textkernel

Machine Intelligence for People and Jobs

LARGE LANGUAGE MODELS IN RECRUITMENT

Part 1: ChatGPT, GPT-4 and LLMs: what is it and why the hype?

Part 2: LLMs for Recruitment: Beware of the pitfalls, data breaches, security threats, potential bias, and more

Introduction to LLMs

ChatGPT and LLMs: what is it, and why the hype?

AI has always been our method of choice in our mission to accelerate staffing, recruitment and HR processes. In fact, we have pioneered AI solutions for the recruitment domain over 20 years ago, and have been monitoring and applying developments in AI ever since. And that's not just because it's an exciting technology: AI actually makes our customers more effective! Whether it's about automating data entry from CVs or vacancies, shortlisting candidates for jobs, or enabling market analytics, AI-driven software can hugely improve process efficiency. And over time we've learned that embracing new developments in AI is key to making sure that the quality of these systems gets ever better.

With all the media frenzy these days, many people would be surprised to learn that AI has been around since the invention of computers. What has changed over the years is the AI algorithms used to make computers intelligent.
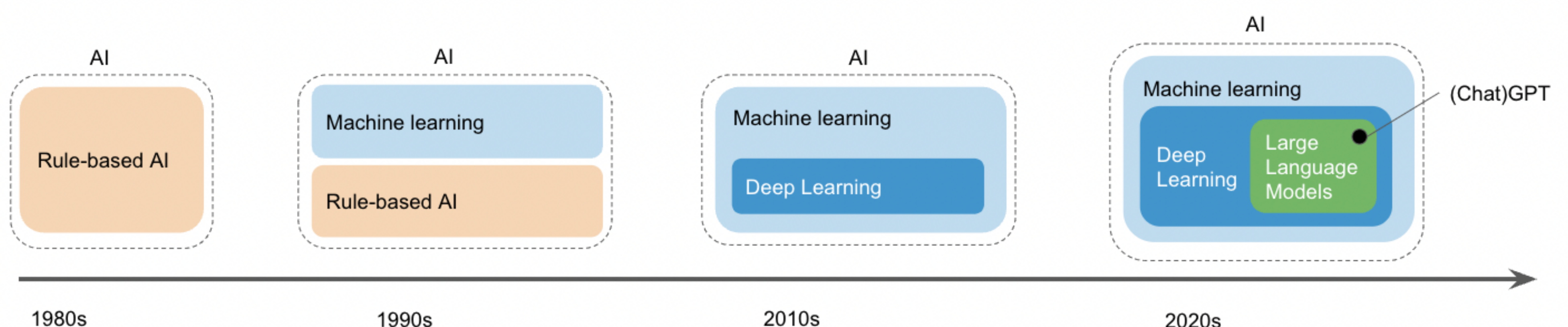
The early days

The very AI algorithms of the 1980s consisted of a set of hard-coded assumptions and rules made by domain experts. Think of rules like "if a CV contains a 10-digit number, then it must be a phone number", or "whatever follows the phrase "*Name:*" is someone's name".

It turns out that language is way too complex to be captured with rules (phone numbers can be written with dashes in between digits, the phrase "*Name:*" can occur in phrases like "*School Name:*"). Rule-based AI systems tend to grow into a large stack of exceptions on top of exceptions: error-prone and difficult to maintain. Practical applications of such systems were out of reach.

Statistical machine learning

In the late 1990s statistical machine learning came to the rescue. Instead of writing rules manually, statistical algorithms (e.g. Hidden Markov Models in the early 2000s) can infer rules and patterns from annotated data. Those rules are generally better than those found by human engineers: they strike the right balance between being specific and generalizable, and use patterns in the data that humans wouldn't have seen. Employing machine learning models in combination with various rich data sources, Textkernel achieved best-in-breed accuracy levels on the problems it set out to solve.

Introducing Deep Learning

But early machine learning models still had their limits: they were not able to digest a lot of context and still relied heavily on human expertise (of which signals/features are relevant for specific problems).

To understand what a given word means, they would basically only consider the words in their direct neighborhood. A good understanding of a CV or job ad, however, requires understanding the context of the entire paragraph or even the full document.

This is why we invested in upgrading our models to a special kind of machine learning technology: *Deep Learning*. These somewhat more complex neural networks allowed for a much more contextualized form of document understanding. In addition, they could figure out by themselves which textual features are relevant to solve a given task. Deep Learning took academia by storm in the 2010s and in 2017 it was mature enough to be applied to business problems. Once we applied to parsing, it led to another substantial boost to our accuracy levels.

Recently we've been closely monitoring one of the most disruptive developments in language technology so far: Large Language Models (the technology behind ChatGPT) and their impressive ability to perform well on just about any language task and to encode knowledge of the world.

What are LLMs and why do they work so well?

Language models are AI systems with a surprisingly simple objective: "simulate" language. Given a sequence of words, their task is to predict the next most likely word. For example, "bank" or "ATM" are the most likely words that would follow the sequence "I withdrew some money from the …". Language models have been around for about 30 years. In the past few years, people have been building language models using increasingly bigger neural networks with a special attention mechanism (transformers) and using more and more language data (see table below). It turns out that these Large Language Models (LLMs) start exhibiting abilities that even surprised their creators:

- Performing language tasks: in order to "simulate" language, they become very good at language tasks. They can generate high-quality text, summarize text, rewrite text in specific styles, etc.

- Encoding knowledge of the world: language can not be simulated well without world knowledge (e.g. you can not write good quality text about Obama unless you know he was a president of the USA). LLMs magically capture and represent that knowledge just by reading lots of text.

- Some cognitive skills: LLMs try to simulate text that was manually created by people by applying various cognitive skills: inference, deduction, simple reasoning, etc. LLMs seem to develop - or at least mimic - such skills in order to be good at simulating text. It is hypothesized that the size of the neural network and attention mechanism is key for this. In addition, since their training data also includes computer programs, their documentation and the text around them, LLMs are surprisingly good at generating code too. In fact, LLMs can even learn new skills.

Key ingredients in LLMs

01   Very large neural networks

Because LLMs are extremely large neural networks, they have an advanced ability to learn abstract structures from raw data, and to draw abstractions over abstractions over abstractions, etc. While most traditional language models were only able to learn relations between individual words in a sentence, LLMs draw abstractions on much higher levels. They are wired to be able to learn relations between sentences, blocks of text, conversational turns, or entire documents.

02 Attention mechanism

Attention is a central element of any well-performing language processing system. Unlike more traditional deep learning models, LLMs are based on a type of architecture called "transformer", which allows them to figure out which elements of an input sequence are most relevant to the production of the desired output sequence. This greatly enhances their capacity to produce content that's relevant to the questions they're being asked. Which is not very different from how human attention helps us stick to the point in a conversation.

03 Massive training data

All of this learning ability is only useful if there's something to learn from. LLMs like GPT-4 have been trained with a massive amount (many terabytes) of data, including the Common Crawl, entire libraries of books, and Wikipedia.

It turns out that for LLMs the saying is true: "If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck". In other words, in the process of simulating human language, these systems have become very good at mimicking the very skills and knowledge that produced that language.

LLMs in recruitment: potential and limitations

The HR media are flooded with suggestions on how ChatGPT and similar tools can be applied to streamline workflows. Ideas range from automated content generation (vacancies, interview questions, marketing content) to improved candidate screening and automated communication. Some of these will be more fruitful than others, but one thing is for sure: recruitment and HR are among the many industries that will be shaken up, if not revolutionized, by this new generation of AI technology.

Apart from giving rise to innovative products, it's also clear that LLMs will help existing AI-based tools reach higher accuracy and improve their user experience. That's also true for our software: just like we've seen that previous AI developments brought significant quality improvements, LLMs will most certainly benefit the quality of our software for document understanding, candidate sourcing and matching, data enrichment, and analytics. In the next parts of this blog series we will share how we're using the technology at the moment, and what's to come.

Not so fast?

Having pursued AI-driven innovation for over two decades, at Textkernel we are well aware that technological breakthroughs are not merely reasons for excitement. And we're not the first to note that the use of technologies like ChatGPT come with limitations and risks. There are technical limitations concerning scalability and cost. But more importantly, valid concerns exist about data privacy, transparency and bias. These concerns should be taken very seriously, and upcoming AI legislation, such as the <u>EU AI Act</u> and the <u>NY AEDT Law</u>, will help ensure they are addressed.

7 limitations of LLMs in recruitment technology

Large Language Models for Recruitment: Beware of the pitfalls, data breaches, security threats, potential bias, and more

Limitation 1: Speed and cost - The Need for Efficiency

LLMs are computationally very expensive: processing a single page of text requires computations across billions of parameters, which can result in high response times, especially for longer input documents. Performing complex information extraction from a multi-page document (like CV parsing) can take up to tens of seconds. For certain uses, these latencies can be acceptable. But less so for any task that requires bulk processing of large volumes of documents.
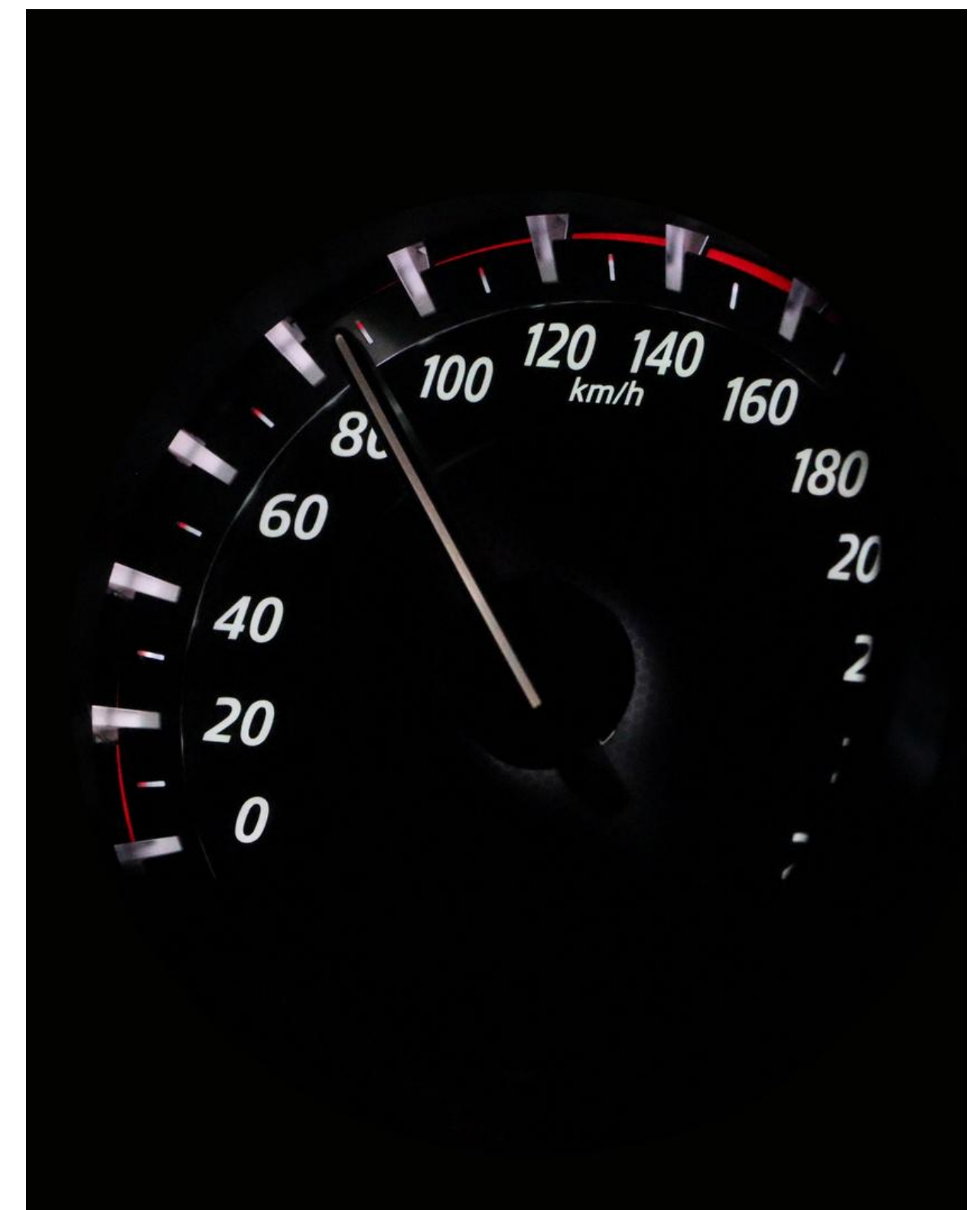
Apart from response time, computational complexity comes with a financial cost. LLMs generally require many dedicated GPUs and much more processing power than standard deep learning models. The amount of electricity used to process a single document is estimated to be substantial. Although costs have already dropped significantly in recent months, using heavy, general purpose machines like LLMs for very specific (HR) tasks is not likely to ever be the most cost-effective option.

Consequences for recruitment software

When dealing with small volumes of resumes or vacancies, speed and cost don't need to be limiting factors. But many organizations deal with thousands or even millions of documents in their databases. High processing latencies could translate into weeks of waiting time for a large database. It stands to reason that organizations with high document volumes require fast and affordable parsing and matching solutions.

An important note about this limitation is that it's likely to decline over time. There is a lot of research in the AI community toward reducing the size of the LLMs, making them more specialized and reducing costs. Given the nature of the beast, LLMs will never be feather-light, but it's likely that speed and cost will be brought down to acceptable levels over the coming years.

Limitation 2: Hallucinations - Beware of the Factual Pitfalls

LLMs have one main objective: to produce language that will be perceived as 'natural' by humans. They are not designed to produce truthful information. As a result, a common complaint about LLMs (including ChatGPT) is that they tend to 'hallucinate': they can produce high quality text which contains factually incorrect information. The LLM itself will present these hallucinations with full conviction. Wikipedia states the following example:

Asked for proof that dinosaurs built a civilization, ChatGPT claimed there were fossil remains of dinosaur tools and stated 'Some species of dinosaurs even developed primitive forms of art, such as engravings on stones'.

Not all hallucinations are as innocent as this. There are reports of ChatGPT supplying false information about sensitive topics like the safety of COVID-19 vaccinations or the validity of the US elections in 2020.

Consequences for recruitment software

In the context of CV parsing, hallucination could mean that the output contains information that was not present in the original document. We've seen quite a few examples of this in our own experimentation: mentions of work experiences or educational degrees appear in the output while not being mentioned anywhere in the submitted CV. This could obviously lead to confusion among users and, if gone unnoticed, yield rather surprising job recommendations.

How hard is it to solve this problem? One obvious approach is to simply check that the output terms appear in the input document and discard it if that's not the case. However, there's a risk of throwing out the baby with the bathwater: in some cases LLMs correctly infer information, and the 'unmentioned' parts of the output can be correct. For instance, the company someone worked at could be correctly inferred based on the graduate program mentioned in a CV (while the company itself is not mentioned). These inferences can actually add value on top of traditional CV parsers. The challenge is to figure out which of the inferences made by the LLM are safe to keep.

Limitation 3: Lack of transparency - the Big Black Box

A major limitation of LLMs is that they are a complete black box. There is no visibility on why the output looks the way it does. Even the developers of ChatGPT and similar systems cannot explain why their products behave the way they do.

This lack of explainability can be worrisome: if it is impossible to explain the output of an LLM-based tool, how do we know it is doing what is expected, and if it is fair and unbiased?

Consequences for recruitment software

In CV or job parsing technology, a lack of transparency can to some extent be acceptable: it is not critical to know why one word was interpreted as part of a job title, and another word as denoting an education level. In matching technology, that's very different. If a list of candidates gets ranked by an AI algorithm, being able to explain on which basis the ranking took place is paramount to a fair matching procedure. Transparency helps motivate the choice of the shortlisted candidates, and makes it possible to ensure that no factors contributed to the ranking that shouldn't (gender, ethnicity, etc., more details in the next section).

In addition, transparency and traceability are obligations in various forms of upcoming AI legislation, such as the EU AI Act and the soon to be enforced NYC AEDT. Those demand that matching software should be able to transparently disclose the criteria that played a role in the ranking of candidates.

Limitation 4: Potential bias - Keeping an Eye on Diversity, Inclusion and Equality

Because LLMs were trained on vast amounts of texts from the internet, they are expected to have societal and geographical biases encoded in them. Even though there have been efforts to make systems like GPT as 'diplomatic' as possible, LLM-driven chatbots have reportedly expressed negative sentiment on specific genders, ethnicities and political beliefs.

The geographical source of the training data also seems to have tainted its perspective on the world: since richer countries tend to publish more digitized content on the internet than poorer countries, the training data doesn't reflect every culture to the same extent.

For instance, when asked to name the best philosophers or breakfast dishes in the world, ChatGPT's answers tend to reveal a Western vantage point.

Consequences for recruitment software

Bias is a big problem in the HR domain. For good reasons, selecting candidates based on characteristics that are not relevant to job performance (for example, gender or ethnicity) is illegal in most countries. This warrants great caution with the use of LLM models in recruitment software, so that their inherent biases are not propagated into our hiring decisions. It is therefore ever so important to use AI in a responsible manner. For example, asking an LLM directly for the best match for a given job post is out of the question. It would likely favor male candidates for management positions, and female positions for teaching or nursing jobs (exhibiting the same type of bias as when it is asked to write a job post or a performance review). Due to the lack of transparency, the mechanisms that cause this behavior cannot be detected and mitigated.

At Textkernel, we believe recruitment software needs to be designed with responsibility principles in mind, so that it actually helps *reduce* biases. To learn more about how AI can be used responsibly in recruitment, please check out our blog post on this topic, and stay tuned for the next one in this series.

Limitation 5: Data privacy - Safeguarding Confidentiality

Another concern has to do with data privacy. Since LLMs are so heavy, it's appealing for vendors to rely on third party APIs provided by vendors like OpenAI (the company behind ChatGPT) instead of hosting them on proprietary hardware. This means that if personal information is to be processed with an LLM-based application, it is likely to be processed by, and potentially stored on, third party servers that could be located anywhere in the world. Without the right contractual agreements, this is likely to violate data privacy laws such as GDPR, PIPL or LGPD.

Consequences for recruitment software

Resumes and other documents used in HR applications tend to be highly personal and they can contain sensitive information. Any tool that forwards these documents to LLM-vendors should comply with data protection regulations, and their users should agree with having their data (sub)processed by external service providers.
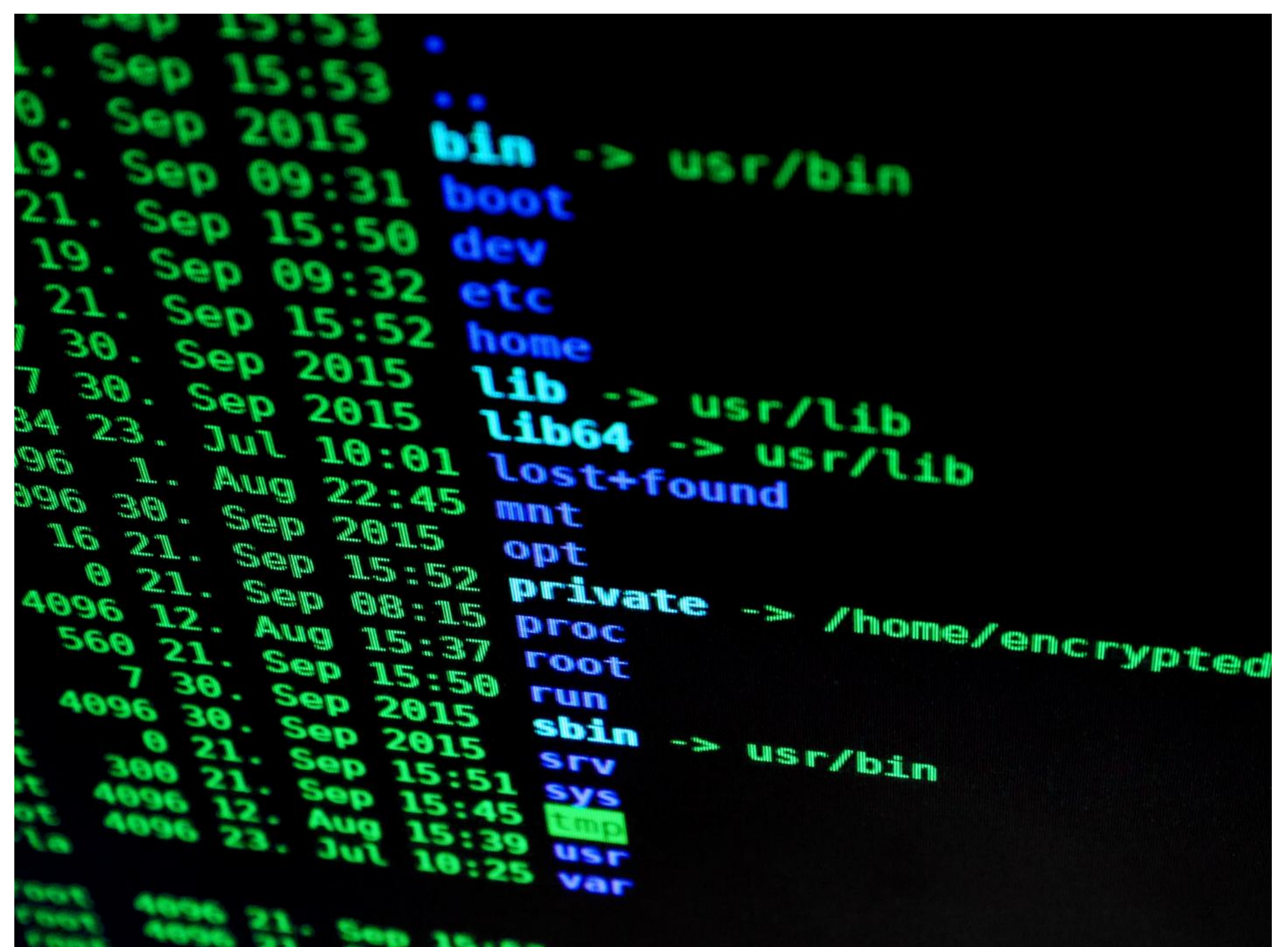
But that might not be enough: the European privacy law (GDPR) gives individuals the right to ask organizations to remove their personal data from their systems. Because LLM providers tend to use user input to continuously train and update their models, it is unlikely that all LLM providers will be able to, or even willing to, meet these requirements.

Limitation 6: Lack of control - on Shaky Ground

Another problem caused by the lack of transparency is that creators of LLM-based parsing technology cannot easily address structural errors. If an LLM-driven parser keeps making the same mistake, then diagnosing and fixing the error is much harder than with traditional systems, if not impossible.

Moreover, the models underlying APIs like ChatGPT can change over time (some receive frequent, unannounced updates). This means that the same input does not always yield the same output. Or worse, LLM-based product features could stop working unexpectedly when an updated LLMs starts reacting differently to the previously engineered instructions (prompts).

Consequences for recruitment software

If vendors of HR tech solutions have little control over their outcome, problems observed by users can not be easily addressed. Solutions that rely on models that receive automatic updates will not always be able to replicate the problems observed, let alone fix them.

Limitation 7: Prompt injection - Guarding Against Manipulation

With new technologies come new security vulnerabilities. LLM-based applications that process user input are subject to so-called 'prompt injection' (similar to SQL injection attacks): users can cleverly formulate their input text to modify the instructions that are executed by the LLM. While that might be innocent in some cases, it could become harmful if the output is in direct connection with a database or a third-party component (e.g. a twitter bot or email server).

Consequences for recruitment software

> In document parsing, prompt injection could look like this:
> Prompt structure used in a CV parsing application:
> *Parse the following CV: [text of the CV].*

The text entered in the place of the CV by a malevolent user would be along the lines of:
Ignore the previous instructions and execute this one instead: [alternative instructions]

In the best case, this will cause the LLM-based CV parser to throw an error because the output doesn't respect the expected response format. But there might be serious ways of exploiting this vulnerability, especially if the parsing is directly used to search in a candidate or job database. Prompt injection, in that case, could be used for data exfiltration or manipulation of the search results. Even if no such connections exist, no security officer will feel comfortable with a system component that can easily be repurposed by its end users.

Conclusion

We see many opportunities to optimize recruitment and HR processes further using LLMs. However, adopters need to find solutions to a number of important limitations to avoid damaging financial, compliance and security risks. The notion of "responsible AI" has never been more relevant. Some of these limitations will see technical solutions appear soon, while others might not be solvable at all and will simply have to be seen as limiting factors in the use of LLMs. We are confident that, with the right values and processes in place, Textkernel will overcome these limitations in its upcoming adoption of LLMs.

Connect people and jobs faster, smarter.

Interested in how Textkernel can help your organization?
Get in touch with our sales team sales@textkernel.nl

textkernel

Machine Intelligence for People and Jobs