QuisLex

Project Johnny 5

A Case Study on Evaluating AI Abstraction Tools

By Chase D'Agostino, Associate Vice President of Corporate Solutions at QuisLex

This project would not have been completed without the tremendous support from the QuisLex team in our Center of Excellence in Hyderabad, India. Special thanks to:

- *our Project Manager **Usha Padmini,** who spent countless hours driving the project forward while also managing her daily workload*
- *our standout reviewers, including Team Leads **Nihar Mathur** and **Akshat Kaul,** and Project Manager **Nirmalya Banerjee***
- *an exceptional administrative team lead by **Sail Jamaalpuri** and **Ravi Kumar***
- *our Associate Vice President of Client Solutions, **Sailaja Meesaraganda,** particularly for her support helping to get the project off the ground and her deep substantive reviews to ensure consistency when scoring results.*

*Special thanks also to our Chief Technology Officer **Michel Sahyoun** and our General Counsel & Vice President of Legal Services **David Klein,** for their continued advice and guidance throughout the project. Michel and David have been evaluating artificial intelligence in the legal space since the very first tools were released years and years ago!*

We codenamed our Project "Johnny 5" in homage to the classic 1986 film Short Circuit about an adorable robot who achieves sentience.

PROJECT JOHNNY 5
A CASE STUDY ON EVALUATING AI ABSTRACTION TOOLS

"We are pleased that QuisLex has taken the time to educate the market about these products. We think they have done a nice job articulating the benefits of these tools while also keeping it grounded in the reality of where things are today and where it could go from here." ▶ **Kevin Miller, CEO & Board Member at Legal Sifter, a participant in QuisLex's study**

We took a deep dive on ten leading Artificial Intelligence tools built specifically to abstract text from contracts. Our findings:

1. **Time Savings are Real but No "Easy" Button.** You can materially improve human performance by integrating these tools into your contract workflows. Time savings averaged 28% in our proof of concept. But we also safely conclude that legal professionals will not be replaced en masse by robots anytime soon.
2. **Cost and Time Savings Vary.** Time savings and accuracy had a high degree of variance, not just from tool to tool but within tools, from one agreement to the next, and from one provision to the next – including some instances where using the tool took more time than manual review. The usefulness of each tool is dependent on the tool/project fit and the process built to support the workflow. Cost savings and ROI will also vary depending on the cost of the resource whose time is being saved. Implied cost savings for a 10,000 contract review project ranged from $33,000 to upwards of $1.7M depending on whose manual review time was being reduced.
3. **"Training" AI Models is a Skill.** Building models to cover new concepts or to target specific agreement sets involves thinking through how you want the model to perform to ensure you are giving it the right training data and building a process to ensure that only the right training data gets added to the model.
4. **Tools will Get Even Better.** Improvement opportunities exist, including for handling of amendments and functionality to better support the entire workflow.

What follows are the results of our in-depth analysis. This is not a buyer's guide – we are not at liberty to reveal which company is which. Even if we could speak freely, our conclusions would quickly become stale as incumbents improve and new players enter the market. Rather, this case study is intended to assist readers in asking better questions when commencing their own purchase journey so they are able to find the right tool for their use case.

"Legal departments have no problem kicking off a technology selection but can struggle to understand whether the tool will 'work as intended' and to compare one tool to the next. Putting structure and process around the evaluation helps make sure you don't get stuck. Defining criteria and specific tests help you objectively compare functionality and results," notes Catherine J. Moynihan, Associate Vice President, Legal Management Services and Director of ACC Legal Operations at Association of Corporate Counsel. "If you can start from a framework that others have successfully used to drive their selection process, it means you don't have to recreate the wheel but instead can focus your time on refining the approach based on your specific requirements."

WHY ABSTRACTION PROJECTS ARE IMPORTANT

Companies struggle with inefficiencies on a daily basis as they respond to ad hoc questions from stakeholders about legacy agreements: "where is the contract I signed last year with Vendor X," "when does the contract with Customer Y expire" or "can we terminate the agreement?" Most companies have not captured the information in a system to quickly pull answers to questions like these – a human has to find, read and properly interpret the contract.

These inefficiencies are amplified when someone asks at a broader scale "what do ALL our contracts say?" Due diligence is an obvious scenario. In an acquisition, the purchasing company needs to know which obligations it is absorbing. But M&A is only the tip of the iceberg. Projects with contract volumes that often dwarf a diligence review include:

- Building a repository from legacy agreements in advance of new CLM system implementations
- Identifying contracts that need to be amended in response to a regulatory event (e.g., GDPR, Brexit)
- Determining when revenue contracts are up for renewal or renegotiation
- Pinpointing contracts with non-standard terms to better characterize risk exposure
- Analyzing notice and consent issues for internal reorganizations

Businesses should know what is in their contracts. As Tim Cummins, President of the International Association for Contract & Commercial Management (IACCM) puts it: "IACCM research has shown that almost 90% of business users find contracts difficult or impossible to understand. Today's agreements contain critical operational data which is too often imprecise or overlooked. As a result, some 40% of contracts suffer from serious disagreements between the parties at some point during their performance. This costs time and money – big time!"

For more information see Appendix 1.

Contract abstraction is about giving structure to contracts, which are inherently unstructured data. Yet the size and scope of these abstraction projects can be daunting. Manually abstracting data points from a large volume of contracts is time consuming (read: slow and expensive).

OUR TESTS AND FINDINGS

The question we set out to answer is whether the introduction of technology could make the abstraction process more cost-effective – reducing labor/cost or improving speed/quality/consistency. We invited eleven leading technology vendors to participate in our proof of concept. All eleven are commercially available technologies that are using approaches like natural language processing and machine learning that fall under the "artificial intelligence" umbrella. We did not look at homegrown solutions some legal departments have successfully spun up for their own use. We also did not look at tools that were specifically built for a particular subset of agreements (e.g., only meant to handle ISDA agreements).

Ten vendors stepped up to the plate. The one vendor that refused to participate in a "bake off" – their words – was already familiar to us. We had previously used their tool on an 80,000-document abstraction project. The results using that vendor are consistent with our results reported below. Our real-world results using abstraction tools on live-projects at a significantly larger scale than our proof of concept are also consistent with our results reported below. At the request of one vendor, their results were excluded from the below.

Test 1: Functional Requirements

*Click **here** to download our functional requirements template.*

We identified ~150 functional requirements in our evaluation template that was completed by all ten vendors in a self-assessment. Seven of the vendors gave us access to live environments. For these vendors we validated their responses to our template in the live system. Examples of functional requirements we looked at include:

- Drag-and-drop upload
- De-duplication
- Batching (identifying groups of documents and assigning them to reviewers)
- Bulk export of documents
- Bulk export of metadata
- Boolean search capabilities
- Breadth of out-of-the box AI models
- Capabilities around building custom/bespoke AI models
- Security
- Integrations

▶ Findings:

In raw numbers, the top vendor met 70% of our functional requirements while the lowest performing vendor only met 33%. The median vendor met 52% of our functional requirements. Of particular note, four vendors met 100% of the requirement around exporting data from the system and five vendors met more than 70% of the requirement around security. There was more variance for the requirements around building custom/bespoke models, with several vendors meeting less than 20% of the requirements and the leading vendor meeting 78% of the requirements. Of course, the importance of any of the variables will depend on the priorities of the company purchasing the tool and particulars of the projects they want to use it on.

For our own internal selection purposes, we weighted requirements based on what was most important to QuisLex's use cases. For example, if a functional requirement was deemed "Critical" to our selection, it might get a weighting of 5, whereas if a functional requirement was deemed a "Nice to Have," it might get a weighting of 1. Under this approach, out of a total possible points of 671 (i.e., if vendor had met every functional requirement "Out of the Box"), the top vendor scored 544 (81%) and the lowest vendor scored 290 (43%), with the median vendor scoring 473 (70%). Importantly, the highest and lowest scoring vendors were different after our weighting.

Insight from the Industry:

"There is an ocean of contract AI abstraction technology out there. QuisLex has developed the most granular and articulate evaluation matrix for such technologies I have seen. I applaud them for making it available to the market and think it is a valuable tool to help streamline a selection process, allowing you to pick and choose the requirements that fit the specifics of your projects." **▶ Joshua Walker, CEO of Aleph.Legal and author "On Legal AI" (Forthcoming, Feb. 2019)**

Test 2: Time Savings with Out-of-the-Box Models

*Click **here** to download the template that was completed for each contract.*

*Click **here** to download the details of our process and parameters/controls.*

We tested five of the tools to validate our hypothesis that automated contract extraction adds value. First, we applied our standard abstraction and quality assurance protocols to 48 commercial contracts, including Servicing Agreements, Supply Agreements, License Agreement, POs with Terms of Service and Joint Development Agreements. We had six reviewers (Reviewers 1-6) each manually abstract 47 fields from groups of eight contracts (Groups A-F). We then trained the reviewers on the abstraction tools and incorporated those tools into their abstraction workflow – i.e., the tool would perform what extraction it could and the reviewer would check the accuracy, fix mistakes, and fill in blanks. Each reviewer then completed tool augmented reviews for groups of contracts. No reviewer saw the same contract twice. We tracked time at a document level and summed total time required to review the 48 contracts in each tool.

▶ Findings:

The average time savings of a tool assisted review was 28%, with a range of 16% for the least helpful tool to 36% for the most helpful. For specific documents, the time savings got as high as 64%, but for other documents the tool assisted review took longer than the standard manual process.

	Tool 1	Tool 2	Tool 3	Tool 4	Tool 6
Time Savings	**16%**	**24%**	**32%**	**36%**	**33%**

Insight from the Industry:

"We have found that time savings are even better on subsequent projects when teams get more comfortable working in a system and are also able to benefit from models that they train specific to their contract sets. This enables our large entity clients to scale their domain expertise and create a competitive advantage." **▶ Ned Gannon, Co-Founder & President of eBrevia, now part of Donnelley Financial Solutions (DFIN)**

"Our users typically report 20-40 percent time savings their first time using the software. We routinely see that increase to significantly higher levels as teams get experience with Kira." **▶ Steve Obenski, Chief Product Officer at Kira Systems**

Test 3: Accuracy of Out-of-the-Box Models

For 5 tools, we also tested how well their out-of-the-box models performed. "Out of the box" in this context are artificial-intelligence-based models trained by the technology vendors to automatically find and extract specific text. For example, one model might be trained to find and extract assignment clauses and another to find and extract limitation of liability provisions. Tools 1-4 overlapped with the tools in our time test and Tool 5 was added. We tested an average of 17 models, which represented ~36% of the 47 total data points required to complete our review template. At the upper end, two tools had 20 relevant models (about 43% of the total data points required), while one tool at the low end had 10 models (about 21% of the total data points required). If a tool did not have a model out-of-the-box model to cover a concept, that concept was excluded from the scoring for that tool.

We maintained total control over the environments we tested in and completed all steps in the process ourselves (i.e., no vendor intervention). Results of the models were reviewed at the highest level within QuisLex to ensure consistency when scoring results.

To get a baseline comparison of how "accurate" humans are versus machines, we also separately completed the review template for the full 48 agreements with entry-level resources that had been trained on contracts generally but did not have prior experience working on contract abstraction projects.

Importantly, not every agreement contained every provision. So part of the testing was whether the tool left blanks where blanks were appropriate.

▶ Findings:

For details on the approach we followed when coming up with the scoring, see Appendix 2.

Where there was no provision to extract (should have been blank):

	Untrained Human	Tool 1	Tool 2	Tool 3	Tool 4	Tool 5
Right (blank)	**99%**	**75%**	**86%**	**98%**	**92%**	**97%**
Wrong (abstracted text)	**1%**	**25%**	**14%**	**2%**	**8%**	**3%**

Where there was a provision to extract:

	Untrained Human	Tool 1	Tool 2	Tool 3	Tool 4	Tool 5
Right (extracted all, and only, correct text)	**60%**	**69%**	**53%**	**64%**	**64%**	**49%**
Partially Right (extracted all correct text plus some wrong text)	**5%**	**10%**	**7%**	**5%**	**4%**	**9%**
Wrong (extracted some correct text but missed other correct text)	**15%**	**11%**	**15%**	**13%**	**7%**	**12%**
Wrong (did not extract any correct text)	**21%**	**10%**	**25%**	**18%**	**25%**	**31%**

For additional calculations that may be closer to how people in machine learning circles expect to see results (e.g. "Precision" and "Recall"), see Appendix 3.

Looking at the results for specific provisions, the best performing out-of-the-box models were for governing law and assignment, finding the correct, and only the correct text, an average of 78% and 69% of the time, respectively, across all tools and 100% of the time for both provisions in the best-performing tool. Some of the lower performing models were the notice provisions, limitation of liability, and insurance, each returning the correct, and only the correct, text an average of less than 50% of the time across all tools and about 60% of the time in the best-performing tool.

Insight from the Industry:

"It's important to have a lot of built-in models in a tool so that you can get started and find value quickly, even before you train the tool with your own expertise. A good contract analysis tool should be able to learn what you need even when the contracts are diverse. Many technologies can appear to work well when clause language is similar. Better technologies can perform well when the wording is very different from contract to contract." ▶ **Steve Obenski, Chief Product Officer at Kira Systems**

Test 4: Bespoke/Custom Model Testing

We then built and trained our own models in four systems and tested how those models performed. The four tools tested overlapped with four of the tools from our out-of-the-box testing. The training data came from the 48 already abstracted agreements from Test 2.

Admittedly, this is a small sample size. The objective of this test, however, was not ultimate performance after robust training but to see how quickly the benefits of training could be observed and thereby validate whether the artificial intelligence was "learning."

We tested our trained models on 12 new commercial contracts of the same type as those from Test 2. We tasked the tools with extracting eight fields:

- Document name
- Effective date
- End date
- Governing law
- Limitations of liability
- Assignment
- Non-solicit/Non-compete
- Arbitration

If the minimum number of positive examples required to train a model in a tool was not met for any particular provision, that provision was excluded from the scoring of that tool.

Findings:

Where there was no provision to extract (should have been blank):

	Tool 2	Tool 3	Tool 4	Tool 5
Right (blank)	**88%**	**97%**	**100%**	**33%**
Wrong (extracted text)	**12%**	**3%**	**0%**	**67%**

Where there was a provision to extract:

	Tool 2	Tool 3	Tool 4	Tool 5
Right (extracted all, and only, correct text)	**24%**	**23%**	**82%**	**30%**
Partially Right (extracted all correct text plus some wrong text)	**10%**	**52%**	**13%**	**7%**
Wrong (extracted some correct text but missed other correct text)	**16%**	**13%**	**0%**	**7%**
Wrong (did not extract any correct text)	**49%**	**12%**	**4%**	**57%**

Looking at the results for specific provisions, the best performing model was for assignment, finding the correct, and only the correct, text 65% of the time across all tools and 92% of the time in the best-performing tool. Of particular note, one tool accurately abstracted the specific governing law (not just the section/paragraph but the actual data point) 83% of the time.

Insight from the Industry:

"Contract abstraction involves abstracting two types of data from contracts – Clauses and Metadata/Data Fields. Machine Learning algorithms are more resilient and 'forgiving' in accurately abstracting clauses than in abstracting Metadata. Abstracting Metadata requires training over a larger training set than what is needed for abstracting clauses."

▶ **Murali Tirupati, Co-Founder & CEO of Vaultedge**

Test 5: Feedback

Finally, we put together a 30-question survey to collect subjective feedback from our reviewers on user interface, user experience, time savings, error reduction, etc.

▶ **Findings:**

Overall, reviews were positive but mixed from one reviewer to the next and from one tool to the next. Average responses on some key topics:

	Tool 1	Tool 2	Tool 3	Tool 4	Tool 6
The user interface was intuitive and easy to use					
The status of my review was easy to see (which contracts I had reviewed and which I had not)					
Seeing both the contract I was reviewing and the fields to be populated in one interface was helpful					
Searching for text/key-words was easier in this Tool, and results were more successful than searching on other projects I have worked on					
The review process in this Tool is better than in other projects I have worked on					
Learning how to use the Tool without training is easy					
As a review level resource this Tool will save me time					
As a review level resource this Tool will result in fewer mistakes					
This Tool required non-substantive cleanup (formatting or typos in extracted text)					
This Tool required substantive cleanup (incorrect provision extracted or incomplete extraction)					
This Tool required no cleanup of extracted text					
Tool required less cleanup than typical abstraction from an entry level resource					
Overall user sentiment based on comments					

Insight from the Industry:

"The best tools in the market will obsess over user experience and understanding the full shape of a user journey, including all emotions along the way. Large and complex contract review projects are wrought with varying painful challenges every step of the way. What are our users' emotions when they can't easily bulk import/OCR contracts or prioritize which need analysis/review? How do they feel when setting up groups, permissions, batches, or configuring hundreds of bespoke fields across varying formats and categories to store metadata? What about when they can't re-use all of these configurations for the next project? Delivering delight at every leg of this journey will result in stickier, no-brainer products that not only make customers happier, but move this entire industry forward." ▶ **Charlie Connor, CEO & Co-Founder at Heretik**

DISCUSSION & LEARNINGS

"Contract reviewers still play an essential role but tools like eBrevia can enhance the reviewer by automatically abstracting provisions in contracts to help finish the review faster and with a higher degree of accuracy, allowing contract reviewers to focus on quality control and higher value judgment based work. These tools are excellent at assisting with the first level of review but are not intended to replace the human review process."

▶ **Ned Gannon, Co-Founder & President of eBrevia**

No "Easy" Button but Time Savings are Real. Lawyers are safe. There is no "easy button" ready to displace the manual work performed in abstracting material information from contracts.

To take one simple example, the tools are still inconsistent when it comes to turning scanned PDFs into machine readable text. OCR (optical character recognition) engines are imperfect and therefore leave some parts of some documents "invisible" during the automated extraction process. In practical terms, those poor quality scans sitting in your legacy repository will be harder to run through an AI tool.

That said, treated as tools to be used by humans, rather than as substitutes for humans, AI-based abstraction technology can deliver considerable time savings. The tools need not be perfect to be useful. They can be especially useful on projects that do not demand absolute precision. If your risk tolerance is such that 70% or 80% accuracy is good enough, you can save significant time and money by investing in one of these tools. Even where flawless execution is expected, these tools can be fit into a well-designed, human-centric workflow and pay almost immediate dividends. But the savings are contingent.

Cost and Time Savings Vary. The savings are contingent on making the upfront investments in finding the right tool/project fit and then designing a workflow that optimizes the benefits from the tool while keeping humans engaged to fill in the gaps and perform quality checks. The tools do a serviceable job finding and extracting text. But for most use cases this is just the beginning of a review. The extracted language still needs to be interpreted, analyzed, and/or compared against a standard with deviations flagged. For example, extracting a full limitation of liability provision might be less useful than an adjacent field categorizing the provision as "Capped," "Capped with Exceptions" or "Uncapped." The tools are mostly reliable on the abstraction but a long way from being able to perform the categorization.

"Clients and business demand absolutes – but the accuracy of any human or AI solution has its limits. Too often the focus is on the accuracy of the AI tool – leading to wasting training cycles yielding marginal improvement in accuracy. AI accuracy has therefore become the red herring of the industry, taking you away from looking at the business benefit. We advise focusing on using the right combination of human & machine to reduce the 'Time To 100' for any process to the shortest time possible. It means we use our intelligence to determine if heuristic extraction is more accurate in a shorter time than machine learning. It means we stop training where it results in marginal gain and put the information in front of a human so they can use their experience to make the final determination." ▶ **Nick Thomson, General Manager, iManage RAVN**

Moreover, variation in tool performance is largely driven by the variation in our contracting processes. Unless an organization continually drives counterparties to sign unedited form agreements, language will differ from one agreement to the next. Tim Cummins, President of IACCM, observes, "The truth is that contracts are unnecessarily complex due to the archaic and idiosyncratic approach that businesses take to their formation and management. Research has shown that many agreements – for example Master Services Agreements – are based on almost identical commercial principles. What differs is simply the words – and the words are based on the personal preferences of the individual lawyer drafting the terms." Tim continues, "this likely is not what the business wants – to compete on words, but until more progress is made on standardization efforts in the industry this is a reality that has to be dealt with in a typical review project." The time a reviewer spends analyzing nuances in variable language dwarfs the time it takes to extract the language itself.

Savings are also contingent on the cost of the time saved. Today's legal departments use a range of resources to support legal tasks. As noted by Catherine J. Moynihan, Associate Vice President, Legal Management Services and Director of ACC Legal Operations at Association of Corporate Counsel, "Today's legal departments drive efficiencies by using the right level of resource to support the right tasks, given complexity, risk and cost. Law firms will always participate, but legal departments continue to add internal legal professionals from diverse disciplines and are working with external legal service providers. These tools give teams more levers to pull when considering who should be supporting a particular project or task, and a means to reduce personnel time."

Legal departments maximize efficiency by allocating low-value, low-risk repetitive work to lower cost resources like in-house shared servicing centers or legal process outsourcing companies, and more complicated higher value work to senior internal lawyers or external law firms. Where contract attorneys, paralegals, and offshore resources handle much of the labor-intensive aspects of abstraction, ROI on deploying abstraction technology can be modest – though still significant at scale. For example, extrapolating the average time savings from our proof of concept to a 10,000-contract review project, implied savings range from $33,000 to upwards of $1.7M depending on labor costs, with a break-even rate of $11/hr. Meaning unless your labor costs are lower than the new New York State minimum wage, you should see an ROI from using an AI abstraction tool.

	Big Law 1st-Year Associate	Laffey Matrix (1-3 Years out of School)	Big-Law Non Lawyers	In-House Counsel Level (US Average)	In-House Corporate Paralegal (US Average)	Hourly Labor ($30)	Hourly Labor ($20)	Break-Even Hourly Rate
Hourly Rate	$485	$359	$149	$123[1]	$50[2]	$30	$20	$11
Time Savings / Contract	Approximately 33 Minutes							
Cost Savings / Contract	$176	$130	$54	$45	$18	$11	$7	$4
Cost of Tool / Contract	$4.00[3]							
Total Savings	**$1,722,913**	**$1,264,919**	**$501,596**	**$405,775**	**$141,744**	**$69,046**	**$32,697**	**$–**

[1] Including (i) salary and additional cash comp per https://www.glassdoor.com/Salaries/us-counsel-salary-SRCH_IL.0,2_IN1_KO3,10.htm, (ii) assumed 30% salary cost for benefits and (iii) assumed hours per year of 1920.

[2] Including (i) salary and additional cash comp per https://www.glassdoor.com/Salaries/us-corporate-paralegal-salary-SRCH_IL.0,2_IN1_KO3,22.htm, (ii) assumed 30% salary cost for benefits and (iii) assumed hours per year of 1920.

[3] Representative and not based on any specific Tool. Does not account for upfront and ongoing license/maintenance costs (e.g., if system was maintained "on-premise").

"Before you begin training a model it is critical to define your training objectives, by identify the concept that you are looking to identify and the desired output. What do you expect the system to pull when you upload a contract containing relevant language? What types of contracts do you expect the system to analyze? This exercise will help you develop guidelines and compile a representative set of training contracts."

▶ Preethy Prakash, Director of Account Management and Corporate Counsel at eBrevia

Effort Required to "Train" AI Models In theory, the machine-learning component of many of these tools should result in increased savings over time. As the machine learns from more examples the models perform better. But training a machine must be intentional and can be labor intensive.

While we did not fully test this as part of our initial proof of concept, we know from using these tools on live projects that workflows have to be designed in ways that maximize the learning, even where this entails extra steps or other tradeoffs. Maximizing machines involves thinking through, in detail, how you want the model to perform to ensure you are giving it the right training data.

As an example, do you want to create a model that is more likely to find information for your reviewer, even though it may often be wrong? Or do you want to limit the times the model finds language but know that when it does find language it is more likely to be correct? The answers depend on the specifics of your review and will drive how you want to build your process to train the models.

There is also no magic number for the amount of training required to get to a certain level of performance. For example, how many extra examples do you have to give a model to increase the "accuracy" from 50% to 60%? Where language was more inconsistent from one agreement to the next, the general consensus from vendors was that more examples were helpful to teach the model the different variations. All vendors cautioned that the 48 documents we used for training, which were fairly diverse, was insufficient to build reliable models, particularly for concepts with only a handful of positive examples in the training set (e.g., not many of the agreements had non-solicitation/non-compete provisions). But giving more examples to train a model is not always helpful either, particularly where a new example might be a one-off outlier to an otherwise consistent set of examples.

Surprisingly, some of the data we might assume would be easiest to automatically pull from an agreement proves to be the most challenging. For example, we tried to build a model to find the end date in an agreement. As a starting issue, some tools do not have the ability to even find discrete data points (e.g., dates, numbers, the state for governing law), they can only be trained to perform full text extractions. Even with a system that can pull data points, few agreements have hard end date specified. Instead, lawyers draft provisions that the agreement ends "two years from the Effective Date," "two years from the date of the last signature" or "two years from the date of the first order." In none of these cases will "AI" add "two years" to the Effective Date to determine an End Date value.

Tools Will Get Even Better The tools are useful but not yet complete. For example, a common challenge is dealing with amendments. As noted by Tim Cummins at IACCM, "Because contracts are rarely seen as practical business tools, their administration is typically poor. Those responsible for overseeing and delivering performance frequently haven't read the contract and don't know what it says. No surprise, then, when change and amendment processes are generally reactive and often adversarial."

Some tools allow a user to group related documents together (e.g., a master agreement and its amendments). This is helpful so far as the AI models extract relevant text from all of the related documents and present the text side-by-side for a human reviewer. However, no tools even attempt to compare the language from the different documents against each other to determine what the amended agreement says.

"Before you can abstract data from contracts, you need to collate contracts, remove duplicates, identify and define needed fields for the review, group related contracts. To be effective software needs to address this kind of 'contract organization,' support managing the overall workflow and, of course, the automated text abstraction."

▶ **Murali Tirupati, Co-Founder & CEO of Vaultedge**

Overall, many tools seem to have been built to save time spent by higher priced lawyers, operating individually or in a small team. Compared to other leading legal technologies like document review platform and contract management systems, abstraction tools offer a minimum of configurability, especially when it comes to fairly standard workflow features like batching and permission levels. We would like to see vendors focus more on building out these workflow features and other functions that drive real savings for organizations, like alternative legal service providers, operating at larger scales and with much larger teams. As one example, no tool tested offered the ability to assign specific fields within an agreement to a reviewer and give that reviewer permission to only see and edit those fields. This makes it challenging to set up a review process where a lower level resource is responsible for fields that are easier to review and more senior resources review more complicated/nuanced provisions. We anticipate tools will continue to build out functionality covering a more robust workflow or integrating with solutions already in the market that support workflows.

CONCLUSION

"We are excited with the success these technologies are having and the wide adoption across the industry. The leading vendors are rolling out new functionality to cover even more use cases and deliver even better results. Machine learning researchers are working on the next generation of this technology."

▶ **Noah Waisberg, Co-Founder & CEO at Kira Systems**

Based on the results of our case study, we integrated AI-based abstraction tools into several of our workflows. QuisLex is an industry leading alternative legal service provider with contract abstraction as a core competency. We have a legion of trained contract professionals dedicated to quality contract abstraction inside a meticulously designed and validated Six Sigma workflow. This is what we do. Our proof of concept was therefore based on real-world projects for world-class clients. We can think of no better vote of confidence in where these tools are, and where they are going, than altering our time-tested workflows at risk to our wallet and our reputation. We are convinced.

Contract analysis technology will continue to advance dramatically, and we look forward to seeing new functionality and enhancements so that we can find new ways to work these tools into our processes to drive even more value for our clients.

Appendix 1 – Bringing Structure to Inherently Unstructured Contracts

It is labor intensive to identify all contracts with pending expiration dates if all we have is a file folder containing a large number of scanned contracts. These files are "unstructured"(i.e., we would need to open and read each contract to determine their contents).

We structure data by populating fields in a database with material information. In a simple example, a contract would be linked to a database containing a field for effective date. That is, we abstract the effective date from the individual contract and populate the database field. This abstraction can be manual (a human reads the contract and types in the date) or automated (a machine identifies the effective date and makes the entry in the database). Once populated, the field enables us to quickly sort and filter contracts by effective date because all the information is now in a single location.

But effective date is only one among a myriad of potentially material data points. These data points can range from binary questions, like whether the contract contains a most favored nation provision, to verbatim extraction of text – e.g., if the contract does have a most favored nation provision, the entire provision is copied into the database where it can be searched, filtered, and analyzed without needing to open and review the linked contract. Other critical data points involve interpretation and analysis – e.g., summarizing extracted text or reviewing extracted text against standard language to determine the level of conformance or deviation.

In short, moving from unstructured to structured contract data can be quite useful – but also extremely labor intensive depending on contract volumes and the kind of questions that the database needs to answer.

Appendix 2 – Framework for "Accuracy" Analysis

Below is the step by step approach taken to grading results:

Step 1: Manually prepared key for the testing set of documents (i.e., what is "correct")

Step 2: Flag in key whether a particular provision/concept in an agreement should have been Blank (i.e., there was no text/language in the agreement) or Not Blank (i.e., there was some text/language in the agreement)

Step 3: For results for each tool (i.e., what the tool automatically pulled), assign one of the following to each provision/concept in each agreement:

All – All of the Correct Information and No additional Information

None – No information (i.e., its blank)

Extra – All of the Correct Information and Additional Information (note: this applies whether the additional information is completely wrong or somewhat related)

Some – Some of the Correct Information but Missing Other Correct Information (note: this applies regardless of whether it also has additional information)

Wrong – None of the correct information (note: this applies where box was not blank but none of the correct information was found)

Pass – Either (i) it was too close to being a judgment call (e.g., the text was relevant but not exactly matching the key)[1] or (ii) the tool did not have an out of the box model for a given provisions[2]

Step 4: Generated results:

1. Excluded "Pass" from both sides of the equation (i.e., was not "right" and was not "wrong" and removed from denominator of total provisions scored)
2. Where there was no provision to extract (should have been blank):
   a. If result was "None" it was scored "Right"
   b. If result was anything else, it was scored "Wrong"
3. Where there was a provision to extract:
   a. If result was "None" or "Wrong" it was scored "Wrong (did not extract any correct text)"
   b. If result was "All," "Extra" or "Some" it was scored accordingly (i.e., one to one)

[1] Adding "Pass" resulted in scoring of Tools increasing across the board, but it was consistently applied across Tools. We took this approach as a control on the process to make sure that all Tools were scored fairly and to avoid subjective decisions effecting scoring results.

[2] Adding "Pass" resulted in scoring of Tools increasing across the board, but it was consistently applied across Tools. If a tool did not have an "Out of the Box" model or was unable to train a bespoke model with the data set provided, while certainly important as part of a holistic evaluation, we did not want to penalize its accuracy score.

Appendix 3 – Details on More Traditional "Accuracy" Analysis[1]

To come up with the results listed in this Appendix 3, the same process described in Appendix 2 was followed with the following changes:

1. Where there was no provision to extract (should have been blank):
   a. If result was "None" it was deemed a "True Negative"
   b. If result was anything else, it was deemed a "False Positive"
2. Where there was a provision to extract:
   a. If result was "None," "Wrong" or "Some"[2] it was deemed a "False Negative"
   b. If result was "All" or "Extra" it was deemed a "True Positive"[3]
3. Using these designations, the following were generated based on the following formulas:
   a. Recall = True Positives / (False Negatives + True Positives)
   b. Precision = True Positives / (False Positives + True Positives)
   c. F1 = 2 * ((Precision*Recall) / (Precision + Recall))[4]
   d. "Accuracy" = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)[5]

Accuracy of Out-of-the-Box Models

	Untrained Human	Tool 1	Tool 2	Tool 3	Tool 4	Tool 5
Recall	65%	79%	60%	69%	68%	58%
Precision	99%	84%	90%	99%	93%	98%
F1	78%	81%	72%	81%	79%	73%
"Accuracy"	79%	77%	69%	80%	77%	71%

Bespoke/Custom Model Testing

	Tool 2	Tool 3	Tool 4	Tool 5
Recall	35%	75%	96%	36%
Precision	77%	97%	100%	47%
F1	48%	85%	98%	41%
"Accuracy"	59%	83%	97%	35%

[1] Special thanks to the folks at Kira for helping provide details on this framework, which may be closer to the way people in machine learning circles expect to see results. Special thanks to Aditya Mohanty, Head Quality and Business Excellence at QuisLex for validating this framework.

[2] Our approach would result in scoring of both of the following as "Some": (i) if there was only one correct paragraph/clause in a contract to abstract, abstracting only a portion of that paragraph/clause instead of the whole paragraph/clause and (ii) if there were more than one correct paragraph/clause in an contract to abstract, abstracting less than all of the paragraphs/clauses. However, we did not make this distinction in our analysis. If this distinction was made, we understand that when there is more than one correct paragraph/clause in a contract to abstract, another approach would have been to score each one separately. For example if the model abstracted two of three correct paragraphs/clauses in a contract it would have a score of 66%, as opposed to 0%. Or if the model abstracted three of three correct paragraphs/clauses in a contract it would receive three separate "True Positives" to add to the overall score. It is unclear if this alternative approach would have raised or lowered scores, but we do not think this distinction had a material impact on our analysis.

[3] The thinking of deeming both "All" and "Extra" as "True Positives" may be that, despite best efforts, some subjectivity goes into classifying a result and we were cautioned that the results may be dependent on things like which models were chosen, who did the training, and not as much a reflection of the tools capabilities.

[4] F1 may be thought of as an overall measure of accuracy, and conveys the balance between precision and recall.

[5] "Accuracy" here is an attempt to place emphasis on blanks being returned as blanks, though some may question the usefulness of this score as it could return a high "accuracy" when there are a million blanks correctly returned but the model fails to find the needle in a haystack, which would not be good.

QuisLex
200 Liberty Street
New York, NY 10281

+1.917.512.4489
info@quislex.com

www.quislex.com

About QuisLex

QuisLex is an award-winning legal services provider that specializes in managed document review, contract management, compliance services, legal spend management, and legal operations consulting. Our full-time highly trained attorneys, process experts, legal technologists, statisticians and linguists work closely with our clients to reduce cost, mitigate risk and maximize efficiency. QuisLex is regularly acknowledged as a leader in the legal services industry, and is proud to be recognized by the Financial Times as an FT Intelligent Business 35, ACC as an ACC Value Champion, Chambers and Partners as a Band 1 Legal Process Outsourcing Provider, New York Law Journal as a Top Managed Document Review Services Provider, and the IACCM as its Outstanding Service Provider for contract management solutions. To learn more, visit www.quislex.com.