eDISCOVERY AI

LARGE LANGUAGE MODEL AI FOR DOCUMENT REVIEW

Contents

eDISCOVERY AI

ABSTRACT

This white paper explores the revolutionary
impact of eDiscovery AI's next-generation
predictive coding tool, which leverages
large language models in AI for eDiscovery.
Traditional predictive coding tools have
demonstrated minor improvements in efficiency
and accuracy since their adoption more than a
decade ago, but there have been few, if any,
significant enhancements over that time.

The advent of large language models has
brought unprecedented advancements to
the analysis, classification, and creation
of legal documents, and we're using
these advancements to solve many of the
longstanding issues with traditional predictive
coding and document review.

This paper examines the capabilities, benefits,
challenges, and best practices associated with
implementing next-generation AI in predictive
coding and document review, providing
valuable insights for legal professionals seeking
to harness the power of these cutting-edge
technologies.
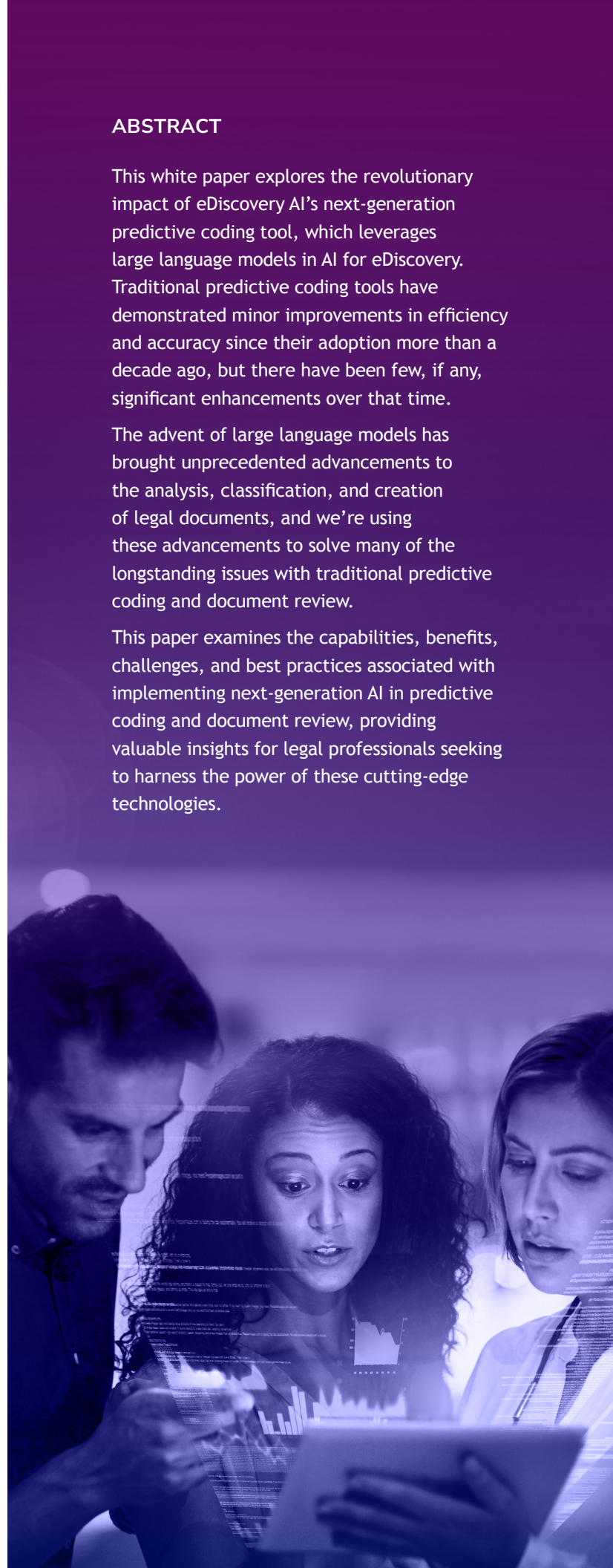
01 INTRODUCTION

BACKGROUND ON THE EVOLUTION OF PREDICTIVE CODING IN eDISCOVERY

In the early 2010s, the first generation of predictive coding was beginning to gain traction in the eDiscovery industry, and court approval of the use of the technology in the Da Silva Moore v. Publicis Groupe helped propel this forward. Predictive coding, also referred to as Technology Assisted Review (TAR) was split into two different workflows, TAR 1.0 and TAR 2.0.

TAR 1.0 involves training a machine learning algorithm on a subset of documents that have been manually reviewed and categorized by human experts. The algorithm then applies what it has learned to automatically classify the remaining documents based on their relevance to the case. As acceptance grew, best practices were adopted and minor improvements to the existing tools were made, but the technology remained relatively static.

In TAR 2.0, the process is enhanced by leveraging continuous active learning. It starts similarly to TAR 1.0, with a small set of manually reviewed documents used for training. However, in TAR 2.0, the algorithm interacts with human reviewers in a feedback loop. The algorithm selects batches of documents for the reviewers to manually review and categorize, focusing on the documents that are most likely to be relevant. The reviewer's feedback is then incorporated into the algorithm, iteratively improving its performance and prioritizing the most likely relevant documents at the front of the review queue. This process continues until the rate of relevant documents reviewed in the most recent batches drops so low that further review is no longer justified.

Despite the ever-growing acceptance of predictive coding and the development of standardized workflows and best practices, there has been little substantive change to the underlying technology over the last decade. The effectiveness of the tools remained flat, and the burden of manual review of large training document sets is a constant requirement regardless of which tool is used.

OVERVIEW OF eDISCOVERY AI'S TECHNOLOGY UTILIZING LARGE LANGUAGE MODELS

The recent introduction of large language models and their rapid improvement have created a much more powerful tool for the analysis, classification, and summary of data than anything previously used in eDiscovery. eDiscovery AI has harnessed the incredible power of these new tools and created a program to allow the use of these new tools for eDiscovery document review.

PURPOSE AND STRUCTURE OF THE WHITE PAPER

This paper will introduce you to eDiscovery AI's new predictive coding tool by explaining key concepts, benefits, challenges, and best practices. With this information, you'll be empowered to use eDiscovery AI's technology to dramatically improve the effectiveness of your document review with minimal manual effort.

02   NEXT-GENERATION PREDICTIVE CODING: KEY CONCEPTS

EXPLANATION OF NEXT-GENERATION PREDICTIVE CODING AND ITS ADVANCEMENTS

The only way to train traditional predictive coding tools has always been through documents that were manually reviewed and provided to the machine as both positive and negative examples of each type of document you were looking for. The typical process required thousands of training examples before the system could begin to reliably classify documents. The next-generation technology used by eDiscovery AI is already trained using vast amounts of data from countless sources and only requires detailed, plain-language directions to identify, classify, and summarize what you're looking for.

INTRODUCTION TO LARGE LANGUAGE MODELS AND THEIR ROLE IN DOCUMENT REVIEW

Large language models utilize advanced natural language processing and contextual understanding to analyze and comprehend vast amounts of data. By learning from incredible volumes of data, eDiscovery AI can make classifications based on the responsiveness of documents, the application of various issues, and the identification of key and privileged documents. Leveraging eDiscovery AI's tools to classify all of this information can nearly eliminate the time and cost associated with manual review while improving the overall accuracy compared to traditional predictive coding or human review.

03   BENEFITS OF NEXT-GENERATION PREDICTIVE CODING IN eDISCOVERY

UNPRECEDENTED ACCURACY: LEVERAGING ADVANCED LANGUAGE UNDERSTANDING CAPABILITIES

Traditional predictive coding tools tend to maximize recall and precision somewhere around 70% after thorough and repeated training efforts. The power of eDiscovery AI allows previously unreachable recall and precision levels at or above 90% to become commonplace without any of the costly and time-consuming training effort.

ENHANCED COMPREHENSION OF COMPLEX LEGAL CONCEPTS AND JARGON

By starting with a model that's already trained, eDiscovery AI can actually find that needle in the haystack or smoking gun that has typically been nearly impossible to locate. Traditional predictive coding required hundreds of positive training examples for each category of document sought. In practice, there often aren't that many truly key documents, let alone an effective method to find them in order to feed them into the predictive coding tool so that it can find more. eDiscovery AI allows you to write a plain-language description of the key content that you're looking for and, without using a single example document, locate that critical email.

AUTOMATICALLY CREATE EXPLANATIONS FOR CLASSIFICATION DECISIONS

While eDiscovery AI is analyzing your documents, it can also create a summary of each document, so at a quick glance you can understand what is being discussed without having to read the entire document. Classification explanations are also incredibly valuable to help you understand why a document is considered relevant or key without having to dig for that crucial bit of information buried in paragraphs of otherwise irrelevant content.

ACTUAL EARLY CASE ASSESSMENT WITH SUMMARIES OF ALL KEY OR POTENTIALLY HARMFUL DOCUMENTS

Classify your documents right away and generate document summaries and explanations without having to wait for round after round of training or review with a team of contract attorneys. Being able to summarize the strengths and weaknesses of your case and the data you've collected can prove invaluable for settlement negotiations and overall case strategy.

DEFINING SPECIFIC GOALS AND PARAMETERS FOR YOUR INSTRUCTIONS TO EDISCOVERY AI

eDiscovery AI can be incredibly effective at finding what you ask for, but it's only as good as the instructions that you provide. Just like instructing a human document review team, the instructions to eDiscovery AI need to be as clear, detailed, and specific as possible. It can also be very effective to include details about certain types of documents that are not relevant but may be similar to what is actually relevant.

DEVELOP CAREFULLY DESIGNED INSTRUCTIONS TO EDISCOVERY AI ON EXACTLY WHAT YOU'RE LOOKING FOR AND ANYTHING THAT YOU'RE SPECIFICALLY NOT INTERESTED IN SEEING

In order to ensure that your instructions to eDiscovery AI convey the accurate and specific meaning of each relevance and issue category, it's best to test and refine your instructions before applying them to your entire data set. Start by crafting your first set of instructions for eDiscovery AI and having it analyze a small set of data, ideally a few relevant and a few irrelevant documents. Run this small set through eDiscovery AI, evaluate the results, update your instructions based on the initial results, and repeat until you're comfortable with the effectiveness of your instructions.

ESTABLISHING A HYBRID WORKFLOW COMBINING HUMAN EXPERTISE AND MODEL PREDICTIONS

Often, the early shortcomings of an instruction to eDiscovery AI come from details or aspects of an issue that were previously unknown. By testing your instructions and reviewing the results, you can clarify your instructions while also learning more about your case and issues that are helpful even beyond this phase of discovery.

DOCUMENTS SUMMARIES AND CATEGORY EXPLANATIONS HELP FAST TRACK THE PROCESS OF REFINING EDISCOVERY INSTRUCTIONS AND YOUR EARLY CASE ASSESSMENT

While testing your instructions with eDiscovery AI, it is extremely useful to turn on category explanations. These features will provide you with short, concise explanations of what the document is and why it was considered relevant or not. This makes it quick and easy to understand why a document was missed or why it was erroneously included and improve your instructions.

COLLABORATING WITH LEGAL PROFESSIONALS AND TECHNICAL EXPERTS FOR OPTIMAL RESULTS

For years, lawyers have been drafting memos to instruct document review teams on how to code documents, and that process is very similar with eDiscovery AI. Despite that similarity, there may be small differences needed to communicate with this new audience, and we have experts standing by to assist in crafting the perfect instructions so you can fully maximize the potential of eDiscovery AI.

05 CASE STUDIES AND SUCCESS STORIES
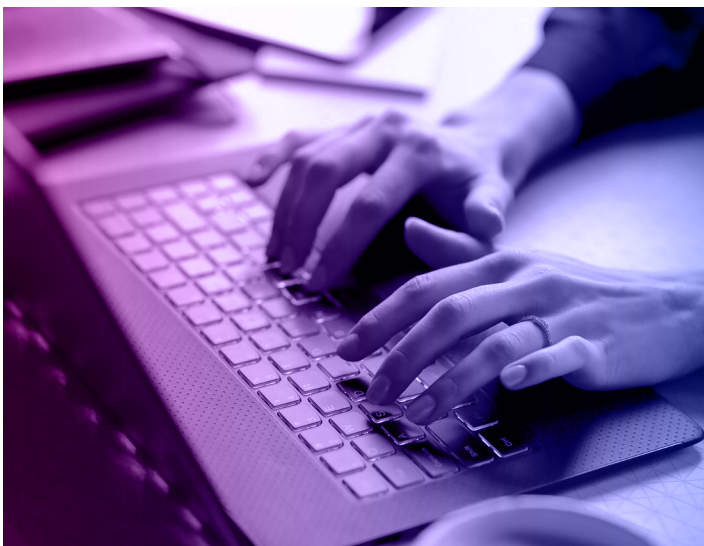
REAL-WORLD EXAMPLES OF NEXT-GENERATION PREDICTIVE CODING IMPLEMENTATIONS

A law firm used eDiscovery AI to replace manual document review on a collection of 180,000 documents related to litigation involving various business practices of their client. The firm designed initial instructions for eDiscovery AI to describe what contents make a document relevant and not relevant. The attorneys then tested the instructions to determine their effectiveness and areas for improvement. Through iterative refinement, the instructions are modified based on the analysis of missed documents, improving their effectiveness. Continuous evaluation and optimization were performed, utilizing the explanations provided by the AI to fine-tune instructions. After four iterations, finalized instructions were used to classify the entire document set.

DEMONSTRATING SIGNIFICANT IMPROVEMENTS IN ACCURACY AND EFFICIENCY

After running the entire document set through eDiscovery AI, the law firm reviewed a random sample of nearly 2,400 documents to create a control set. Recall and precision metrics were created by comparing AI categorizations with those of subject matter experts in this control set, generating a recall of 99.01% and a precision of 96.57%. This case study demonstrates the successful application of eDiscovery AI in document review, offering improved accuracy and efficiency for legal professionals.

HIGHLIGHTING THE IMPACT ON LEGAL OUTCOMES AND CLIENT SATISFACTION

By eliminating the time, cost, and human error involved in traditional document review, eDiscovery AI gives you significant strategic advantages. You'll now be able to understand the strengths and weaknesses of your case much earlier in the process. With the speed of eDiscovery AI, you can fully devote your time to settlement negotiations or arguing the merits of the case rather than spending months managing eDiscovery reviews.

06 CONCLUSION

RECAP OF THE BENEFITS AND CHALLENGES OF NEXT-GENERATION PREDICTIVE CODING

eDiscovery AI is powered by next-generation, large language model AI that can generate more accurate results than previous predictive coding tools without the burden of thousands of training documents. In order to achieve those results, it's essential to create specific and accurate instructions that are typically the result of time spent testing and refining.

ENCOURAGEMENT FOR LEGAL PROFESSIONALS TO EMBRACE AND EXPLORE THESE TRANSFORMATIVE TECHNOLOGIES

Large language models like those that power eDiscovery AI are new and powerful tools, but they are also incredibly easy to use. No coding or special formatting is required; simply provide plain-language instructions and analyze the results eDiscovery AI can bring you.

GLOSSARY OF KEY TERMS

Traditional Predictive Coding – also known as technology assisted review (TAR) describes the use of machine learning algorithms to assist with the identification and classification of documents in eDiscovery matters. Traditional predictive coding requires subject matter experts to manually review and classify documents from their dataset in order to train the predictive coding model to effectively classify the remaining, unreviewed documents.

TAR 1.0 – A predictive coding workflow that involves training a machine learning algorithm on a subset of documents that have been manually reviewed and categorized by human experts. The algorithm then applies what it has learned to automatically classify the remaining documents based on their relevance to the case.

TAR 2.0 – A predictive coding workflow that leverages continuous active learning. Unlike TAR 1.0, the algorithm interacts with human reviewers in a feedback loop. The algorithm selects batches of documents for the reviewers to manually review and categorize, focusing on the documents that are most likely to be relevant. The reviewer's feedback is then incorporated into the algorithm, iteratively improving its performance and prioritizing the most likely relevant documents to the front of the review queue. This process continues until the rate of relevant documents reviewed in the latest batches drops so low further review is no longer justified.

Large language model - Advanced artificial intelligence systems that are designed to process and understand human language on a vast scale. These models leverage deep learning techniques to learn from massive amounts of textual data and generate contextually relevant and coherent responses. Large language models often have millions or even billions of parameters, enabling them to capture intricate patterns in language and produce highly sophisticated outputs. These models have the potential to assist users in understanding and generating human-like text.

Instructions – specific guidance, or prompts written to the AI in order to describe the type of document that the user is looking for and any type of summary or explanation that the AI should create. Instructions must be written with attention to detail and should typically be tested and refined before running across a large data set.

Control set – a random sample of documents that are both manually reviewed, and AI reviewed in order to measure the effectiveness of the review and generate effectiveness metrics by comparing manual and AI classifications.

Recall – a statistical measure that is used to calculate the proportion of relevant documents that are retrieved from the use of predictive coding, often described as a measure of completeness. High recall means that a predictive coding system is able to accurately find most of the relevant documents.

Precision – a statistical measure that is used to calculate the proportion of documents classified as relevant that are truly relevant. If a predictive coding model has high precision, it means that that when the model suggests that a document is relevant, most often it is truly relevant.

Da Silva Moore v. Publicis Groupe – A landmark court case where Judge Peck's decision established a precedent for the use of predictive coding in eDiscovery that recognized the efficacy of the technology. This decision led to an increase in acceptance of predictive coding across eDiscovery matters. (Da Silva Moore v. Publicis Groupe et al, No. 1:2011cv01279 - Document 96 (S.D.N.Y. 2012))

eDISCOVERY AI

NOTE: This white paper provides a starting point for understanding the potential of eDiscovery AIs next-generation predictive coding tools using large language models. As the field and our tools rapidly evolve, it is essential to stay updated on the latest research, advancements, and ethical considerations to ensure the effective and responsible implementation of these technologies. Stay up to date at eDiscoveryAI.com and follow along with us on social media.