All.Net Analyst Report and Newsletter

Welcome to our Analyst Report and Newsletter

Problems with using steganography to self-identify AI-generated content

When Anthropic announced that it was using selection of next words to allow its AI output to be identified as large language model (LLM) produced content, it revealed a major difference between AI-generated and well thought out human generated content.

Words have meaning

The basic scheme can be thought of as choosing from among synonyms for word selection based not on the probability of the next specific word appearing, but rather modifying that probability based on a signature method to signal origin of the sequence. So instead of picking “likely”, it might pick “probable”, with the idea that this conveys the same content but that the latter is the longer (more characters) of the major other options at this point in the sentence, and that the sequence of longer / shorter / middle / shorter / shorter could later be analyzed as a sequence of [10-00] where longer=’1’, shorter=’0’, and middle=’-’. As long as these words are reasonably similar in context, it should make no different to the reader and allow the LLM to be identified. With enough word selections, the string of {1,-,0} values will be extremely unlikely to appear from a human and therefore be an effective signature of the AI generating it.

Sometimes words have two meanings[1]

Of course part of the problem here is that actual human crafted words have actual meanings in sentences and differentiate different concepts. It seems likely that probable will be a ready replacement for likely in almost any sentence. It seems probable that likely will be a ready replacement for probable in almost any sentence. So let’s look at some synonyms in sentences, and as an example, an AI response to a requests for synonyms for “bad” came back with this:

> Synonyms for "bad" depend on what you mean. You can explore full lists on the Merriam-Webster Thesaurus or Thesaurus.com. **Poor Quality**: Awful Poor Terrible Lousy Substandard **Harmful**: Dangerous Damaging Detrimental Hurtful Unhealthy **Wrong or Immoral**: Evil Wicked Wrong Sinful Unethical **Rotten Food**: Spoiled Rotten Rancid Putrid Sour

So suppose the sense of the term “bad” being substituted is “Rotten Food” and the AI picks the synonym “Sour” for its middle length. Is “this lemon tasted bad” the same as “this lemon tasted sour”? I don’t think so. Rancid is not the same as putrid or sour or rotten or spoiled. But how does the LLM tell the difference? By likelihood of appearance in similar sentences. And “this lemon tastes sour” being the most likely selection, modified by the staganographic need for a different word effectively must lead to a different word being selected.

> If you think this is hurtful as examples go, or sinful as a way to express the putrid nature of this sort of selection method, try to figure out what I really meant to say here.

But to be clear, this is not only a problem associated with the use of steganography for forming covert channels in text for LLM identification. It points out a more basic problem.

1 From “Stairway To Heaven”, the song by Led Zeplin

LLM generation methods misuse terms all the time

In writing a recent proposal, an LLM was used to generate text,

> which a human who knew what was being expressed then edited,
>
> > the LLM was then asked to update the proposal based on these edits,
> >
> > > and it promptly (no joke here but it could be one) changed the specific improved word selections with worse ones (from the perspective of the human seeking to convey actual meaning).

This is almost certainly reflective of a shared experience that most LLM users who wish to be reasonably precise in their expressions have also encountered.

> *This almost certainly reflects a shared experience among LLM users who strive for reasonable precision in their expressions.*

Rewritten by AI at my request “Rewrite this sentence: …” the AI does a better job of expressing most of the content, but misses the “most LLM users” replacing it with “among” which does not express “most” (more than half). Asked to rewrite again and again, select words change, “almost certainly” becomes “likely”, “clarity and precision” replaces “precision”, and so forth. And over time the selection process produced more and more variations.

The underlying problem here has to do with meaning, which is something LLMs do not actually understand in context.

It may just sound picky

I’m almost certain that many of my readers will find this to be just picking nits, and to some extent it is just that. But having said so, I think it’s important to note that in evaluating proposals reviewers often have to differentiate between nearly equivalent alternatives and it’s the nits that differentiate the picks.

> *I’m almost certain that many of my readers will regard this as mere nitpicking, and to some extent, it is. But it’s worth remembering that when evaluating proposals, reviewers often have to distinguish between nearly equivalent alternatives - and it’s precisely those seemingly minor details that can ultimately determine the choice.*

Rewritten by AI it just misses the poetic nature of my expression, which is the joy embedded in my writing. And sucking the joy out of life is somehow not my objective here.

Conclusions

Steganographic content embedding for LLM identification is problematic:

- It changes the meaning of the text (not that there was meaning in the first place)
- It counters the very thing that made the LLM output look like human writing
- It’s likely trivial to bypass, especially for short texts (not detailed herein)

But I think the best part of it being announced is that it clarifies to an even larger extent the lack of actual meaningfulness in the selection of words and expressions by LLMs. Having been on the receiving end of copy editors for years, I can tell you that their readings and rewritings often change the meaning of my expressions, so I often reject the changes.

