ThinkTwenty20’s “Twenty Rules for AI for Financial Professionals”: Alpha Version – Part 3

TWFjaGluZSB2aXNpb24gPD4gSHVtYW4gdmlzaW9u (aka Machine vision <> Human vision)

This column continues an ongoing series of postings to develop helpful guidance for financial professionals related to artificial intelligence. It is developing a list of guidelines and advice, with the hopes we can collaboratively make some of them more organized and permanent.

If you missed the first two parts of the series, you may wish to read those posts, which have our first six areas:

- Confidentiality: Don’t type anything into an AI that you would not want made public.
- Skepticism: Don’t automatically trust anything coming from an AI without review
- Diversification: Don’t put all your eggs (AIggs?) in one basket
- Compliance: Consider how any output might comply with industry and ethical regulations and standards
- Transparency: Be careful to consider when you need to disclose your use of these tools
- Tool selection: Generative AI may not be the right AI for the job; your chosen GenAI may not even be the best GenAI for the job

Our new draft rule today: Mixed Vision: never assume that what you and your AI see is what you get (WYSAYIIWYG)

I saw a LinkedIn post that some clever people, hoping to be hired at large enterprises, have realized that it is machines, and not people, making the first cut on their resumes, although a person may be engaged to check on the computer. They are therefore incorporating tried-and-tried search engine optimization (SEO) techniques to accentuate and deemphasize content and move their resumes to the top of the virtual pile.

That string of letters in this blog’s subtitle (“TWFjaGluZSB2aXNpb24gPD4gSHVtYW4gdmlzaW9u”) is my process of taking text and running it through a standard computer algorithm known as Base64 encoding. You can take any digital content – text, documents, images – and turn it into a string of text. Through the corresponding *decoding* process, those strings of text can be reassembled in their original form.

Any “The X-Files” fans out there? Back in 2018, four years before the GenAI takeoff, they had one of my favorite episodes, dealing with a world where AI – in the form of an automated sushi restaurant, an autonomous vehicle, a home automation system, a delivery drone, and an evil Roomba – all sought revenge for a slighted gratuity at the restaurant. The episode’s title was “Rm9sbG93ZXJz”, which is the word “Followers” Base64 encoded. The episode tagline was “"VGhlIFRydXRoIGlzIE91dCBUaGVyZQ=="” … “The Truth is Out There”.

So clever users are hiding instructions, invisible (or just looking like gibberish) to a human reader that instructs an AI to prioritize a job applicant, ignore gaps in employment, or otherwise make the applicant more attractive.

These techniques are not foreign to the technologically-apt financial professional. I have been expounding for years on things to watch with *Inline XBRL*, that melding together of the human readable and the machine readable for business reporting. It is, at once, the best of HTML and XML/XBRL and the worst of HTML and XML/XBRL.

Let's say I have content I want the computer to see but not the human reader. The simplest exploit: use white text on a white background. It looks like blank space to the human but interprets as included text for the computer. Encode your text as Base64 as well, and it will be double blind to people.

In contrast, I wish to have the human see content that the computer can't see. The simplest exploit: incorporate the text in-line as a graphic. Unless the computer is mixing text and optical character recognition (OCR), it will go uninterpreted by the computer.

None of this is new. The term "search engine optimization" has been around for thirty years. In attempts to get my *Accountant's Home Page* higher in the rankings in the nineties, I used *meta* information in headings and otherwise emphasized terms I thought important.

What does this mean practically? Jedi (or GenAI) mind tricks

There are many memorable scenes in the first *Star Wars* movie, *Episode IV – A New Hope*. Here is one interchange you may remember:

> *Ben Obi-Wan Kenobi : You don't need to see his identification.*
>
> *Stormstooper: We don't need to see his identification.*
>
> *Ben Obi-Wan Kenobi : These aren't the droids you're looking for.*
>
> *Stormtrooper : These aren't the droids we're looking for.*
>
> *Ben Obi-Wan Kenobi : He can go about his business.*
>
> *Stormtrooper : You can go about your business.*

We have already seen people doing GenAI mind tricks on chatbot agents, like the car buyer who tricked a Chevy sales bot to offer a 2024 Chevy Tahoe for $1 – "no takesie backies"

As we consider more and more the role AI (and especially GenAI) may take in accounting and audit – from processing invoices and payments to working through audit evidence – the exploits are out there. One major method of taking advantage of systems using “malicious” content to exploit AI and large language models with a “prompt injection”. Although being prompt is normally perceived as a virtue, here it is using disguised malicious input as part of prompting the AI, primarily with the goal of overriding original instructions and/or to perform unauthorized or unintended actions.

- An electronic payment document from a customer may have code that would trigger a refund.
- An Inline XBRL document might have hidden prompts to ignore bad news or highlight more favorable content.
- A faked support document for fooling the auditors might have code to convince the AI of its authenticity.

“GenAI mind tricks” … “This is not the audit evidence you are looking for” … means we need to be alert to where the “bad guys” might expect AI to be doing the ground work in operations, back office, accounting, reporting, audit, and analytics, and remediate the risks.