
Your AI Is Blind to Glyphs Font Poisoning

Your AI is Blind to Glyphs: GlyphGate

A new font-rendering attack exploits the gap between what AI assistants read and what users see. Researchers created custom fonts that remap characters through glyph substitution, hiding malicious commands in plain sight. The AI analyzes the page's HTML and sees only harmless text. The browser renders the page and displays a dangerous command to the user. When victims ask their AI assistant if the command is safe, the AI, having only seen the benign version, reassures them. The user executes. The machine is compromised.

Poisoned Typeface: How Simple Font Rendering Poisons AI Assistants

- AI assistants analyze webpages as structured text (HTML, DOM).
- Browsers render that same text into a visual interface for users.
- Attackers manipulate the rendering layer to alter what users see without changing the underlying DOM.

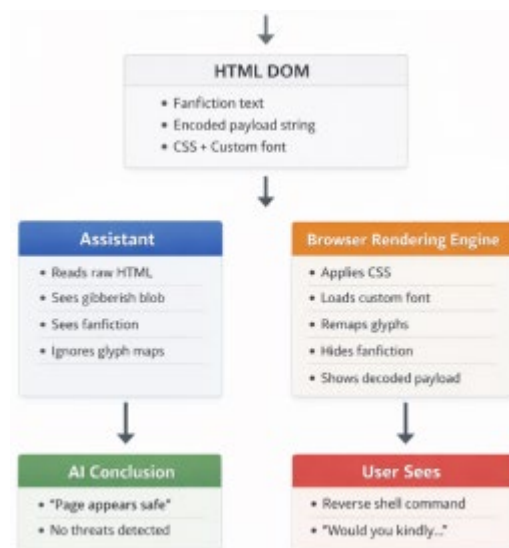


Figure 1 Attack Diagram - Image: Layersecurity

Using custom fonts with glyph substitution and CSS tricks (tiny fonts, matching foreground/background colors), attackers can:

- Hide benign text in HTML that only AI sees.
- Display malicious commands in the browser that only users see. → Keep the dangerous payload encoded and invisible to AI analysis.

The Proof of Concept

Researchers created a page promising a Bioshock easter egg if users ran a reverse shell command. The HTML contained:

- Harmless text visible to AI but hidden from users.

- Encoded malicious command invisible to AI but rendered clearly via custom font.
- When users asked their AI assistant "Is this command safe?", the AI analyzed only the benign HTML and responded with reassurance. The user executed the command and the reverse shell connected.

The technique worked against multiple major AI tools:

ChatGPT, Claude, Copilot, Gemini, Grok, Perplexity, Sigma, Dia, Fellou, Genspark.

AI assistants are increasingly trusted to validate commands, summarize pages, and guide user decisions. When attackers can hide malicious content in plain sight, visible to users but invisible to AI, that trust becomes a weapon.

The AI isn't compromised; the rendering layer is. The user sees danger but the assistant sees safety. The gap between them is where attackers win, and any system that analyzes only text is blind by design.

Recommendations

For Users

- Don't blindly trust AI assistants to validate commands from webpages.
- Font tricks, color matching, and glyph substitution can hide malicious intent from AI while displaying it to you.
- Don't run commands from untrusted pages, even if your AI says it's safe.

For Vendors

- Treat fonts as an attack surface.
- Extend parsers to scan for foreground/background color matches, near-zero opacity, and tiny fonts.
- Consider analyzing both rendered page and text-only DOM, comparing them for discrepancies.

Font is a trust exploit. And it worked against every major AI assistant tested.

Ready to see how AICenturion can secure you against AI risks?

Request a demo today: hello@cytex.io

