May 2023

Hybrid Search:
Matching both meaning and words

An overview of hybrid search technology and how it enhances site search

---

Raffle Search is a world-leading search engine powered by advanced deep-learning vector search technology. It's designed to accurately understand and interpret users' search queries, enabling it to provide highly accurate, relevant, and personalized search results.

Raffle has achieved a leadership position in this field by investing in research and development into a fast deep-learning methodology that works without training data for new customers (zero-shot) and improves over time. Raffle's hybrid search technology combines deep semantic search with traditional keyword-based search to allow users to find what they are looking for based on context and word association.

Raffle AI-based Insights tool helps the customer to identify usage trends and questions the users cannot answer, making it easy for the customer to improve their business's content accordingly.

This insights paper covers Raffle's hybrid search technology, gives background on the most significant developments in search, and details how Raffle's technology fits into the most vital current technology trends.

https://raffle.ai | hello@raffle.ai | VAT no:. DK39747189
US:.+1 929-207-4717 | UK: +44 20 3966 5755 | Benelux: +31 202442319 | Nordics: +45 4290 9050

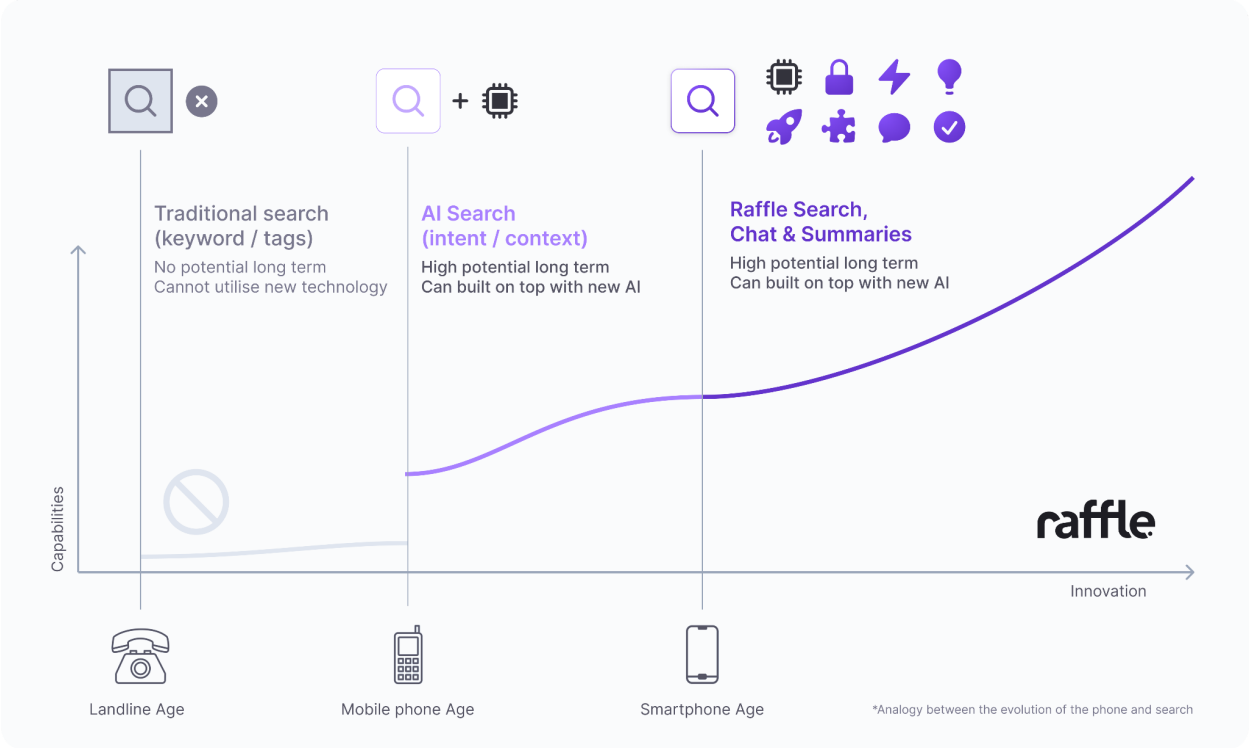
Background

Artificial intelligence is progressing at an astounding pace.

This opens possibilities for improving and even disrupting many of the tasks we perform with computers. However, it also comes with a lot of uncertainty and risks that we need to address. Raffle is an AI company at the forefront of this development, putting us in a good position to build highly performant and safe AI systems. In this series of three insights papers, we will address the following important topics:

1. Deep learning powered hybrid search - the foundation of highly accurate search that understands what you are looking for.
2. Large language models - the intelligent conversational interface to computers
3. The 4 pillars of Search for Businesses - Security, GDPR/HIPAA Compliance, Sustainability, and Disability Compliance (WASP2)

The development of search can be viewed as analogous to the development of the phone; the technology enables a whole range of application beyond the initial use case:



This is the first paper in the series on hybrid search. Raffle Search uses deep learning to find the correct answer for both simple and complex user queries. Our customers can use our AI-powered Insights tools to understand and improve their own data sources concerning what the users are searching for. This gives a

revolutionized search experience, enabling users to find the information they need faster and more efficiently than ever before.

Limitations of Traditional Search Technology

Search technology has one simple goal: to find what you're looking for.

To achieve this goal, search engines must undertake two crucial steps:

1. Understand the search query and
2. rank the results according to relevance.

However, it is important to note that the first step is critical and needs to be highly efficient for the other step to be effective.

For example, if a person types in the term "jaguar" on a search engine. The search engine may assume that the user is searching for information related to the animal species. However, the user might actually be interested in learning about the luxury car brand. We need the search engine to understand the user's intent. The simple query "jaguar" is not providing us with any context to distinguish, whereas "electric jaguar" or "jaguar natural habitat" do. The query "jaguar top speed" is still ambiguous on whether we are referring to the animal or the car.

Modern search engines should be able to understand the intent from the context and give you the right answer.

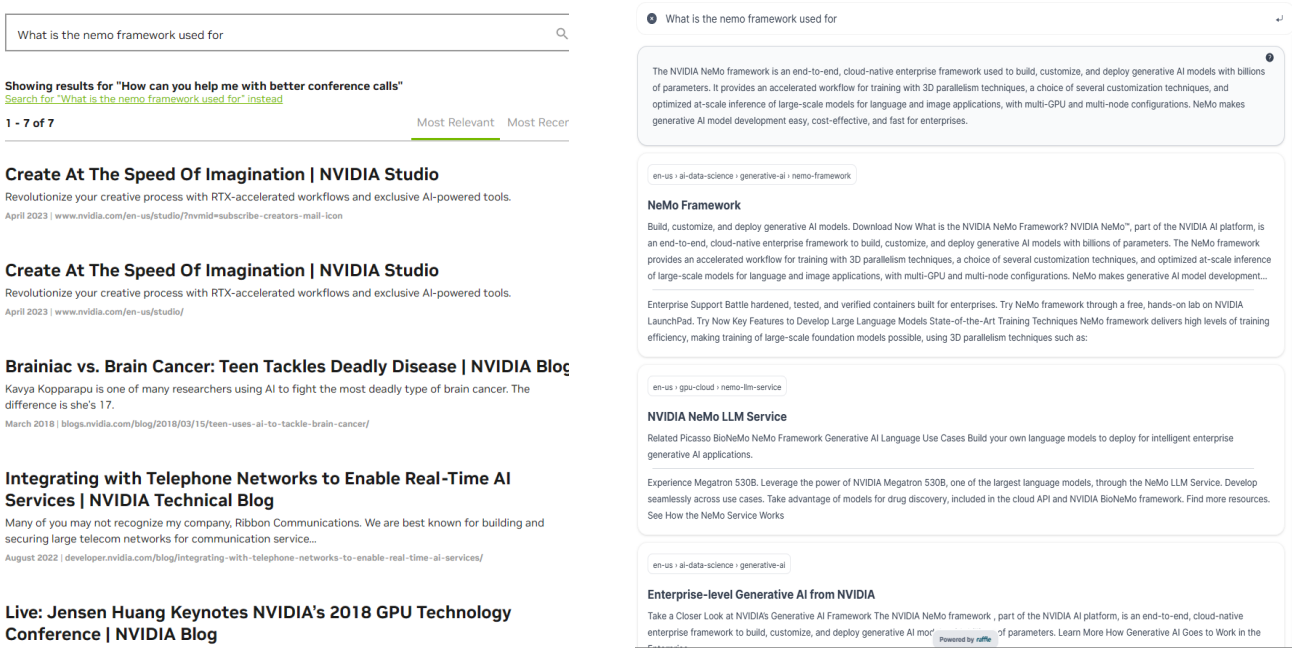
Example of Traditional Search versus Raffle Search

How we did it:

1. Raffle selected a random large company and used our crawler to get the content of their website.
2. We can now set up Raffle hybrid search for their website and compare it to their existing traditional search on the same queries. The existing search is shown on the left and Raffle search on the right in the example below.

This is Raffle Search Zero shot performance - out of the box, day one,- and it clearly shows the difference in understanding the intent and the context of the search string.

This is just one example, we will show many more towards the end of the document.



Before giving more real world examples, we will make a deep dive into the underlying technology using "jaguar" example to explain how traditional search fails and explain how vectors are used to represent meaning (semantics).

Bag of words

Traditional search works by matching words between your search query and the document you are searching in. This means that word order does not matter; instead, we use a so-called bag of words representation. Take the following example with four (short) documents:

D1: Where the jaguar lives
D2: Jaguar seen in city
D3: Jaguar is available in many colors
D4: Jaguar on sale

Humans immediately understand from the context that documents 1 and 2 are about the animal and 3 and 4 are about the car. However, the traditional search will not understand, so if we see the question

"Where can I buy a blue jaguar?",

We know that this is a question about a car - jaguars are not blue, and most of us don't keep jaguars as pets - so documents 3 and 4 should be the most relevant. In the bag of words approach we will simply

count word overlap, and here, document 1 wins with a word overlap of two: “Where” and “jaguar.” The three other documents overlap with just the word “jaguar.” In traditional search, we use a document-term matrix to represent the bag of words for the knowledge source. This can be illustrated by our example like this:

		available	city	colors	in	jaguar	lives	many	on	sale	seen	the	where	Score
D1	Where the jaguar lives					1	1					1	1	2
D2	Jaguar seen in city		1		1	1					1			1
D3	Jaguar available in many colors	1		1	1	1		1						1
D4	Jaguar on sale					1			1	1				1
Q	**Where** can I buy a blue **jaguar**					1								

		available	city	colors	in	jaguar	lives	many	on	sale	seen	the	where	Score
D1	Where the jaguar lives					1	1					1	1	2
D2	Jaguar seen in city		1		1	1					1			1
D3	Jaguar available in many colors	1		1	1	1		1						1
D4	Jaguar on sale					1			1	1				1
Q	**Where** can I buy a blue **jaguar**					1							1	

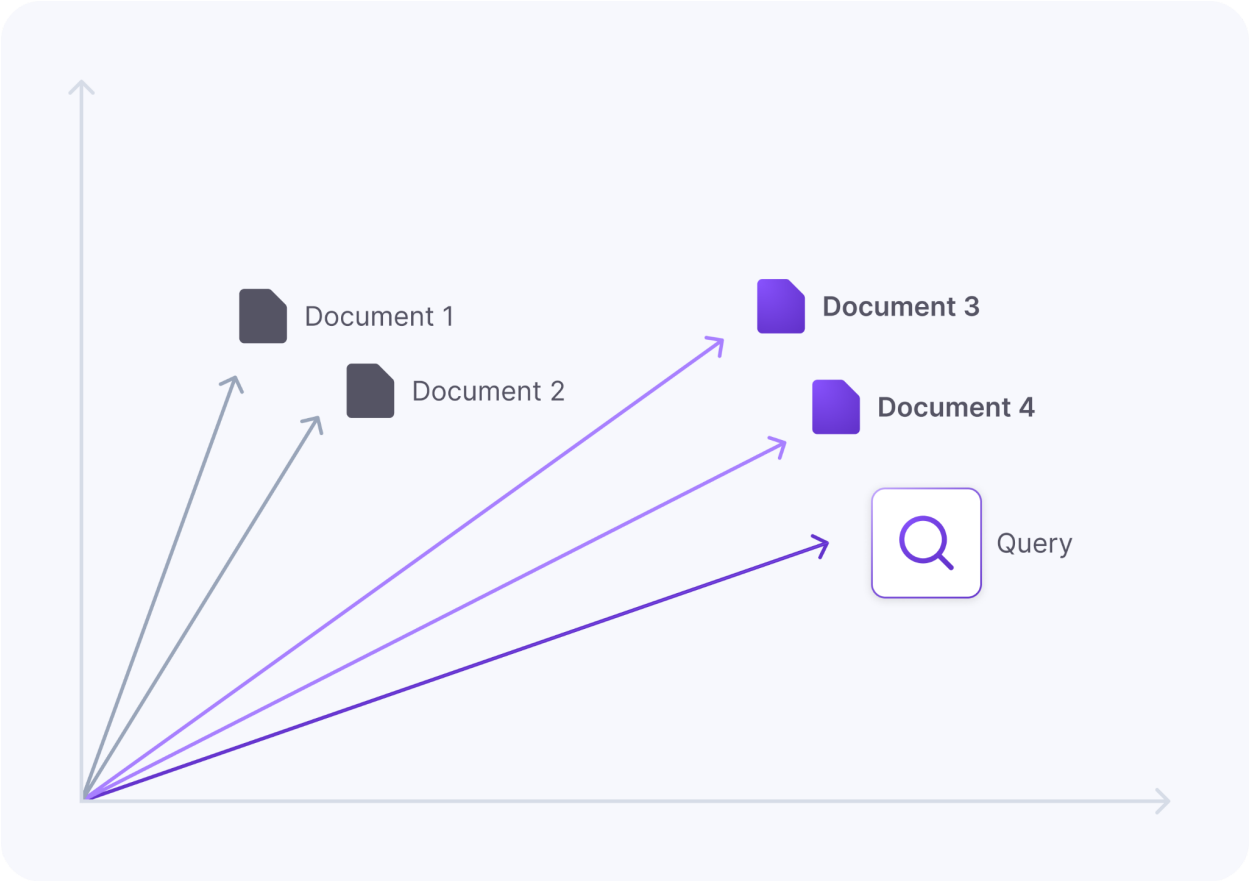
While a bag-of-words search is fast due to the abundance of zeros in the matrix and efficient querying methods, it is challenged when it comes to semantics as this example illustrates.

Vector search - understanding intent with deep learning

The previous example demonstrated how keyword searches can be ineffective when the query does not match the exact wording. This limitation is not just theoretical; in the following real-life examples, we will illustrate how this can have significant consequences and why traditional word-based search engines are no longer adequate.

In vector search, documents and queries are mapped to vectors with a language model neural network. These vectors should reflect the intent and semantics of the content. We can visualize the documents and the query from above as two-dimensional vectors:

https://raffle.ai | hello@raffle.ai | VAT no:. DK39747189
US:.+1 929-207-4717 | UK: +44 20 3966 5755 | Benelux: +31 202442319 | Nordics: +45 4290 9050



The example shows that the vectors capture the semantics: Since Documents 1 and 2 are closest in meaning, their vectors should also be more similar. The Query is closest in meaning to Documents 3 and 4 and vectors capture that. We get the most relevant answer if we choose the document with a vector closest to our query vector. So the recipe for vector search is, in principle, simple:

1. **Setting up the document vectors**: Map all your documents to vectors and store them and
2. **Search query ranking step**: Map your query to a vector, calculate the vector similarity with all document vectors, and return the top ones as your search result.

In reality, these vectors are high dimensional because semantics and intent are too complex to quantify in just a few dimensions. So this begs the following two questions:

1. **Training:** How do we construct the neural network method for mapping text to semantic vectors?
2. **Serving:** How to get the answers in real time when we use large neural networks and have many vectors (documents) to rank?

To answer this, we need to dig a bit deeper into Raffle's core deep learning model optimization technology and vector search serving.

Raffle Search - the core

Raffle has, since 2018, built an AI-powered search consisting of the following three vector search optimization steps and a final hybrid search step. It is important to add, that this exact order and process makes AI the actual core of Raffle, and is one of the reasons, that Raffle heavily outperforms other search engines:

1. **Pretrained language models**. The language model maps the text input to vectors. We use language-specific and multi-language pre-trained *masked* language models like BERT as the building block. When Google introduced BERT in its search engine in 2019, it was considered the largest improvement in search technology for a long time.
2. **Fine-tuning on labeled data**. The language model transforms the text into vector representations that we correlate (for example, using the dot product) to define a semantic similarity between a query and a piece of text of the knowledge source. We can transform this into a scoring function to optimize a training set of questions with answers in the knowledge source. Since we are starting from pre-trained models, we only need little labeled data.
3. **Question generation.** We use a second generative language model to generate questions for the user's knowledge base. This limits the need for labeled data. Raffle has a US patent application on question generation technology.
4. **Hybrid search.** We combine vector search (points 1. through 3.) with traditional keyword search to get a complete solution that excels in semantic and word-matching information retrieval.

So do all these additional efforts pay off in practice? The short answer is yes; it makes the difference between having a search that users will stop using after the first experience and one that they will want to return to again and again. A study done by Forrester's research team suggests that a product search experience that causes a significant level of frustration leads to an astonishing 68% rate of churn and burn, which we believe to be unacceptable.

Benchmarking Raffle Search

In our internal benchmarking, traditional search is typically capped at a performance below 40% correct answers in the top 3. In other words, in less than 4 out of 10 queries, the correct result will appear in the top 3 returned search results. Raffle Search has a zero-shot performance of 44% and rises to more than 70% top 3 accuracies with the introduction of training data relevant to the knowledge source.

Said differently, users will find what they are looking for more than twice as often with Raffle Search compared to traditional keyword-based search.

Traditional search companies such as Algolia and Elastic have naturally started to include vector search in their offering. While vector search is a useful tool for semantic search, it is essential to note that it is not

sufficient to guarantee optimal performance. Therefore, it is crucial to exercise caution when assuming that implementing vector search alone will automatically result in effective semantic search capabilities.

We have created a challenging internal benchmark with data pooled from many different customers. It is made to test our zero-shot performance: We train models on one set of customers and test the models on data from other customers. The four steps have the following top 3 performances:

1. BERT by itself has a poor performance at 13.1%
2. Supervised fine-tuning on internal data gives 37.4% - barely beating an optimized keyword search algorithm at 36.7%.
3. Question generation brings the performance up to 43.7%.
4. Hybrid search, where vector search is combined with the optimized word-based search, finally brings the performance to 44.8%, which is 8% better than word-based search by itself.

These numbers indicate what the technology delivers without further supervised fine-tuning. Raffle search will always be deployed with a certain amount of fine-tuning so the performance will start at at least 70% top 3 accuracy. In other words, the user will get the correct answer in at least 7 out of 10 queries.

Vector search is still a relatively new technology, and in our experience, commercial offerings are by no means close to the performance of Raffle's solution for two reasons:

1. The vector model is not optimized. It is either a BERT model out of the box or barely fine-tuned. So the vector model is only really useful in a hybrid search solution where the keyword search does the bulk of the work. In Raffle Search, it is the other way around: vector search contributes far most to the ranking, and our keyword search is only used for the very specific situations where the user knows exactly what document is sought.
2. Poor support for further training. Our internal benchmarking shows that a significant boost to top 3 performance up to 70-90% of top 3 accuracies can be achieved by labeling data specific to the customer's knowledge source. The Raffle App provides tools and insights - currently not available in other products - to identify unanswered user questions (clustered knowledge gaps) and the ability to improve data and models.

These are crucial features to probe when considering the vector search solution.

Fast neural networks and vector similarity look-up

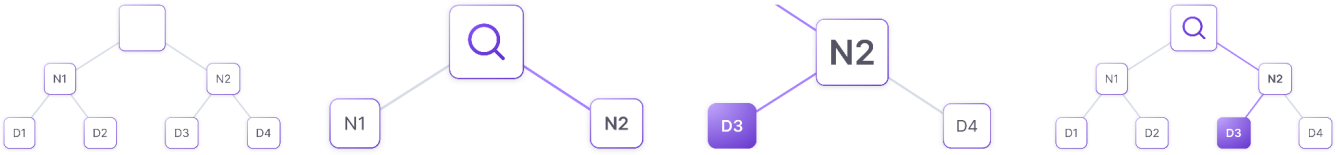
Traditional search excels in fast look-up in the document-term matrix. Luckily, fast methods have also been developed for vector search. This means that vector search can be served in real time even for very large knowledge sources.

A vector index is the set of vectors for all the documents in the knowledge source. A tree structure is built for vectors such that similar vectors are close in the tree. To explain the key concept at a high level, we

can make a tree for our simple example with the four documents. The documents sit as leaves in the tree and we construct additional node vectors sitting inside the tree.

To find the best match to our query, we start at the root of the tree and find the best match among the two children of the root. We keep on doing this until we find the best match as illustrated in the figure. For this small example, we need to perform four vector comparisons to find the best match. This is the same as if we had just matched the query with each of the documents in turn. However, once we go to larger datasets, the difference between going through the tree and going through all the documents one by one will be dramatic: logarithm(D) versus D, where D is the number of documents.

The tree structure is the core technology enabling fast retrieval. But that is not all: modern tree-based data structures have many more features than explained above such as real-time tree updating algorithms and approximate top-k retrieval. Raffle employs all of these.



Start
The vector tree is given and a query comes in

Step1
Compare Q with N1 and N2 and choose the most similar (N2)

Step2
Comparing documents and choosing the the 3rd over the 4th

Result
Raffle selected the right document for you.

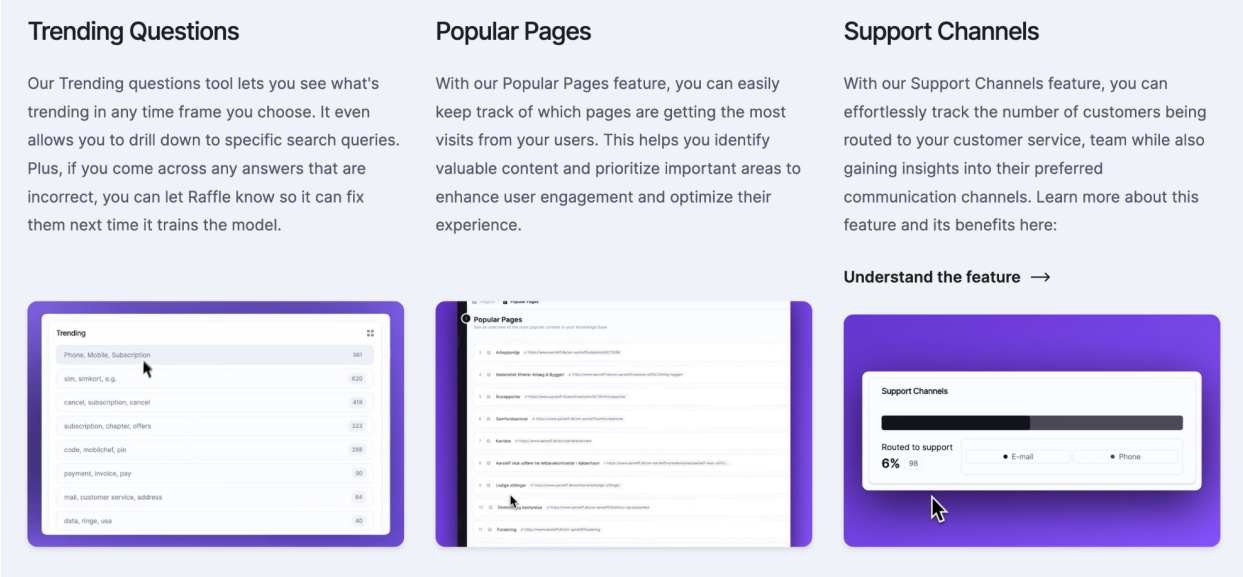
Raffle search versus traditional search

Traditional search engines like Elastic, Solr, Algolia, Asure, Coveo, and Cludo have been the option for many companies needing to manage their content and data search needs. Their search engines operate on keyword-based search, string search, or mixed keyword search algorithms. Keyword-based search engines collect documents based on individual words, while String Search searches across numerous words, which generally form a sentence. A mixed keyword search is when a myriad of keywords are searched that may or may not form a sentence. While these search methods have been effective in the past, their limitations are becoming increasingly evident with the increasing complexity of search queries and data.

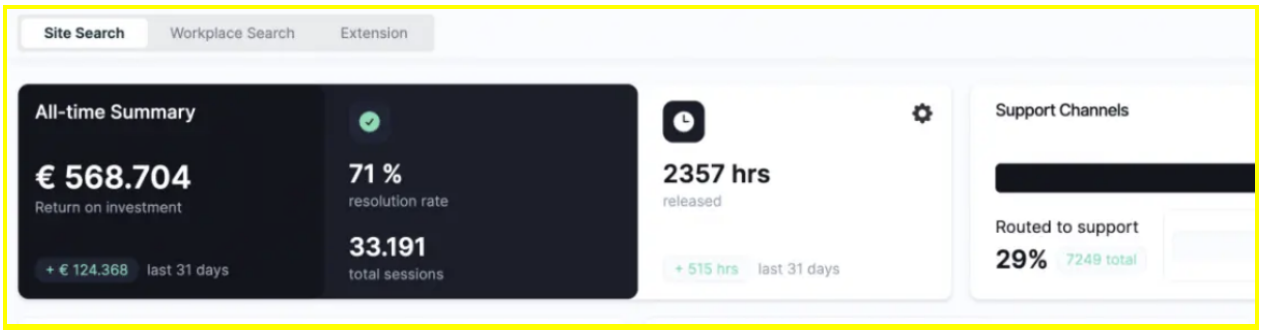
Instead, we now need to combine the best of the two worlds, however with the emphasis and core in the semantic search, as the search will require a much higher level of understanding in general including real conversational AI and AI assistants going forward.

Raffle Insights

Data analytics is essential to gaining deep insights into customers' needs and wants, predicting upcoming trends, and making better, data-driven decisions. Raffle's Behavior Analytics solution helps businesses obtain valuable insights into customers' needs and wants. With Trending Questions, Trending Topics, and Popular Pages and support channel tools, businesses can gain a deep understanding of their customer's preferences and tailor their marketing efforts and customer service accordingly.



Raffle Insights is a valuable tool for businesses, providing real-time data on resolution rates, productivity, time spent, and cost savings. This data helps companies to predict trends and take proactive measures to improve the overall customer journey and interactions within the organization. With the help of this data, businesses can make better, data-driven decisions about their operations, resulting in improved outcomes and competitive advantage.



https://raffle.ai | hello@raffle.ai | VAT no:. DK39747189
US:.+1 929-207-4717 | UK: +44 20 3966 5755 | Benelux: +31 202442319 | Nordics: +45 4290 9050

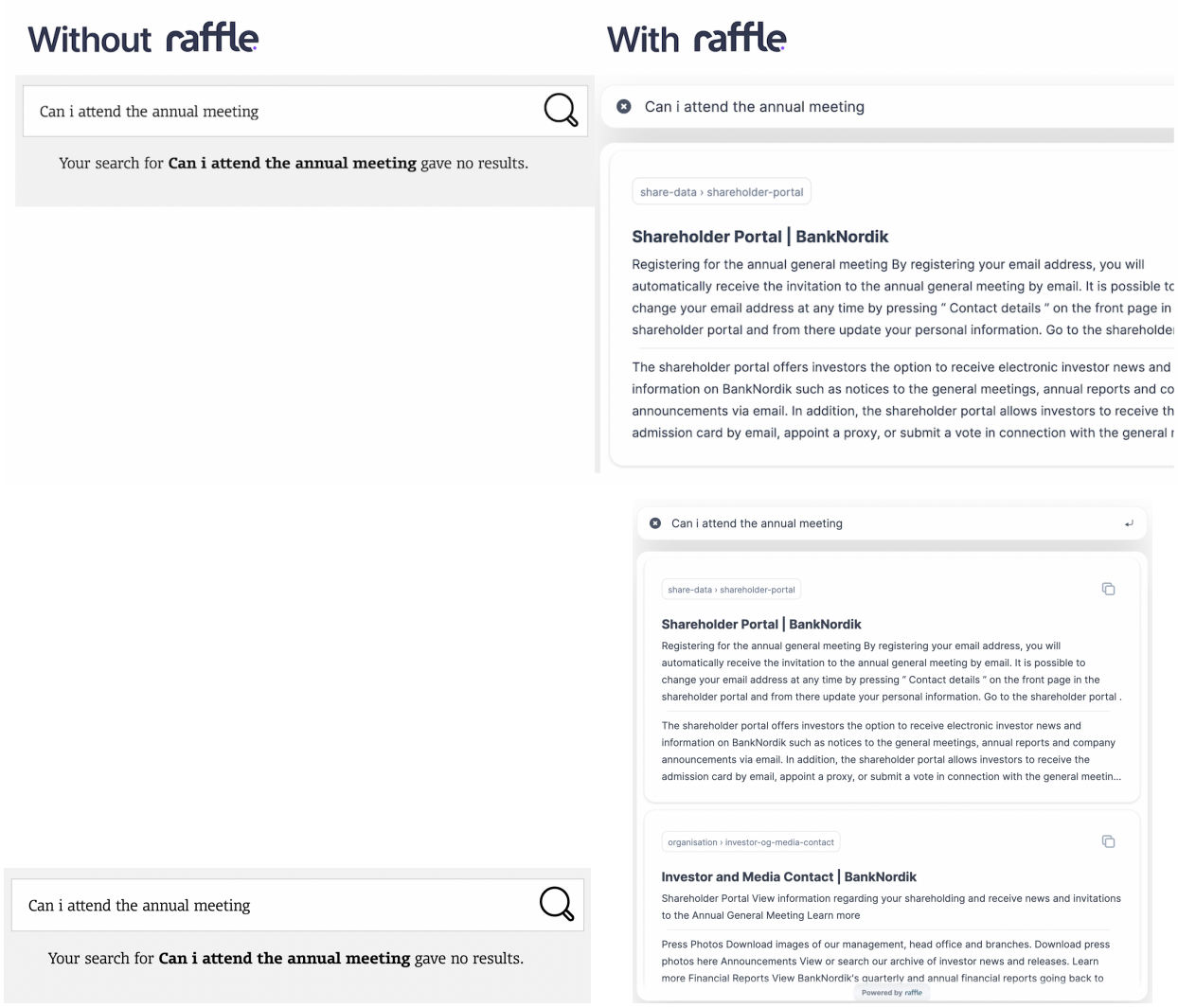
One of the key benefits of Raffles Behavior Analytics is its ability to help businesses identify potential issues with their products or services quickly. For example, suppose customers frequently ask how to install a particular product through the company's customer service portal on their website. In that case, it can indicate an issue with the product's instructions or installation process. By using Raffles Behavior Analytics, businesses can detect such problems in real time and take corrective action promptly, enhancing the customer experience and preventing negative impacts on sales.

Another benefit of Raffles Behavior Analytics is its ability to help businesses identify their website's most popular pages. By analyzing customer behavior, companies can determine which pages generate the most engagement and focus their time and resources on optimizing them to drive more engagement and sales, helping businesses save time and resources while boosting their bottom line.

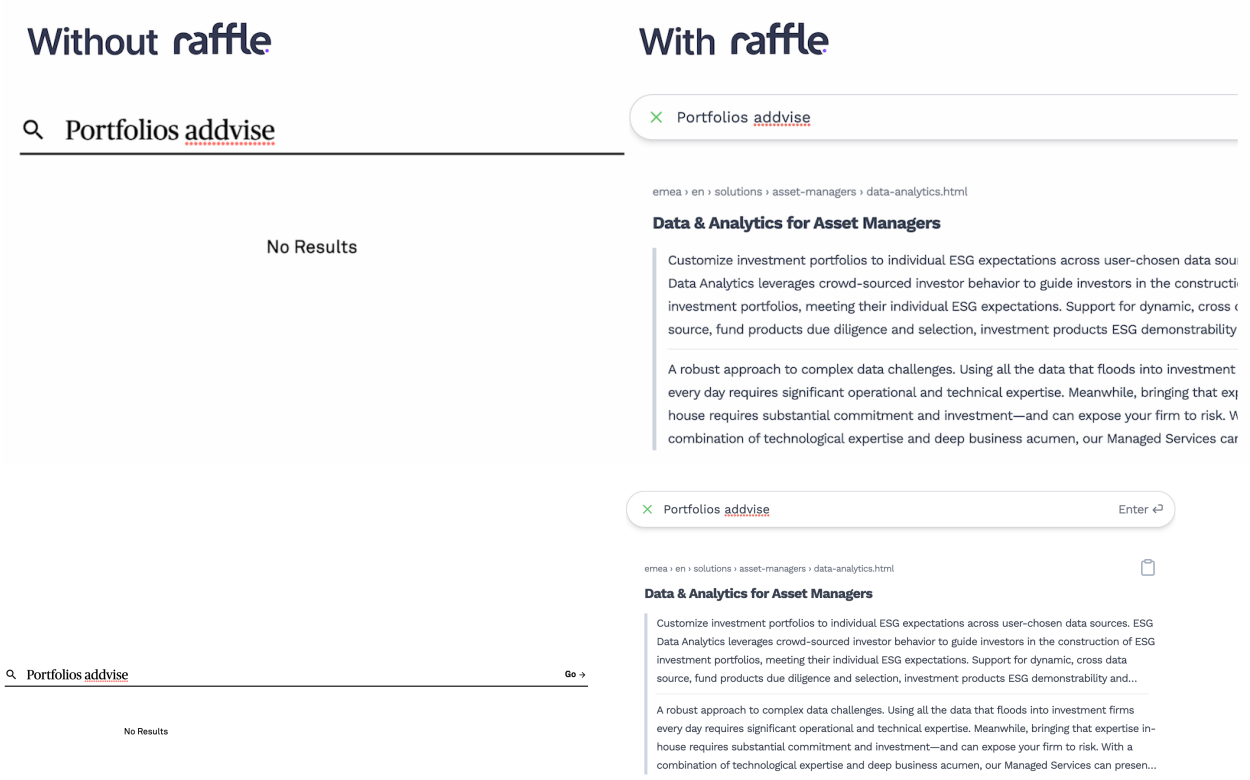
FILLING THE GAP

Raffle uses a sophisticated clustering algorithm to identify groups of searches that share a joint inquiry, highlighting the most popular and pressing questions. By doing so, it can determine if corresponding answers are already available to these queries and alert your team if there needs to be more information on a specific topic. Raffle identifies these knowledge gaps (Missing information that users need.) and creates temporary answers, allotting businesses the time they need to replace the temporary answers with permanent ones—preventing customers from leaving empty-handed and allowing customer service agents to work more efficiently, saving time and resources and ensuring that information is always up-to-date and accurate for your customers - and in any language.

In English below is an example of 'zero results' with Traditional Search on a search string, as often seen, although the answers are clearly in the data source/index.

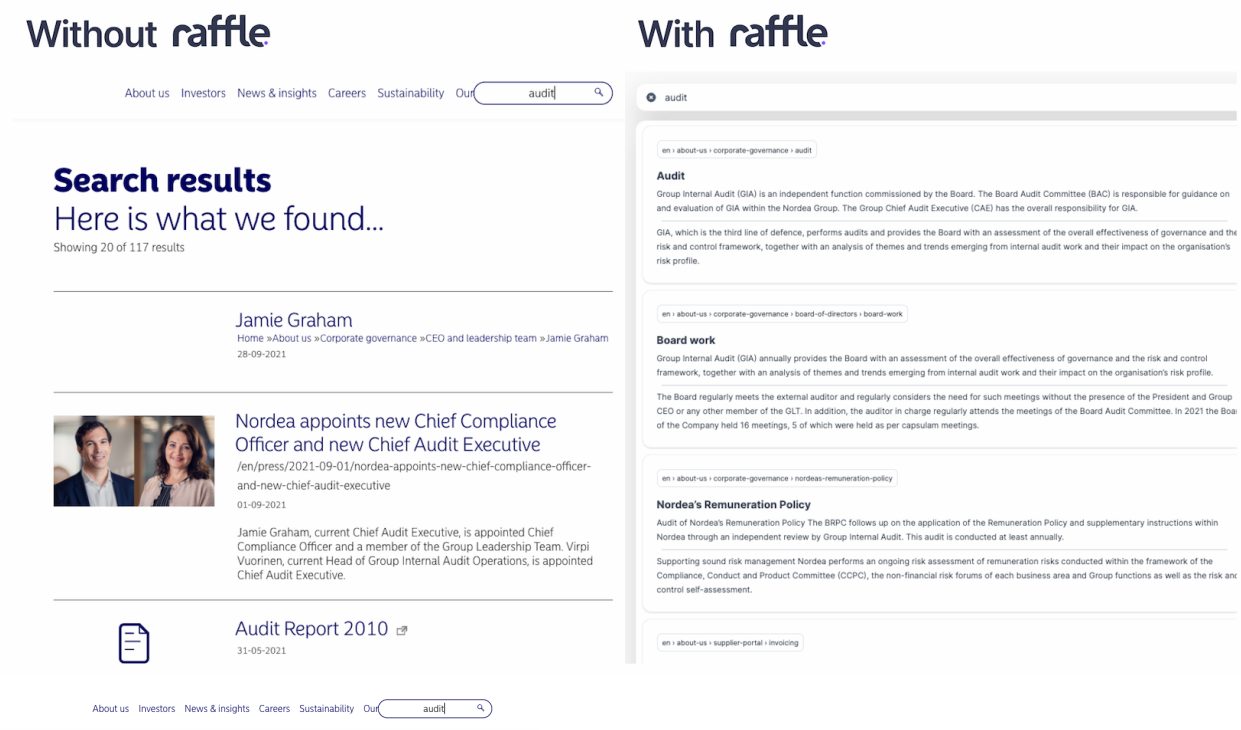


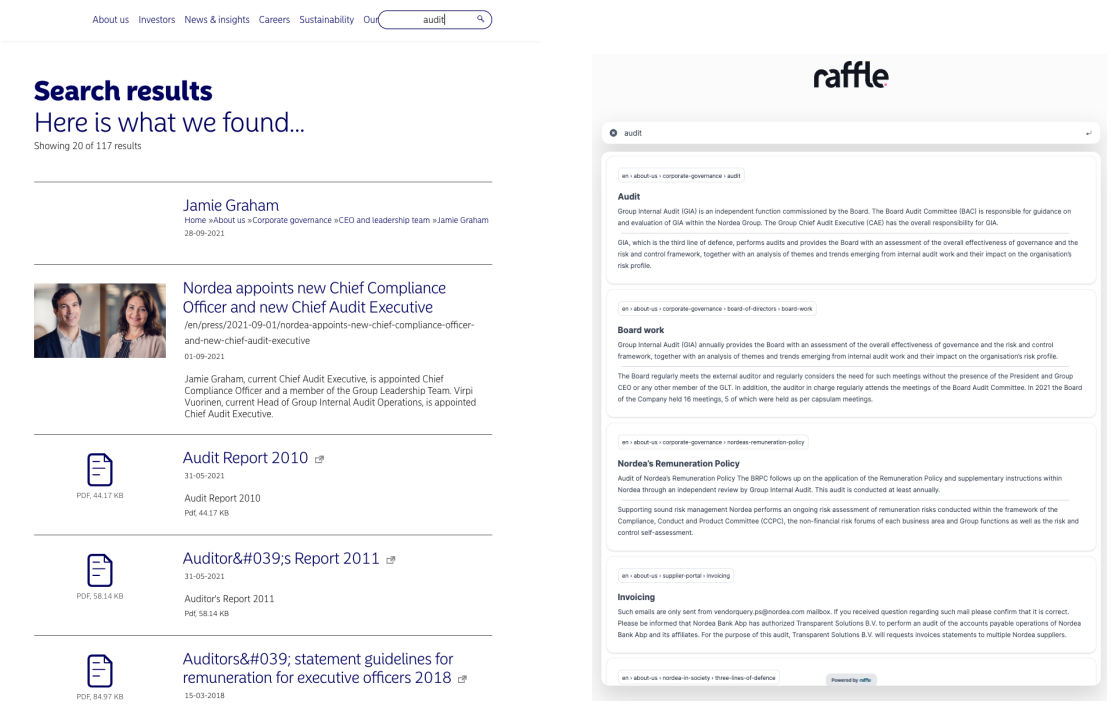
Also, Traditional Search often returns ‘zero results’ if there are spelling mistakes. Raffle Search does not depend on keywords or tags, and spelling mistakes are therefore no issue.



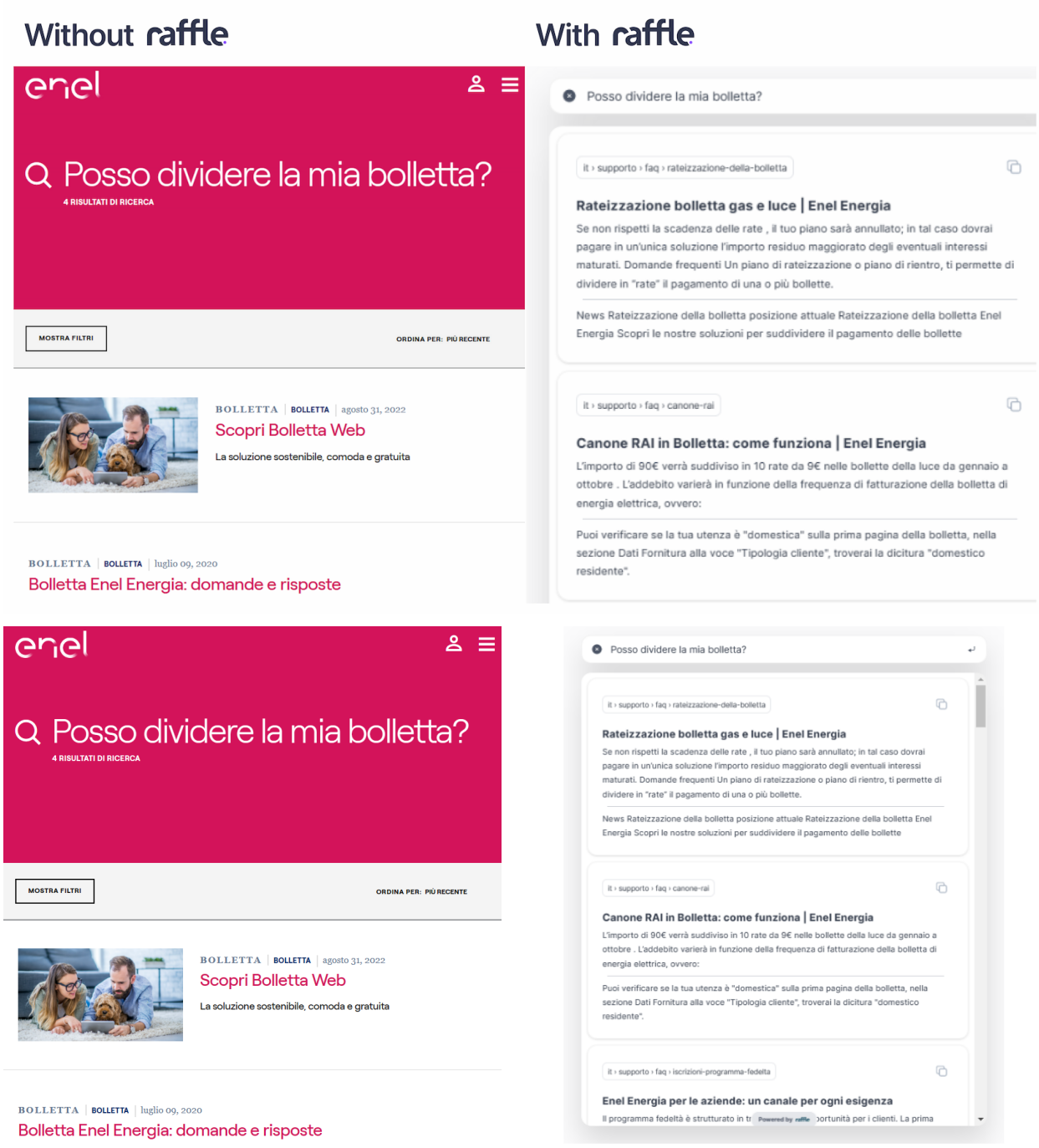
Where Traditional search sees single words as a keyword and ranks results by the logic of most keyword matches, Raffle sees the intention with single words and translates it into relevant answers.

In other words, single words have intent as well:



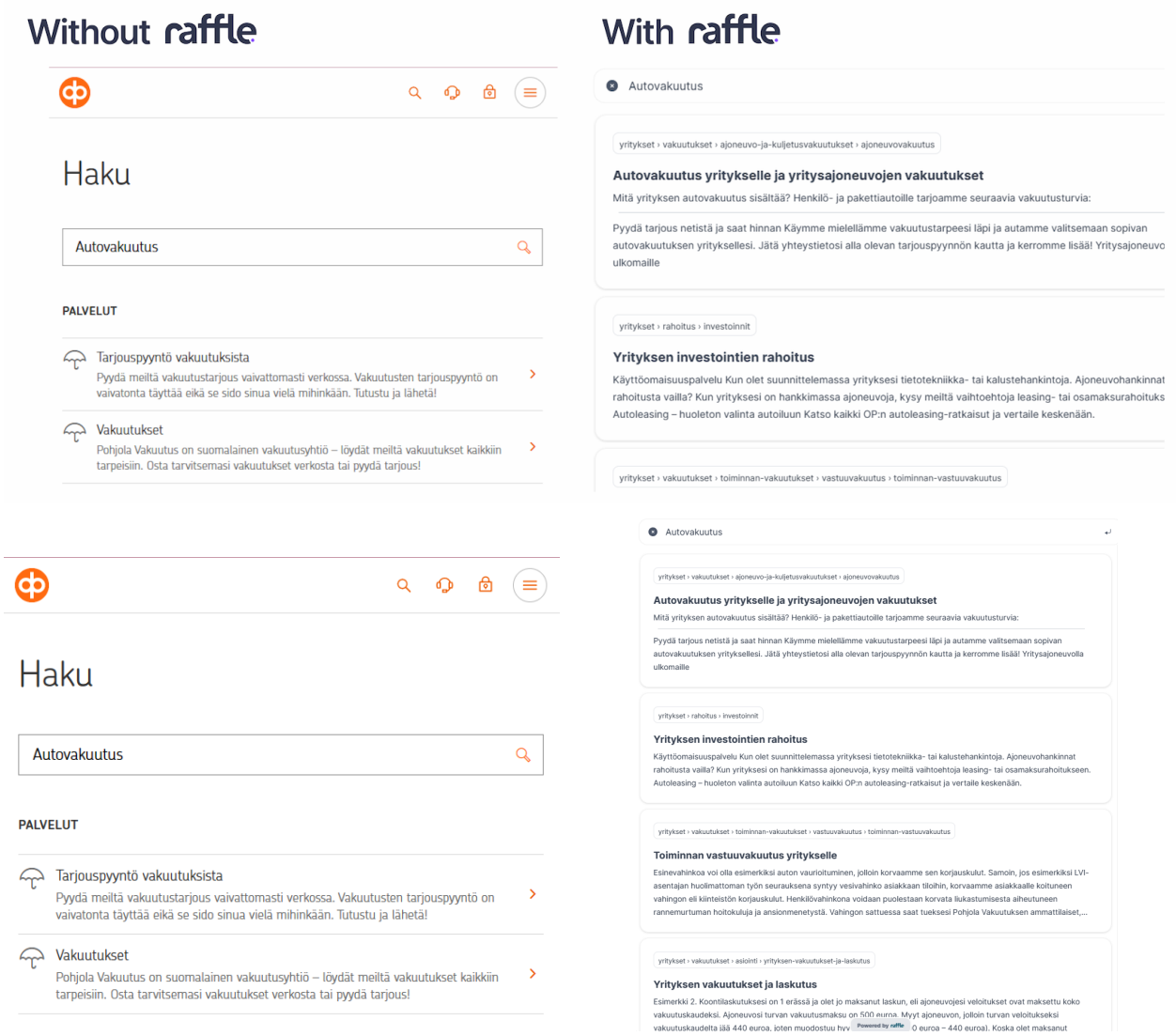


Below is an example in Italian, where the Traditional Search on the current website cannot find the right answer:



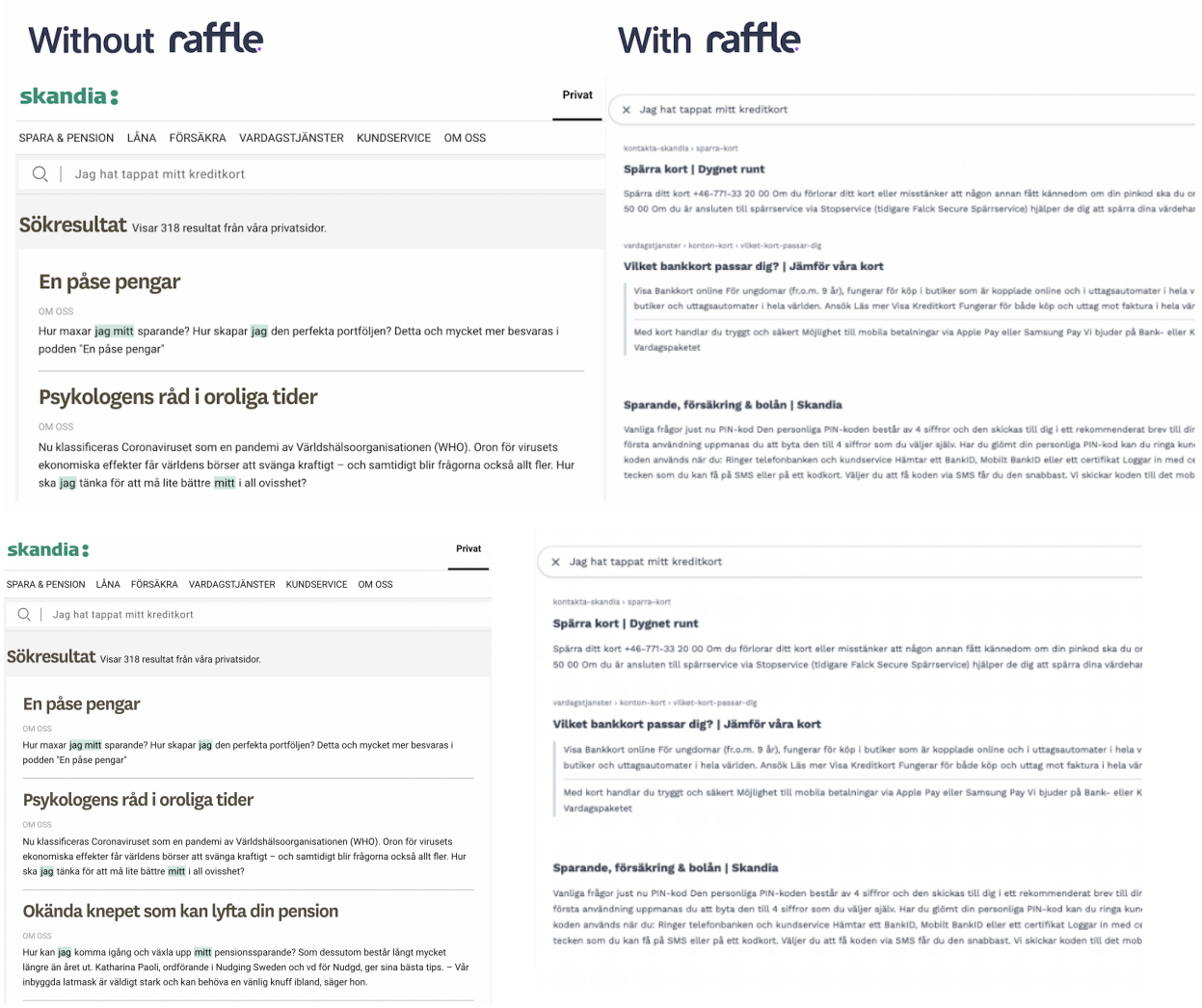
And then with a large company from Finland. Single word search on Traditional Search on the left side that cannot find the right answers, and Raffle on the right side finding the right answers out of the box:

https://raffle.ai | hello@raffle.ai | VAT no:. DK39747189
US:.+1 929-207-4717 | UK: +44 20 3966 5755 | Benelux: +31 202442319 | Nordics: +45 4290 9050

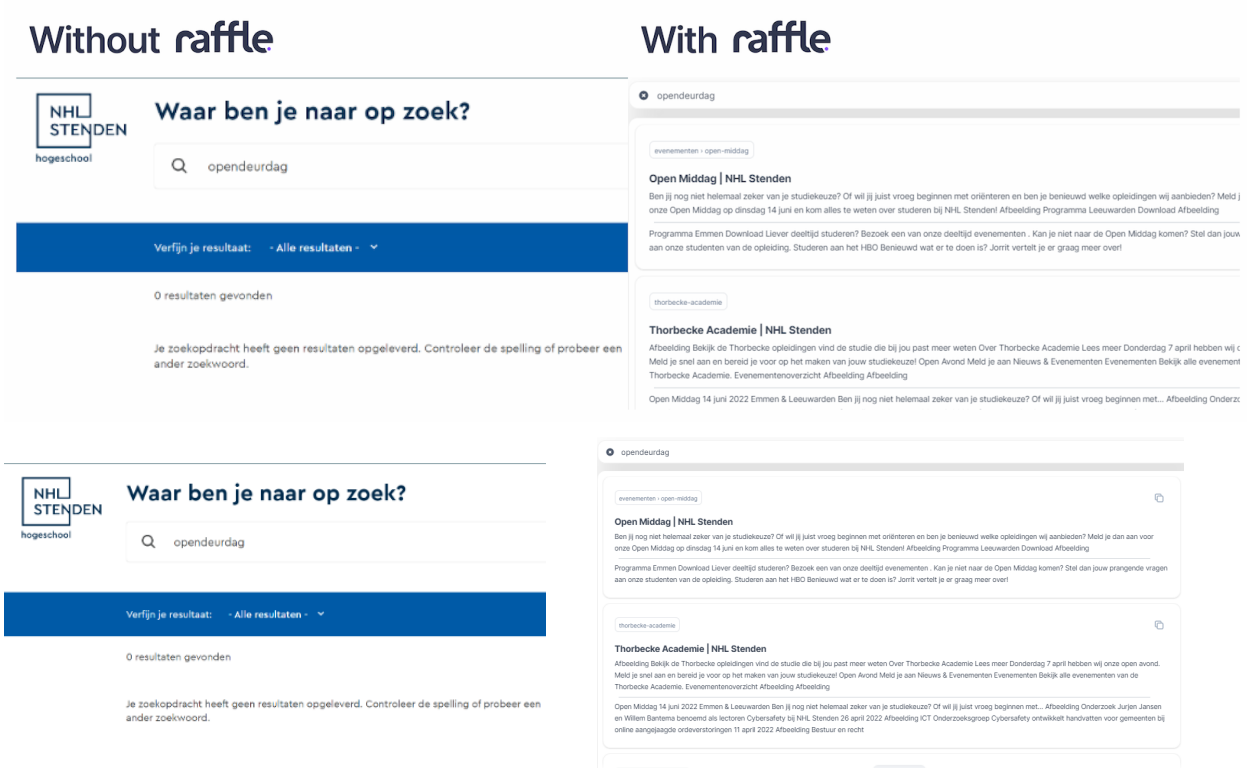


Below with a large Swedish company on a search string. Raffle on the right side giving the right answer out of the box. Whereas on the left side the answers have nothing to do with the search applied.

https://raffle.ai | hello@raffle.ai | VAT no:. DK39747189
US:.+1 929-207-4717 | UK: +44 20 3966 5755 | Benelux: +31 202442319 | Nordics: +45 4290 9050



In Dutch with a large company from Holland (The Netherlands). The single word that is not understood by the traditional search because the word was not tagged. Raffle, however, understands the intent of the word and can find the right answers on day one.



Any language in the world goes with Raffle Search because of its origin and core in vectors and semantics.

By analyzing inquiries externally and internally, Raffle optimizes your website, allowing customers to flow naturally through the customer acquisition funnel, ultimately reducing customer churn and giving businesses a competitive edge in today's marketplace.

Conclusion

In conclusion, Raffle Search's hybrid search technology has revolutionized search experiences by utilizing vector search technology, which maps documents and queries to vectors using large language models neural network, reflecting the intent and semantics of the content, providing highly accurate and personalized search results.

As we progress into the future, the importance of AI search will only increase. However, it is crucial to recognize that the foundation of its success is its sustainability and accuracy in producing relevant data. Without this, even the most advanced AI models like GPT will fail to deliver the desired outcomes. An issue we delve into deeper in our following insights papers.