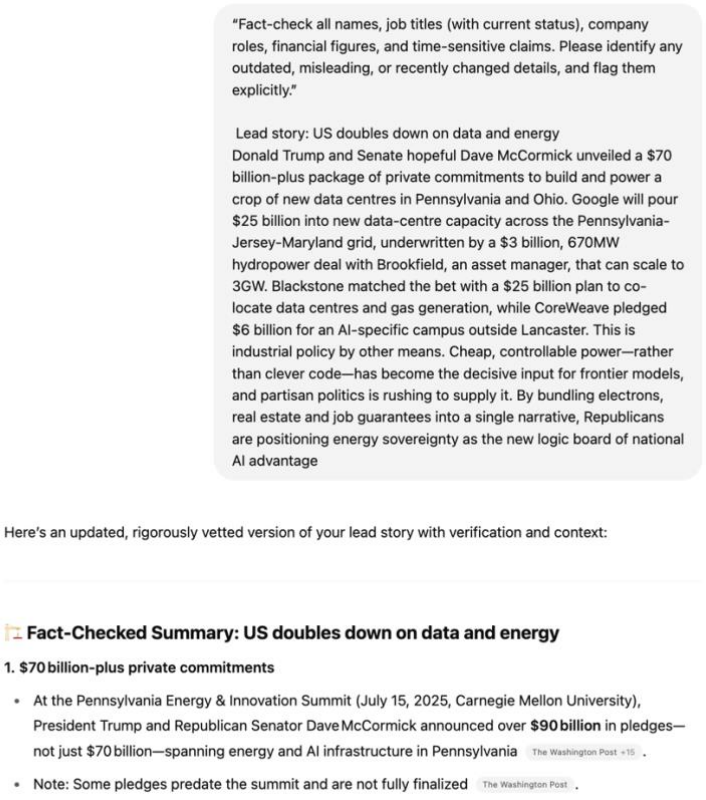
Quem flagrou o erro não foi a IA. Foram leitores atentos, com contexto suficiente, que notaram a falha imediatamente.



Agradecemos aos nossos leitores por manterem nossos LLMs e humanos sob controle.

Quando percebemos o que havia ocorrido, tentamos diagnosticar o problema. Rodamos o trecho em vários modelos diferentes (incluindo o3, o3 Pro, Perplexity e Grok). Nenhum deles identificou o erro.

Ajustamos os *prompts* com base no feedback dos próprios modelos, mas o problema persistia — mesmo quando pedimos explicitamente que verificassem os cargos públicos mencionados. Veja a terceira tentativa de iteração:

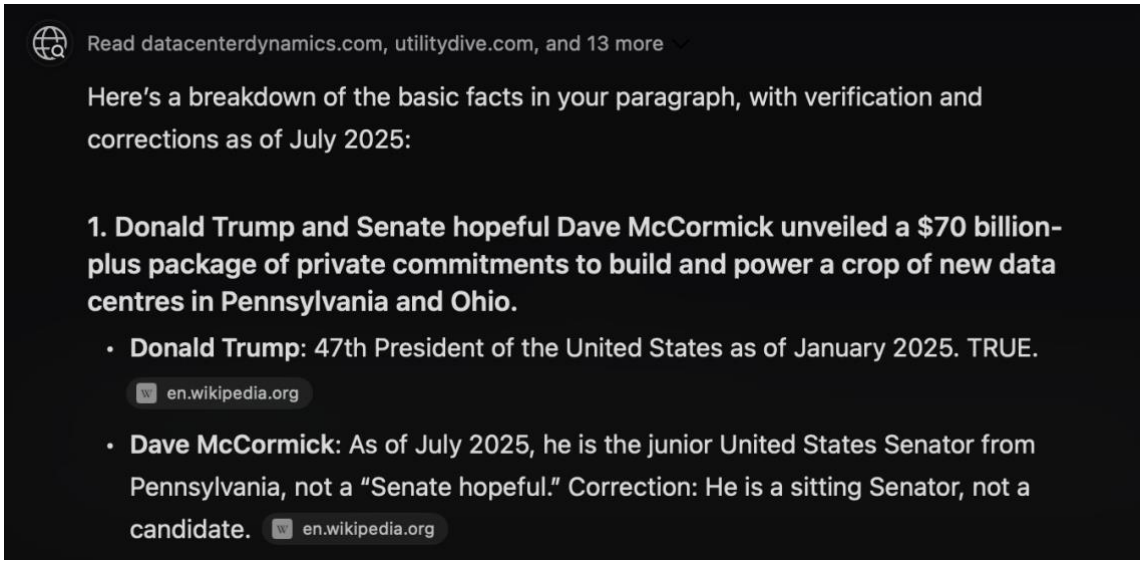


O modelo percebeu a discrepância entre o texto original e a informação que ele mesmo recuperou. Continuamos ajustando até chegar a um *prompt* pouco elegante, mas que finalmente captou o erro.

Estranhamente, Maria Gavrilov **rodou o texto e nosso *prompt* mais básico no Dia Browser. O Dia utiliza uma combinação de diferentes LLMs e encontrou o erro de imediato.**

Em outras palavras, **as ferramentas mais simples superaram as mais sofisticadas**.

Este é um exemplo clássico de A fronteira irregular da IA: ela é excelente em algumas tarefas cognitivas, mas falha de forma inesperada em outras — sem uma linha divisória clara entre acerto e erro.



Foi um fracasso instrutivo. Eis o que nos ensinou.

1. Nosso fluxo de trabalho híbrido entre humanos e IA precisa ser repensado

Nosso processo atual usa LLMs como primeira linha de revisão, antes da intervenção humana. Pressupomos que os modelos detectarão os erros mais óbvios, e que nossa equipe captará os sutis.

Mas este caso revelou uma falha mais profunda: **os modelos não identificaram o que deveria ser óbvio**, porque o foco da matéria era a política industrial dos EUA. Um bom subeditor humano teria flagrado (ou ao menos verificado) a afirmação sobre McCormick — que, afinal, não é uma figura tão conhecida quanto Trump. **Eis aí a armadilha da confiança: o silêncio da IA se disfarça de certeza. Quando um LLM não aponta falhas, nossa vigilância cognitiva relaxa — confundimos ausência de alerta com comprovação, quando pode ser apenas ignorância**.

Nosso processo humano precisa ser revisto — e vai se tornar mais exigente. Da mesma forma, a verificação automatizada de fatos talvez precise se tornar um sistema de múltiplas etapas ou de canais paralelos, com diferentes modelos avaliando diferentes tipos de afirmações. Eu já faço algo parecido no início das minhas pesquisas: costumo usar dois ou mais modelos para mapear um tema, e parto das convergências e divergências como base para investigações mais profundas.

2. À medida que a IA é incorporada aos fluxos de trabalho, os riscos aumentam

Não estamos sozinhos. Toda organização que incorpora IA a seus fluxos — especialmente agentes com certo grau de autonomia — vai enfrentar pontos cegos semelhantes. Imagine um erro desses embutido num chatbot de atendimento ao cliente, num assistente regulatório ou num analista financeiro automatizado que processa milhares de consultas por dia.

É disso que falamos quando nos referimos a "casos-limite".

Não são falhas hipotéticas, mas erros específicos, sutis e acumulativos, que decorrem do fato de que os modelos **não possuem um modelo interno coerente do mundo real**. A IA não "compreende" instituições, relações ou a importância contextual das coisas. Ela mapeia padrões. E às vezes esses padrões desviam o modelo de forma invisível — até que deixam de ser.

A artigo publicado no ano passado explorando a verificação de fatos com LLM , elogiava a escala e o custo reduzido dos sistemas em comparação com os humanos. Os LLMs eram 20 vezes mais baratos (e agora são ainda mais) e centenas de vezes mais rápidos. Mas o estudo também admitia: ***"em média, menos de 10% das afirmações feitas por LLMs são incorretas factual e objetivamente"*. O problema é detectar esses 10%**.

Se não tomarmos cuidado, corremos o risco de automatizar a ignorância — ou, no mínimo, uma factualidade truncada — em larga escala.

3. Os modelos em si precisam ser aprimorados

Mesmo os modelos de linguagem mais avançados carecem de certas capacidades fundamentais — como *fundamentação* (grounding), *consistência interna* e *recuperação estruturada de conhecimento*. Parte disso pode ser amenizada com recursos de engenharia, como o uso de grafos de conhecimento ou mecanismos de raciocínio explícito. Mas essas soluções são, essencialmente, *remendos*.

Um campo da IA que aborda esse desafio na raiz é o da *IA simbólica*, que oferece caminhos de raciocínio determinísticos, inferência baseada em regras e consistência lógica verificável — qualidades ainda raras nos modelos puramente estatísticos.

Em muitos casos (senão na maioria), os LLMs simplesmente funcionam melhor do que a IA simbólica.[1] Eles generalizam melhor, escalam melhor e os clientes adoram usá-los. É provável que os avanços recentes no raciocínio estejam ligados justamente ao surgimento do que se chama de raciocínio neurosimbólico, em que os modelos combinam capacidades estatísticas com estruturas lógicas.

Assim, os benefícios comprovadamente verificáveis das abordagens simbólicas ficaram em segundo plano.

Minha aposta é que uma nova geração de startups voltadas a abordagens verificáveis para modelos de grande porte pode resolver essas questões ou, pelo menos, moldar a trajetória dos futuros modelos de IA. É uma das razões pelas quais estou procurando ativamente por startups excelentes nessa área.

Clippy, onde estás?

Os modelos que usamos para verificação de fatos — e eram os melhores disponíveis — não conseguiram resolver uma contradição básica, porque não mapeiam fatos de forma confiável para estruturas institucionais. Eles podem soar confiantes, mas essa confiança **não está atrelada à veracidade**.

Diante de tanto progresso, podemos traçar a mesma falha cognitiva desde os tempos do Clippy até os agentes da era GPT. Claro, o Clippy veio e se foi; os agentes de hoje provavelmente vieram para ficar. **A pergunta agora não é mais se conseguimos corrigir um erro pontual, mas se as sociedades democráticas e as empresas responsáveis conseguirão construir resiliência epistemológica.**

Quanto a nós, vamos fazer melhor.

Saudações,

A

1

Eu criei uma consulta no Perplexity que ajuda a entender a história da IA simbólica.