Second Submission

## What Can Regulation Do?

It can and has turned what was a perilous journey on an aircraft a hundred years ago into a highly reliable service, with about forty million flights per year and an amazing safety record, and nonstop flights of halfway around the world being promised. This came about from unrelenting regulatory effort – airworthiness certification, better understanding of materials and failure modes, redundancy of control systems, pilot health checks – every aspect of commercial airline operation in every country, to the point where even structural cracks are managed and not immediately reacted to.

An aircraft has a solid basis in physics – if the lift exceeds the weight and the thrust exceeds the drag, it will fly. Regulation was always going to work. Sometimes the regulatory shield is pierced by duplicity, but only at great cost.

## What Can’t Regulation Do?

It can’t turn a sow’s ear into a silk purse.

And LLMs are very much a sow’s ear.

LLMs rely on a probability that the English language does not have. It eschews knowledge of parts of speech, grammar, and meaning, with the hope that it can induce such knowledge by inference from a large body of text. But text is not a simple line of words, where every word has only one meaning. A word can have half a dozen different parts of speech, and many meanings, up to seventy. The POS/meaning depends on the words around it, so finding the words you are looking for in a piece of text is no guarantee that the text is talking about what you are looking for.

There is the notion that a large number of nodes in the LLM data set (576 billion for Bard!) will lead to accuracy. The problem with that is what a natural language can describe is vastly larger, and someone can utter a few words – say 10 – and communicate something that has never been said in the history of the world. “Cold fusion” was a good example – it was nonsense, but those two words energised the science establishment to thoroughly debunk it. Which brings up the problem of how up-to-date the text in the LLM is, and how do new ideas get introduced, and under whose control?

Natural language is about transmitting useful information, not whether the next word has high probability in a large set of textual data. “The house is on fire” has very low probability but, when it is used, ranks as information of vital importance.

Here is an example of Bing’s use of Generative AI (LLM).

My young daughter is terrified of dogs.  
Which parks are safe?

The words it had to work with – My, young, daughter, terrified, dogs, parks, safe.

There was a message of condolence about “terrified”, but that didn’t mean it had understood anything.

It could find some text linking dog and park, so it provided a list of dog-friendly parks – exactly the wrong answer.

Running the same input in Google Search gave 64 million hits, together with Google’s Must include



Looking at some of the hits, it becomes clear that plucking words from the search string is not a good idea. A child can be terrified of a dog, but small dogs can be terrified of large dogs, and owners of dogs can be terrified that their dog will be attacked by other dogs. Without the ability to shape the search and manage synonyms (“frightened”, “afraid”), the answer using probability is likely to be about something else altogether. If the person using the tool has to get into the statistics being used for individual words, the tool becomes unworkable.

Even simple words can be treacherous – he turned on the light, he turned on a dime, the dog turned on its owner – the word “on” switching from being an adverb to a preposition with different meanings. This is handled for us by our Unconscious Mind, so we are not aware of how much language processing we are doing, and don’t immediately question something which does no language processing.

Could Generative AI be made more accurate? Not without killing its “creativity”. If it were only permitted to return part or all of one piece of text or an amalgam of text in a limited domain, it would be more accurate, but still not perfect – even in a single domain, you have the problem of multiple meanings – cervical cancer or cervical vertebra, he had a cold shower in the morning, he had a cold in the afternoon. When it attempts to assemble pieces of text from different domains, it is cobbling things together without any understanding of what they mean, which can easily end up a “crazy quilt”. A person’s Unconscious Mind is used to untangling garbled messages, so it does its best, assuming the message has been created by another human mind, with knowledge of the structure of language, parts of speech, grammar and meaning.

What is the risk of using such a tool? Obviously, any application, such as health advice, that demands high accuracy is out of the question. It was once thought that searching huge databases would yield useful results, but comorbidities mean that many cases are distinctive, requiring much knowledge about the patient, not probabilities based on millions of patients.

Are there many mid-level applications that would benefit? Can the language model be shrunk so the risk of going off-point is small? It would need examination on a case-by-case basis. Bringing the person’s profile into it – concession card holder, etc. – may help to reduce the errors, but the user may be asking about people in general, not themselves in particular – it is rather hard to escape the necessity of understanding what words mean.

What is the risk when there seems to be no risk? This is the risk of groupthink – everyone gets the party line, even when the statistical difference between the dominant message and a competing message is only a few percent. Until the only message is the dominant message. Unthinking acceptance will likely be the unfortunate default.

## Problems With Collaboration

Let’s assume for a moment that it would be possible to develop regulations that would be effective in improving the reliability of the LLM.

One suggestion is that a group of people be assembled, with diverse backgrounds – a lawyer, an ethicist, one or more public servants, a software engineer, one or more scientists. This group will have very little common technical vocabulary and, given that humans have a Four Pieces Limit, no-one will understand the regulation output in toto. Examples are an economist and an epidemiologist talking past each other during Covid, or lawyers and software engineers with Robodebt. We think it would take at least several months for the non-software engineers to become au fait with the operation of LLMs.

The ethicist and the public servants may be hoping for a system that could be described as “safe, trustworthy, loyal, with an ethical backbone”. The LLM has no element in it which could respond to such words, or their meanings spelt out in more detail, with the software engineer reminding the others that you can change the text, and you can introduce a few hacks, but otherwise there isn’t anything to be changed, that is how an LLM works.

During the collaboration phase, we would recommend the use of an AGI (Artificial General Intelligence) tool, which does “understand” the vagaries of the English language in all its frustrating glory. The regulations can be created using complex text, with the particular meaning of each word accessible with a click, and clumpings of words, and wordgroups (examples of such objects are given in the supporting documents), so the structure of the regulation can be seen, discussed with much greater understanding, and approved, while the software engineer reiterates – “You can only change the text in the model – it doesn’t understand parts of speech, grammar, or meanings – it will still use probability”.

We would expect the regulations to end up limiting the application of the technology, with any life-critical application strictly verboten, but this in itself presents a problem. How do you know the output of the LLM breaches the regulation, without it being read by a person who is sufficiently expert in the areas of technicalities that the output touches on? Again, this would require the use of an AGI tool to analyse the output text, which would be much slower in operation than the LLM, and destroy the usefulness of it (if you have to check the output using something that is slow, you might as well use the output of the slow system). Could the AGI system only be used where the LLM is cobbling together different pieces of text, while allowing it to run unchecked while it stays within one piece of text? – possible, but unlikely.