



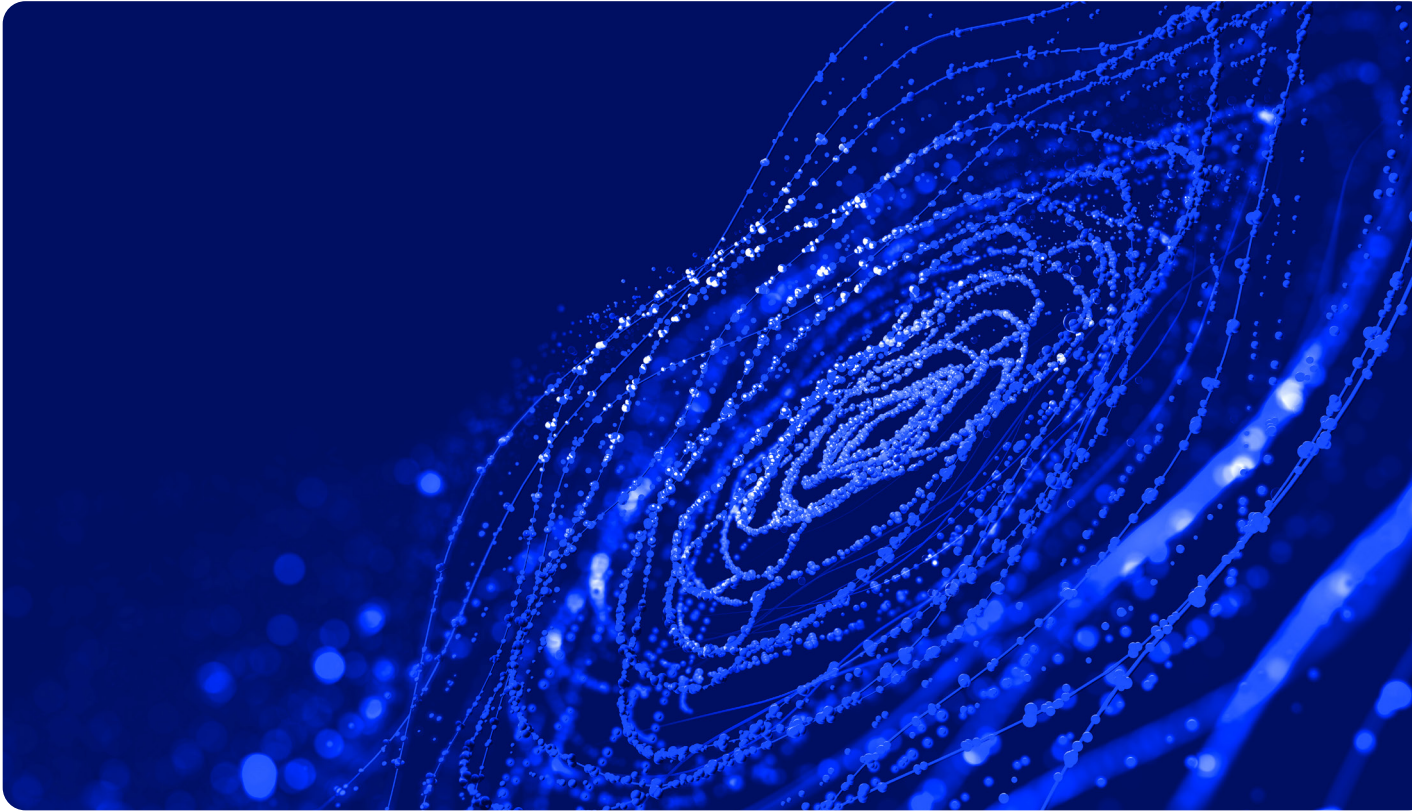
# Building Luminance's Artificial Intelligence

WHITE PAPER

# | Contents

Building Luminance's Artificial Intelligence	1
A Disruptive AI Technology	2
Building the Most Sophisticated Language Model for Legal Document Processing	3
Embedding Applications	6
Transformers	9
AI in Practice: Luminance's AI-Powered Features	10

# Building Luminance's Artificial Intelligence



Luminance is the world's most advanced AI technology for the legal processing of contracts and documents. Founded by mathematicians from the University of Cambridge, Luminance's AI reads and forms a conceptual understanding of documents in any language. Luminance uses this understanding to augment the spectrum of legal matters, from AI-powered contract drafting, negotiation and review to investigations and eDiscovery. Luminance is used by over 500 customers in 60 countries, including all of the Big Four consultancy firms, over a quarter of the world's largest law firms and multinational organisations.

Luminance's core intelligence is the Legal Inference Transformation Engine (LITE), the first true application of machine learning to the legal industry, which uniquely combines pattern-recognition technology with supervised and unsupervised machine learning. Exposed to over 150 million documents so far, Luminance's language model rapidly reads and forms an understanding of documents before clearly displaying the results of its analysis to the user across a series of easy-to-navigate interactive widgets, labels and 3D visualisations. Luminance's AI combines several disciplines within the field of machine learning including inference, deep learning, natural language processing and pattern recognition. With its advanced machine learning technology, Luminance provides legal professionals with the most robust, powerful platform for legal document analysis and process automation.





Luminance's AI for legal process automation and augmentation offers lawyers and organisations a highly intuitive and intelligent platform for legal document analysis and comparison. Luminance automatically finds patterns and correlations in data that would take weeks to uncover via manual review methods. Luminance's AI learns from every document it sees, so that it can adapt to a legal team's individual strengths and specialisms without being explicitly trained. Luminance helps lawyers to understand an entire project at a glance, as well as giving granular analysis of specific clauses, documents or concepts.

## A Disruptive AI Technology

Older or more rudimentary technologies have attempted to alleviate the burdens imposed upon organisations facing the dual strain of increased legal data production and growing bodies of localised regulation governing the way that data must be processed. However, the inherent rigidity of these older rules-based systems means that they can do little more than query legal data for keywords during the course of document review. Furthermore, they require intensive pre- or re-programming following every legal shift, rendering them fundamentally unsuited to the ever-changing environment of regulatory compliance and/or the legal industry. They are even more incompatible with a modern, globalised economy that has seen years of unpredictable world events which require a rapid assessment of an organisation's contractual standing across a range of unforeseen issues; the Covid-19 pandemic and the introductions of sanctions against Russian entities are just two recent examples which have required drastic evaluation and action.

**"With Luminance, we have everything we need in one place to enhance our legal processes."**



**Laura Pickle**  
CIO and Discovery Director



Luminance's AI approach allows for complete flexibility in legal document processing as it develops a fundamental understanding of language; this allows Luminance to render paragraphs of text into numerical vectors and thereby comprehend them in terms of patterns. This underlying technical architecture ensures Luminance's thriving research team in Cambridge can apply the latest machine learning techniques to the 'real world' problems faced by legal professionals today. Nevertheless, it is not always the latest models borne out by academic research that will be the right tool for the job; the real innovation lies in Luminance's application of both existing and novel techniques to the spectrum of legal document review.

## Building the Most Sophisticated Language Model for Legal Document Processing

AI-generated language models are becoming increasingly sophisticated in their ability to understand the written word. Computers are fundamentally ill-equipped to understand human language, so the essential first step in developing any machine learning model is to make human language intelligible to the computer. Hence the emergence of a subfield of artificial intelligence known as Natural Language Processing (NLP), which sits at the intersection of computational linguistics, machine learning and deep learning. This field is dedicated to developing techniques that help computers analyse and synthesise human language and forms the basis of Luminance's language model, allowing it to both derive and understand meaning from any corpus of text.

### **OPTICAL CHARACTER RECOGNITION**

For users uploading scanned documents or other image-like files to Luminance, this process begins with Optical Character Recognition (OCR). OCR is the method by which a computer turns an image of printed text into parseable, highlightable or editable text. The computer must be able to read the document before it can begin to understand it. Luminance has an in-house team working to constantly improve this process, developing tools to recognise new languages and formats.

### **NATURAL LANGUAGE PROCESSING**

Machine learning models generally require numerical inputs, and so any approach to NLP begins with embeddings that convert text into vectors that help computers process language. The simplest vectorisations are syntactic, which relates to the structure of the sentence itself, and involve techniques where text becomes a count of how many times each word occurs within a body of text. Other vectorisation techniques are semantic, which relates to the meaning of the words and sentences. Word embedding is a commonly used semantic vectorisation that numerically maps text into a high dimensional vector space where words of similar meaning have a similar representation.

### **SYNTACTIC NLP**

Luminance uses *n*grams to ensure that vectors are informed by contiguous sequences of words rather than by individual words in isolation. These syntactic vectors are relatively common within the field of NLP, but Luminance has further developed this methodology to create a custom technique to inform its language model.

Luminance's unique 'genome' vectors are inspired by a fingerprinting technique (originally developed by Google) to identify duplicate websites. Luminance uses a unique technique to sum these vectors together over any overlapping window to get the resulting syntactic vector. These syntactic embeddings underpin Luminance's advanced pattern recognition technology and exponentially speed up Luminance's text comparison algorithms.

**“The fact that the technology is language agnostic makes it uniquely capable of digesting the geographically wide-ranging documents we upload to Luminance.”**



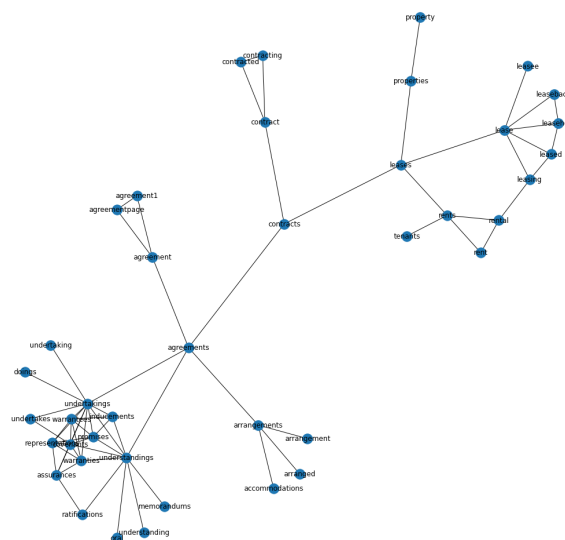
**Yan Pecoraro**  
Partner



Ngram vectorisation techniques can also be refined with term frequency-inverse document frequency (TF-IDF), a statistical method of assessing the importance of a word both within a specific document and the larger corpus. TF-IDF is, in effect, being used to reweight the words mapped into the vector space according to their importance. This is used predominantly for Luminance's legal process automation platform, Luminance Corporate.

## SEMANTIC NLP

Whilst the methods described give syntactic embeddings, the language model also needs to obtain semantic understanding. This is where the Word2vec technique is applied. Word2vec uses a neural network model to learn word representations and associations from a large corpus of text. The model is asked to predict words to fill gaps in a sentence, basing its prediction on the surrounding context. This technique embeds meaning into the vector space. Sentences and even paragraphs can then be treated as the average vector of their words. Paragraphs with similar wording and/or meaning thus receive similar vectors.



*This network shows connections between the most similar words. Our models learnt these similarities from a large legal corpus by analysing the contexts in which different words appear.*

Luminance's in-house research team have developed new transformer methods specifically for our Luminance Corporate product. Transformers combine syntactic and semantic methods implicitly within the network, while still being application specific – it can be fine-tuned to suit any problem.

These unsupervised techniques allow us to build embeddings for *any* language out of a corpus of documents and underpin Luminance's unique language agnosticism. Indeed, Luminance is the only technology within the legal space that is being used for document review in over 80 languages, including non-Latin script languages and right-to-left text.

## | Embedding Applications

Word embeddings form the basis for most of the tools in our LITE engine, as outlined below:

### SYNTACTIC SIMILARITY: NEAR DUPLICATE DETECTION AND CLUSTERING

In a situation where a lawyer would wish to identify syntactically similar content in their dataset, whether that be clause phrasing or document type, Luminance uses its genome vectors together with the fingerprinting method. Luminance has developed a novel indexing method that helps find all vectors within a multi-dimensional vector space. The application of this method means that, for instance, once a document of interest is located, Luminance can rapidly identify syntactically similar content on the basis of proximity to it within the vector space. This technique underpins both Luminance's Duplicate Detection, which eliminates duplicate documents from the dataset at the click of a button, and the Clustering Tool, which groups documents according to type and displays this back to the user in the form of an interactive, spatial widget.



*AI-powered document clustering enables lawyers to visually understand their contract templates and easily maintain compliance.*

## PATTERN RECOGNITION: DIFFING/COMPARISON AND ANOMALY DETECTION

Luminance has the ability to compare wording across a document set and identify anomalies within neutral wording. Luminance has developed a custom mapping technique that overcomes the inherent problem involved in identifying all the similarities and differences across a set of documents that look similar and are produced according to a similar template. For instance, across thousands of documents and thousands of paragraphs, there are enormous numbers of patterns. Furthermore, if these are scanned documents, OCR errors will inevitably be encountered, which makes the job of identifying these patterns significantly more difficult. Luminance's mapping technique, however, uses document partitioning to overcome both OCR and the sheer volume of data, identifying patterns across the entire document set and surfacing similarly phrased paragraphs.

## SEMANTIC HYPERSPACE: SIMILAR SUGGESTIONS AND METRIC LEARNING

If a reviewer finds a paragraph within their document set to be of specific interest, Luminance can find paragraphs that contain similar wording and notions by using Word2vec embeddings. Luminance uses the average Word2vec vectors more when looking for suggestions with similar meaning, irrespective of exact wording. Luminance's research team recently demonstrated the benefits of 'metric learning' which is a semi-supervised technique and a very new approach within the field of machine learning. It is based on training the model to have a better understanding of the similarity of documents by explicitly training it to embed vectors so that the cosine distance between two samples is a better proxy for identifying similarity.

**"Tools like Luminance that enable lawyers to quickly and easily see which areas of a contract need to be negotiated can speed up the time to contract – and faster contracting means faster revenue generation."**



**Rosemary Martin**  
Group General Counsel



## CLASSIFICATION (CONTRACT TYPES AND CLAUSES)

A large component of Luminance's technology is devoted to contract and clause classification. It is a tool for lawyers to rapidly navigate their data and is the crucial first step in the development of Luminance Corporate's contract negotiation pipelines. For Luminance's 1,000+ pre-learned concepts that it can recognise out-of-the-box, we use bi-directional attention-based Long Short-Term Memory (LSTM) artificial Recurrent Neural Networks (RNNs). RNNs iteratively incorporate embeddings of words in a sentence into a hidden-state vector which, ideally, eventually represents the meaning of the sentence. In practice, they struggle with memory issues, and balance remembering early words with learning new words, so a 'memory mechanism' is added to allow the model to choose what to 'forget'. For the attention mechanism, which is used to weight an average for Luminance's final representation, Luminance uses the final output as well as the hidden states after each additional word. Furthermore, Luminance does this left-to-right and right-to-left, as this gives the model an opportunity to use text from later in the sentence to understand the meaning of earlier text and, therefore, understand the broader context of the sentence.



**“We’re reaching a tipping point where it will be negligent not to use technology... if you’re not, you’re running a real risk of missing something.”**



**Rob Webb KC**

Former General Counsel for British Airways and Rolls-Royce

Empirically, Luminance has found that all of these additions give the best models where it encounters large datasets, but the models become overly complex if there are only have a handful of examples of certain concepts. In these situations where Luminance receives less training data, the technology adopts a simpler support vector machine (SVM) over *n*grams. This simpler model not only works better on smaller amounts of data but also runs faster, allowing faster iterations of suggestions, so Luminance can quickly narrow in on the concept the reviewer finds to be relevant. This has been especially useful when developing Luminance’s language-agnostic model.

Another key example, particularly for Luminance Discovery, is Named Entity Recognition (NER). The intention with NER is for a model to go through all text identifying the starts and ends of, for instance, names, addresses and dates. Luminance uses a variety of techniques to deal with these. Simple examples of named entities, such as email addresses and NHS numbers, can be found via regex and checksums. More complex entities like company names are identified using a spaCy model which has been extensively trained over a legal dataset. Most recently, Luminance has developed a legal definition extractor which uses our own legal transformer models to power spaCy predictions.

score: 0.39

this oem agreement may only be amended or modified in a written document signed by authorized representatives of both parties

score: 0.41

whereas hastings has been employed by the company as the chief executive officer and chairman of the board pursuant to an amended and restated executive employment agreement effective august as amended by the first amendment to the amended and restated executive employment agreement and second amendment to the amended and restated executive employment agreement as amended the employment agreement attached hereto as exhibit a

score: 0.49

a amendment no to apiant corporations software oem enterprise license agreement effective september amendment

score: 0.63

this amending agreement made the st day of july the amending agreement between university of victoria innovation and development corporation a corporation owned by the university of victoria having its principal office at shut mckenzie avenue victoria british columbia canada vw w idc and the john hopkins university a nonprofit corporation duly incorporated under the laws of maryland having an office at n charles street baltimore maryland united states of america thu and therapeutics inc

score: 0.64

this amending agreement made the st day of july the amending agreement

score: 0.71

therefore in consideration of the mutual agreements provisions and covenants contained in this amendment service provider and client hereby agree that the agreement be and hereby is amended as follows

score: 0.78

amendment this oem agreement may be amended or modified only in written document signed by authorized representatives of and

score: 0.81

this first amendment to the employment agreement is entered between voicephone limited and luminance

score: 0.83

amendment restates amends modifies fifth contained slt

score: 0.85

this amended and restated employment agreement agreement dated as of april is entered into between reinsurance ltd an alberta corporation having its principal place of business at drive north vancouver bc vp s employer and darryl drawings an individual residing at

score: 0.86

this amendment and supplement no this supplement dated as of august amends and supplements the amended and restated security agency agreement the sara dated as of october among bank of america na bank of america as global administrative agent as defined therein on behalf of the global lenders as defined therein certain other creditors or the representatives of such creditors of prologis a maryland real estate investment trust prologis and bank of america as collateral agent as defined therein

score: 0.87

this amendment no to lease agreement amendment no is made and entered into as of the st day of august by and between allergan coal land company a maryland corporation owner and patriot mining company inc a west virginia corporation company

score: 0.88

this amendment this amendment dated june amendment effective date is entered into by and between hca information technology services inc a tennessee corporation its which is a wholly owned subsidiary of hca healthcare corporation a delaware corporation hca formerly known as columbia information systems inc and life point corporate services general partnership a delaware general partnership together with its successors and

*This is an example of how Luminance’s transformer model scores text according to similarity.*

# Transformers

Transformer models have recently come to dominate the field of NLP and have already had a huge impact on machine learning. Luminance has adapted transformers to legal-specific tasks, enabling an efficient use of these models in the context of legal document review and process automation. These state-of-the-art models move away from static word embeddings and instead create contextualized word embeddings. This is where a word's meaning is changed depending on the context in which it is used.

Transformers work by the repeated stacking of a Multi Headed Self Attention block. This processing block allows the model to update a word's meaning depending on the other words in the sentence. Each 'Head' will learn a different viewpoint of the data and incorporate different information into the word. Unlike the RNN, which does this operation sequentially, the transformer is free to use any other word in the sentence, making it well-suited to longer sequences.

**"In the eDiscovery market, where cost is a high priority, this is a gamechanger."**

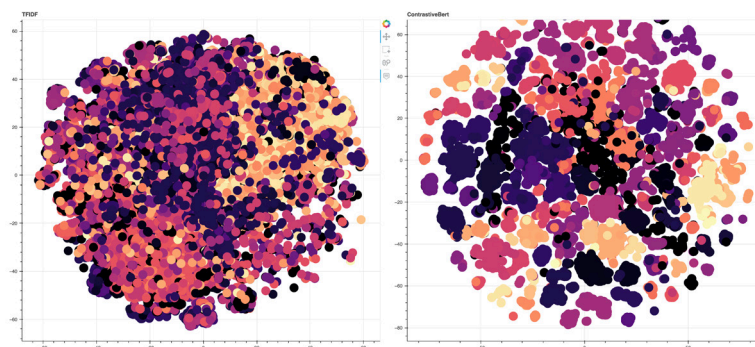


**Pontus Scherp**  
Partner

Norburg & Scherp

These transformer models are trained over huge corpora of text in the public domain, with the masked language modelling objective. The goal, similar to that for Word2vec, is to predict missing words, which is again a sufficient training objective for the model to learn a great deal about language. These models can then be fine-tuned to a specific language domain. In Luminance's case, the training is continued on a huge set (500k+) of unlabelled legal documents so that the model may acquire a special legal understanding. Indeed, building neural networks at scale requires vast data flows: Luminance has now seen and analysed over 150 million documents in over 80 languages across hugely diverse fields of law, from Non-Disclosure Agreements and property leases to emails and WhatsApp messages.

Once Luminance has developed legal-specific language models, they can then be fine-tuned for certain types of legal-specific tasks, such as document classification, clause classification and Named-Entity Recognition (NER). They can even be used to vectorize whole sentences, paragraphs, and documents.



*This comparison is between a standard (relatively simple) method for embedding sentences into a vector space on the left and one of our latest sentence transformer models, which can be seen to group sentences with similar meaning (points of the same colour are more clearly clustered together).*

# AI in Practice: Luminance's AI-Powered Features

The described techniques make Luminance the most robust and powerful platform for legal document analysis and process automation. Indeed, Luminance's core technology powers multiple applications, from taking a first-pass assessment of incoming contracts to automatically flagging contractual anomalies and areas of non-compliance and applying advanced AI-driven ECA and TAR within eDiscovery matters.

## SUPERVISED MACHINE LEARNING AND BESPOKE CONCEPT CREATION

Luminance can be taught new examples of concepts via a simple point-and-click process as part of the normal review workflow, allowing for bespoke tags for projects, clients or business areas. When users add a new label to a clause, teaching Luminance what the clause 'means', Luminance then suggests further conceptually similar clauses that might be relevant to the concept that the clause represents.

Crucially, Luminance's supervised machine learning means that the system can stay current as the company undergoes change, for example, when entering new markets with different regulatory standards, or when unexpected global events require swift evaluation of contractual risk exposure. Users simply need to tag one example of the new standard, and Luminance can then compare how all other examples comply or deviate from that position. Indeed, by adopting a machine learning approach, Luminance ensures that there is no need for downtime to perform costly maintenance and businesses can rapidly find answers to urgent business questions.

**Create Party Tag**

Customer Rep

Contact: **Daria Perez** Address: Semiloo Inc, 1182 Cotton Abbey, Tabasco, New Jersey, 08875- 7181, US

2000-07-01: Semiloo Inc Sales Agreement.docx

Party: **Daria Perez**

**Similar Tag Suggestions**

☐ Select All

☒ Create tag with type *Customer Rep*

Contact: **Van Johnston**

don't suggest again

Premill Inc Sales Agreement.docx

Party: **Van Johnston**

☐ Update type from *Generic Party* to *Customer Rep*

Address: **Coranix Inc**, 2732 Dusty Street, Rapture, Texas, 78079-1403, US

don't suggest again

Coranix Inc Sales Agreement.docx

Party: **Coranix Incorporated**

☒ Create tag with type *Customer Rep*

Contact: **Madeline Rowe**

don't suggest again

Coranix Inc Sales Agreement.docx

Party: **Madeline Rowe**

**+ Create 3 Tags**

*Users can add any number of additional clause models and Luminance's AI will apply this learning across the entire dataset.*

## CASE STUDY



When the in-house legal team at Colombia's flagbearer airline, Avianca Airlines tagged an important 'Proveedor' (provider) party clause in Spanish, Luminance applied this learning to the rest of the document set, instantly flagging all other examples across **1,000** other contracts.

**"With Luminance we no longer have to rely on external counsel for complex reviews but can instead keep the work in-house."**



**Daniel Felipe Morales Martinez**  
Former Contract Manager

## CASE STUDY



The in-house legal team at world-leading biotechnology company, IDEXX Laboratories, used Luminance to analyse its entire contract database containing **20,000 documents** following the sudden introductions of economic sanctions in March 2022.

With Luminance's AI providing a holistic overview of the organisation's business activities, identifying all geographies present within contracts and any language referencing Russian places or legal structures, IDEXX completed the review in just **20 minutes**.

**"With Luminance's AI, we can see exactly where we need to focus our attention in negotiations."**



**Matthew Forsyth**  
VP and Deputy General Counsel

## AUTOMATIC ANOMALY DETECTION

Luminance's unsupervised machine learning automatically identifies anomalies present in the data room and categorises them according to severity, type, and number, allowing legal professionals to effectively prioritise their review. These anomalies are often the 'unknown unknowns' - the hidden risks of which the lawyer was unaware, such as a missing document or minor deviation within a clause, and thus did not think to actively search for.

### CASE STUDY

大成 DENTONS

The world's largest law firm, Dentons, found inconsistencies in the wording of indemnity clauses whilst conducting a due diligence review with Luminance. Without Luminance's anomaly pattern-detection capabilities it would have been incredibly difficult and immensely time-consuming for the team to manually identify the minor deviations in language across the set of 100 near-identical documents.

**“The greatest benefit of Luminance lies within the immediate insight into the contents of the virtual data room. This allowed us to provide a more sound product to our client.”**



**Nick de Rooij**  
Associate

## AUTOMATIC CONTRACT CLASSIFICATION WITH OVER 1,000 LEGAL CONCEPTS

Luminance's AI works out-of-the-box to identify over 1,000 different concepts within executed contracts, such as clauses, document types, party names and governing law. This means that when exposed to a new dataset, Luminance is able to automatically recognise this information and display it back to the lawyer across a series of intuitive 3D visualisations.

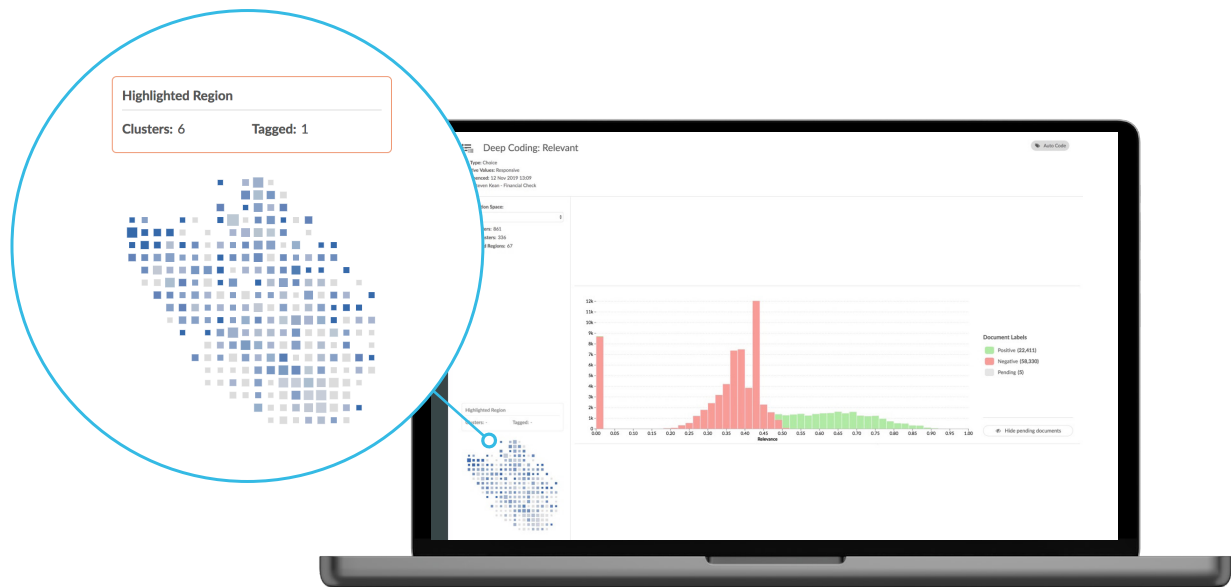
## LANGUAGE AND JURISDICTION AGNOSTIC

Luminance's advanced pattern recognition technology means that it is language-agnostic, ensuring lawyers can use Luminance on documents in any language and even on datasets containing multiple languages.

## DEEP CODING AND TECHNOLOGY ASSISTED REVIEW (TAR)

Luminance's Deep Coding functionality is a powerful application of machine learning to the document review process, whereby the system actively learns from the lawyer's expertise when identifying which documents are relevant to the investigation. Luminance's AI can observe the coding actions of the team and then flag where the coding of a document appears anomalous compared to how similar documents have been coded in case this warrants a further review. Luminance can also predict how relevant yet-uncoded documents will be, ensuring lawyers can focus their time effectively on the most relevant documents.



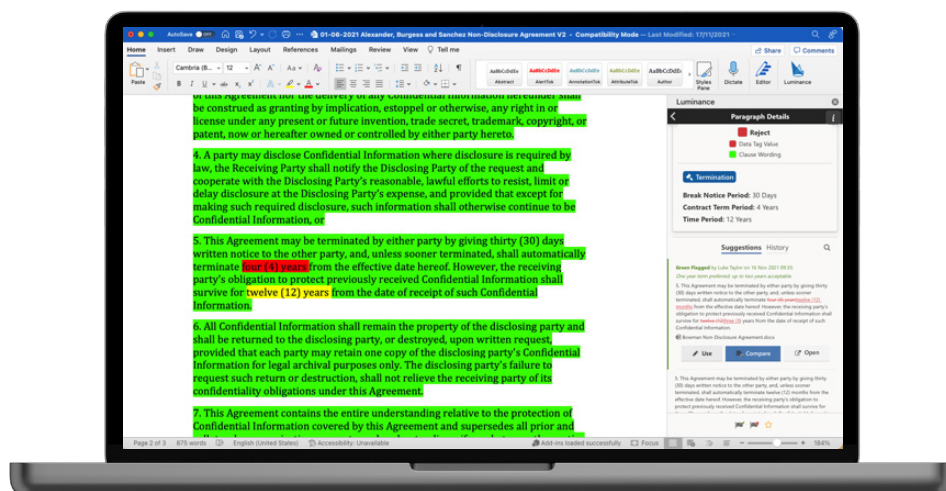


*Luminance's AI organises the dataset by conceptual similarity, displaying the results of its analysis in an interactive Heat Map.*

With no prior training needed, Luminance's Heat Map organises the entire dataset by conceptual similarity, with each tile representing conceptually alike documents. The depth of colour of each tile illustrates how many of these documents have been reviewed, making sure lawyers can allocate resources to under-investigated areas quickly so nothing is overlooked.

## TRAFFIC LIGHT ANALYSIS: FIRST PASS REVIEWS OF INCOMING CONTRACTS

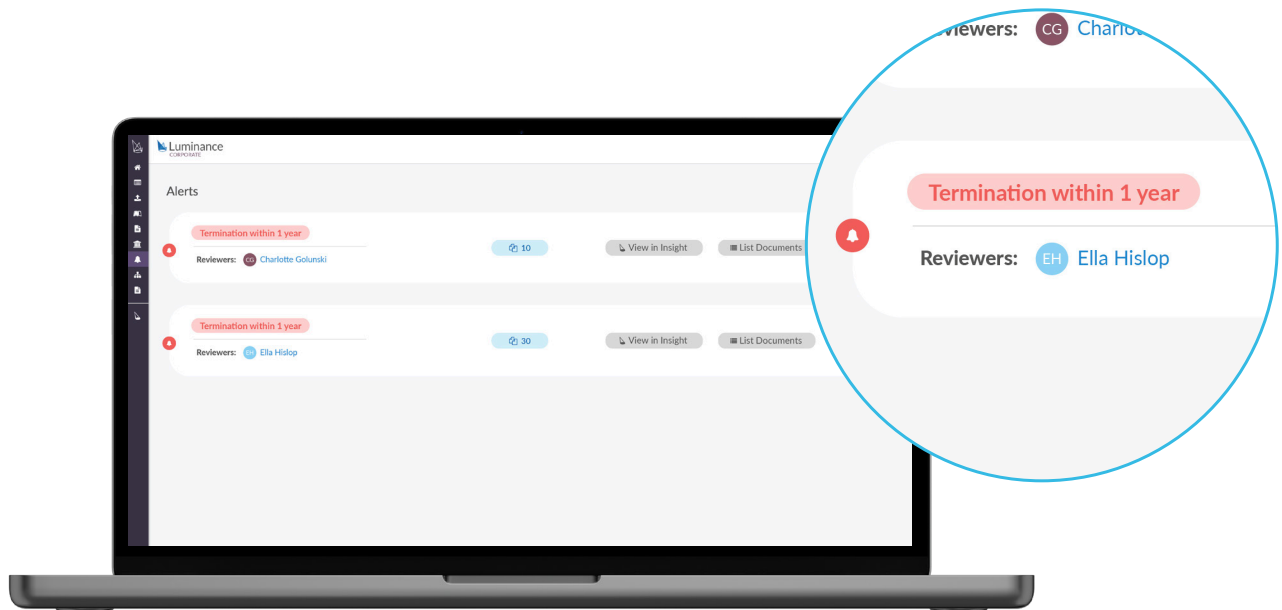
Luminance learns from an organisation's approved contracts and visually indicates via a Traffic Light Analysis colour-coding system where terms differ from previously agreed wording, highlighting within a contract which clauses are acceptable (green), in need of further review (amber) or non-compliant (red). This allows lawyers to instantly understand where to focus resource, reducing reliance on the time-consuming and outdated practice of manually redlining contracts. Where terms do not meet internal standards, alternative acceptable wordings are provided from internal 'precedent banks' of approved wording.



*Luminance's AI will take a first pass review of any document, automatically highlighting areas of compliance and risk via a colour-coded system.*

## AUTOMATIC ALERTING FUNCTIONALITY

Luminance reads and forms an understanding of contracts, automatically highlighting areas of risk or business opportunity, for instance flagging key periods within contracts such as commencement dates, breaks and expiration.



*Automatic alerts can notify users of key contractual periods, including commencement dates, breaks and expiration.*

## EARLY CASE ASSESSMENT (ECA)

With AI-powered ECA, lawyers can quickly and easily apply powerful searches and filters to cull irrelevant documents and find what matters. Luminance understands documents on a conceptual level, allowing lawyers to search for conceptually similar documents to the one under review in just one click. Lawyers can easily navigate between documents, including support for Microsoft Excel, WhatsApp and Skype Conversation files.

## AUTOMATIC PII DETECTION AND REDACTION

With data privacy regulations such as GDPR and CCPA posing a major compliance challenge to organisations, lawyers can now feel confident that nothing confidential goes unredacted as Luminance will automatically recognise patterns of language to highlight text that may constitute PII within a document.

## About Luminance

Luminance is the world's most advanced AI technology for the legal processing of contracts and documents. Founded by mathematicians from the University of Cambridge, Luminance's AI reads and forms a conceptual understanding of documents in any language. Luminance uses this understanding to augment the spectrum of legal matters, from AI-powered contract drafting, negotiation and review to investigations and eDiscovery. Luminance is used by over 500 customers in 60 countries, including all of the Big Four consultancy firms, a quarter of the world's largest law firms and multinational organisations such as Tesco and Ferrero.

[info@luminance.com](mailto:info@luminance.com)

[www.luminance.com](http://www.luminance.com)

