

RAG time

Generative AI can track down and define many words – but not all of them.
Michael Farrell discusses a new system for getting the music flowing again

In a survey conducted early last year on how professional translators use generative artificial intelligence (GenAI) in their workflows, 72.8 per cent of those who reported using it said they did so in order to find the definitions of words. Indeed, if you type ‘What does photosynthesis mean?’ into any GenAI tool, you will virtually always get a pretty accurate definition.

The problem is, though, that translators do not normally search for the definitions of words that everyone learns in school. Instead, they are far more likely to look up highly technical terminology or obscure, antiquated words. And that is where things start to go wrong.

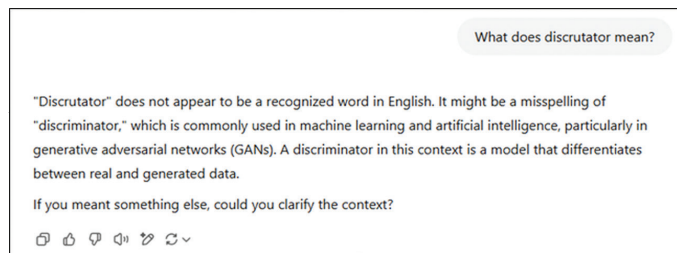


Figure 1: ChatGPT’s answer to what *discrutator* means

The large language models (LLMs) which GenAI draws on are pre-trained transformers. They are not information retrieval systems; they do not simply take your input and look it up in a database in the way an online dictionary does. Instead, they generate responses by calculating the most probable next element in a sequence on the basis of the patterns they identify in the vast amounts of data they have been trained on. (To be absolutely accurate, they randomly select one of the most probable next elements in the sequence based on their relative probability.)

‘But surely,’ you might say, ‘if these models have access to so much data, the word *discrutator* must have ended up in there somewhere.’ I imagine it very probably did, but these tools don’t look for the most probable next *word*: they look for the most probable next *element*. These elements are called ‘tokens’, and are generally smaller than words and do not necessarily correspond to syllables. A helpful rule of thumb is that one token normally consists of about four characters of text.

RAG combines language generation with information retrieval, allowing the model to refer to external sources when needed

Figure 2 shows that GPT-4o actually breaks down *dis-cr-ut-ator* into four tokens. Given the rarity of the word in the training data, it is easy to imagine that the odds of GenAI reconstructing these four tokens into the original full word *discrutator* are very steep; it is a highly unusual pattern of tokens. As a result, the model is more likely to suggest a more common, simpler sequence, like *dis-criminator* (which consists of only two tokens).

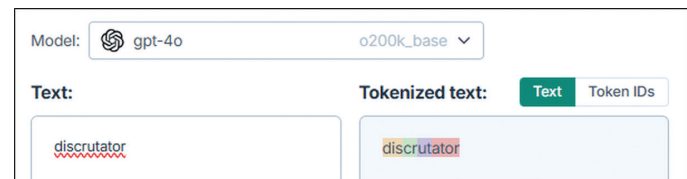


Figure 2: *Discrutator* in GPT tokeniser playground (<https://gptforwork.com/tools/tokenizer>)

Maple Leaf Rag

This is where RAG might prove useful. RAG stands for ‘retrieval-augmented generation’. The principle is that it improves GenAI output by incorporating relevant information from external databases or documents when generating a reply. In other words, it combines language generation with information retrieval, allowing the model to refer to external sources when needed.

RAG was designed to try to prevent situations like the one which occurred in 2022, when an Air Canada chatbot provided a passenger with false information about bereavement fares after his grandmother’s death.

The chatbot incorrectly told him he could retroactively apply for a reduced bereavement rate within 90 days of purchasing his tickets. However, when he applied for the refund, Air Canada denied his request, stating that the company’s policy did not allow refunds on completed journeys. The passenger subsequently filed a claim with the Civil Resolution Tribunal of British Columbia, which ruled in his favour, awarding him damages, and criticised Air Canada for attempting to distance itself from the chatbot’s *hallucination*.

Although Air Canada never made an official statement to confirm it, it is widely believed that their chatbot probably did not incorporate RAG technology.

Essentially, it is as if we had included the most relevant snippets from the reference documentation in the GenAI prompt, along

with the instruction: ‘Refer to the attached material when generating your reply.’ The retrieved information effectively becomes part of the conversation the chatbot continues when looking for the most likely next token, and therefore anything in there strongly influences the calculations. As a result, the likelihood of the reply being out of line with the contents of the documentation becomes significantly lower.

RAG was proposed by Patrick Lewis et al way back in 2020 but has only become more widely available recently as more platforms have integrated it into their models. Today, RAG has already been implemented in several freely available GenAI tools, including Gemini and ChatGPT. To activate it, the prompt often needs to indicate that external information should be referenced. Sometimes this happens automatically because of the nature of the prompt; at other times you can explicitly ask the chatbot to cite the sources it draws on. Once you do this, the model retrieves relevant information, often performing a web search via integrated APIs from commercial search engines or using more advanced technologies like vector searches to pull data from pre-indexed documents or databases.

Hitting the right keys

To see what difference it makes, let’s add the phrase ‘cite your sources’ to our *discrutator* query in ChatGPT-4o (Figure 3).

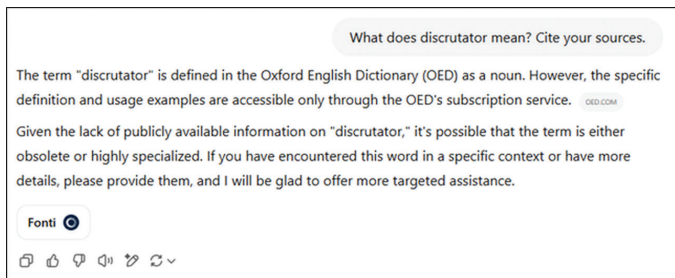


Figure 3: *Discrutator* with RAG

This does at least acknowledge that the word exists, but I am still none the wiser. To make matters worse, I could have found a more useful answer with a simple web search (Figure 4).

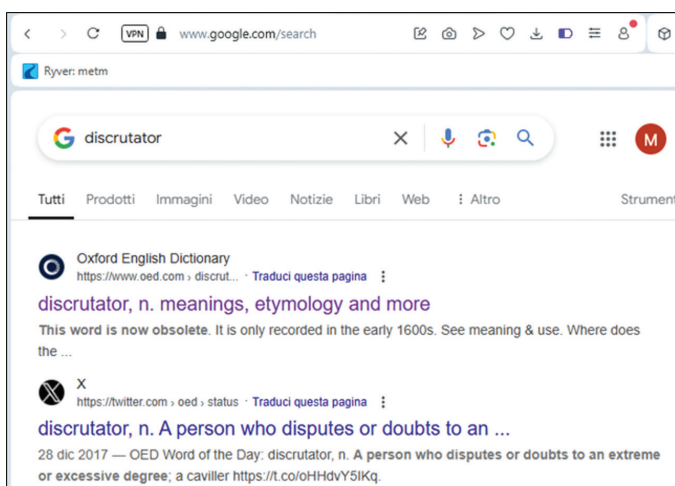


Figure 4: Web search to the rescue

Why isn’t it being more helpful? The problem – with all due respect to 72.8 per cent of my fellow GenAI users – is that finding

the meaning of words is not exactly a chatbot’s greatest strength. But...we already have highly efficient tools, called online dictionaries, which work wonderfully and produce far less greenhouse gas and dry up fewer rivers. And might I even dare mention those hard copy objects sitting on our bookshelves collecting dust and serving as doorstops in various parts of the house? Even ChatGPT would seem to be recommending a subscription to the *OED*, judging by the last example.

To be fair, the same survey also revealed that 84.8 per cent of GenAI-using translators find it useful for looking up ‘words in context’, which plays to GenAI’s strength in pattern recognition and generally works well even when terms are misused (Figure 5).

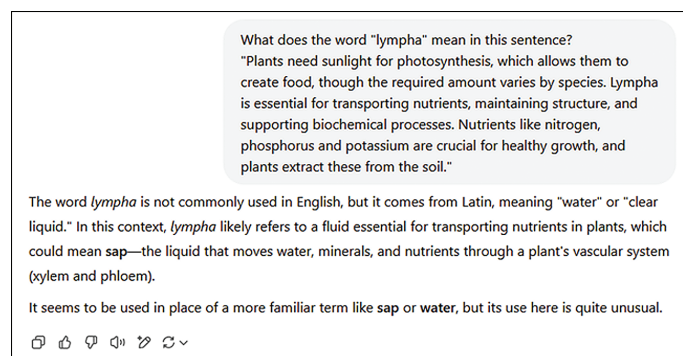


Figure 5: Finding meaning ‘in context’

The Entertainer

When it comes to translation itself, RAG comes into its own because it can be extremely useful. Figure 6 shows the translation of a seemingly simple sentence from Italian to English without the use of RAG.

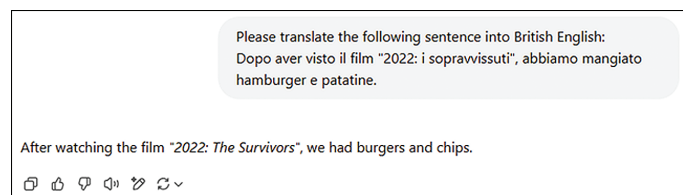


Figure 6: Translation from Italian into British English without RAG

At first glance, it might look like a fairly good translation. The problem, however, is that there is no such film as *2022: The Survivors*. This is another case where the Italian film distributors decided to change the title entirely for the release in Italy.

Let’s see what happens when we use RAG (Figure 7). This translation is spot on and has handled the film title correctly. (The only problem is that, if you have ever seen the film, the thought of stuffing yourself with burgers afterwards has probably left you feeling thoroughly nauseated.)

What about vetting its sources? I cannot find a publicly available OpenAI document that explicitly states that ChatGPT verifies its sources when providing information. So I asked ChatGPT-4o directly whether it vets its sources. It explained that, in practice, when generating responses that include external information, it aims to reference reputable sources, like government websites, academic institutions, well-established news outlets and recognised industry experts. Similarly, Gemini responded that it does not vet each individual source in the same way a human might. However, it added that Google’s systems are

Please translate the following sentence into British English:
 Dopo aver visto il film "2022: i sopravvissuti", abbiamo mangiato hamburger e patatine.
 Cite your sources.

The British English translation of your sentence is:
 "After watching the film "Soylent Green", we ate burgers and chips."

Explanation of Translation Choices:

1. Film Title: "2022: i sopravvissuti" is the Italian title of the 1973 film *Soylent Green*, so I used its original title.
2. "After watching" vs. "After having watched": "After watching" is more natural and commonly used in British English.
3. "Burgers and chips":
 - "Burgers" is a widely used short form of "hamburgers" in British English.
 - "Chips" refers to thick-cut fried potatoes, whereas "fries" (used in American English) would typically mean thin-cut ones.

Sources:

- *Oxford English Dictionary* (for British English usage)
- *Collins English Dictionary* (for British English food terminology)
- *IMDb* (for confirming the original English title of the film)

Figure 7: *Soylent Green* and burgers thanks to RAG

designed to prioritise reliable, trustworthy information. (I was not able to find a publicly available statement supporting this claim.)

Back with the baton

What I personally like about RAG is that not only does it put the translator back in the loop, it places us firmly back on the piano seat – or even on the conductor’s podium.

By providing access to the relevant information, it allows us to fact-check everything, which is what all responsible translators have always done. In this way, RAG does not just enhance our workflow; it strengthens our role as critical evaluators in the translation process.



Michael Farrell is primarily a freelance translator and transcreator. Over the years he has acquired experience in the cultural tourism field and in transcreating advertising copy and press releases, chiefly for the promotion of technology products. Being a keen amateur cook, he has also translated texts on Italian cuisine. In addition, he is an untenured lecturer in post-editing, machine translation and computer tools for translators at IULM University, Milan, Italy; the developer of the terminology search tool IntelliWebSearch; a qualified member of the Italian Association of Translators and Interpreters (AITI); and a member of Mediterranean Editors and Translators (MET) Council.



Diversification strategies for professional translators and interpreters



Who is this course for?

This course is aimed at freelancers and self-employed professionals, specifically translators and interpreters, who are looking to strengthen their business by implementing effective diversification strategies. Whether you’re seeking to expand your service offerings, enter new markets, or reduce dependency on a single income source, this course will equip you with practical techniques to diversify your freelance business and achieve sustainable growth.

What will you learn?

This course consists of five comprehensive modules designed to help you successfully diversify your freelance business. You will learn how to identify new income streams, assess market opportunities, and strategically expand your service offerings. By the end of the course, you’ll be equipped with the tools to diversify your business model, minimise risk, and ensure long-term financial stability and growth.

- **Module 1:** What is diversification, and what does it mean for us?
- **Module 2:** The toolbox approach to diversification
- **Module 3:** The preparation toolbox: Laying the groundwork for diversification
- **Module 4:** The opportunity toolbox: Identify, evaluate and exploit diversification opportunities
- **Module 5:** Operational tools: Operating as a diversified LSP

How will you learn?

The course includes a variety of teaching techniques, such as tools, case studies, activities, and a knowledge assessment, to ensure that you come away with actionable steps to implement straightaway. It is self-paced, allowing you to work through it at a speed that suits you.

How long is the course?

This course takes approximately one hour to complete.

Course leader: Ken Paver



Ken Paver is a translator and interpreter with over 35 years in the nuclear, power generation, and technology sectors, including three years in-house at BNFL and subsequent freelance work. His skills and experience include translating commercial, technical, and legal texts from French, German, and Dutch into English, as well as copywriting and editing in these industries.

He has worked as a translator, interpreter and interpreter team leader for the International Atomic Energy Agency and the World Association of Nuclear Operators and provides training for interpreters on nuclear technology and terminology.

Ken’s academic background includes lecturing on translation and interpreting at Aston University and York University. He is a former vice chair of ITI, with past roles on the ITI board and admissions committee, as well as CIPOL Council. He has also served as chair of the editorial board for *The Linguist* journal.