LITIG: AI Benchmarking Initiative – Workshop (1 August 2024)

Key Takeaways and Insights

Introduction

As AI systems, especially those using generative AI, continue to make waves in the legal industry, evaluating their performance has become a growing challenge. To bring clarity to this, **LITIG**, with support from **Artificial Lawyer** and **Legal IT Insider**, has kicked off an **AI Benchmarking Initiative**. This effort aims to create an open source benchmark for legal AI systems that everyone can get behind – building transparency, trust, and responsible use across the board.

We launched with a bang, holding an inaugural workshop at CMS's London office in August 2024. The event brought together over 40 organisations, including legal tech vendors, law firms, and other stakeholders, all keen to share their insights and help shape the future of legal AI. The goal? Gather feedback, define initial focus areas, and chart a path forward for AI benchmarking in the legal space.

Workshop Highlights

The workshop was designed to be hands-on, with participants breaking into groups to dive deep into AI benchmarking challenges.

Richard Tromans from Artificial Lawyer kicked off the workshop with a presentation highlighting why we need a benchmark and then John Craske a member of the Litig Board and also Chief Innovation & Knowledge Officer at CMS set the scene and facilitated the workshop sessions.

Two group activities formed the core of the evening's discussions:

Activity 1: Benchmarking Options:

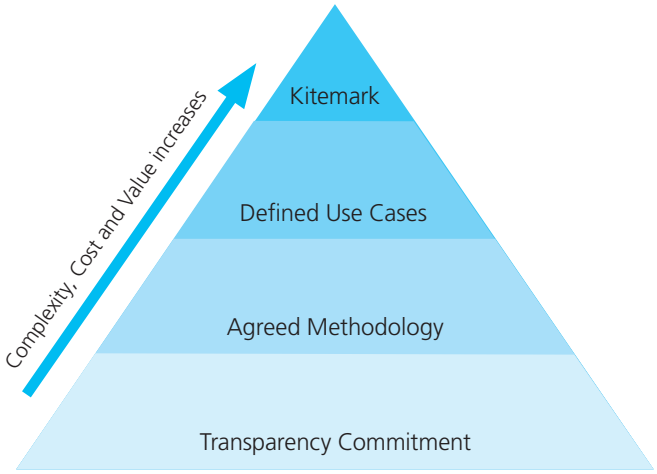
We explored four potential ways to benchmark AI tools, each one each one potentially more valuable but also more complex to build than the last. The groups were asked to consider the pros and cons of each of the options and suggest any options that might be missing.

Activity 2: Drafting a Consultation Paper

Teams put themselves in the shoes of the wider legal community, brainstorming what a consultation paper might look like and identifying key elements to include.

Benchmarking Options: Simplifying the Complex

The heart of the first activity was reviewing four proposed benchmarking models, each with its own level of complexity. Here's a snapshot of what was discussed:



Level 1: Transparency Commitment:

This base option calls for AI vendors who quote any accuracy (or other) metrics to be transparent about the methodology, data and scenarios used for any testing so that anyone with access to the tool can repeat the test. The consensus? Transparency is a must. It builds trust and gives users the ability to verify claims, ultimately helping them choose the right tools.

Level 2: Agreed Methodology:

This approach would mean agreeing a common methodology to testing (but not necessarily consistent use cases or data) – making metrics easier to understand and trust. If we are all testing tools in the same way it also makes it easier to compare. Agreeing a methodology would be more effort than a simple transparency commitment, but would take less time and effort to build (and maintain) than a more robust set of defined use cases.

This sparked lively debate. Some saw the value in a universal testing standard for AI tools, making comparison easier. Others argued that the complexity of legal tasks means one size doesn't fit all. Flexibility might be key here.

Level 3: Defined Use Cases:

This approach builds on the agreed methodology and adds in agreed use cases – scenarios where AI tools might support accompanied by example questions and criteria for valid answers.

Everyone agreed use cases are important. They help users understand where AI tools are most effective. But the challenge lies in defining these use cases clearly and ensuring they are backed by solid data.

Level 4: Kitemark:

A trusted "seal of approval" like a Kitemark or ISO certification from an independent organisation was viewed as attractive but potentially expensive and time-consuming. It could reassure users about the quality of AI tools, but might also stifle innovation if the bar is set too high. It might also take too long to agree to be useful in today's fast moving world.

Common Themes: What Everyone Agreed On

While opinions varied, there were some clear points of agreement across the groups:

— Clarity is King: We need consistent definitions of key terms like accuracy, hallucinations, and usefulness.

— Balance is Key: It's important to find a middle ground between speed, cost, accuracy, and trust (of any benchmark as well as any AI tool).

— Data Matters: Legal AI systems require high-quality, representative data – not an easy feat in an industry where confidentiality is paramount.

— One Size Doesn't Fit All: Legal tasks and AI tools are diverse, and any benchmark must be flexible enough to account for that.

— Shared Responsibility: Vendors, law firms, clients, academics and regulators all play a role in shaping the future of AI benchmarking.

Drafting the Future

The second activity was all about looking ahead. Groups were tasked with imagining the next phase – a consultation across the legal industry to gather feedback and refine the benchmark. Here's a sneak peek at what that consultation paper might include from the work done in the workshop:

— Introduction: Set the stage by explaining the need for a benchmark, the benefits and risks of legal AI, and why transparency and trust are vital.

— Objectives: Establish clear goals – such as defining common terms, identifying key use cases, and proposing practical solutions.

— Stakeholders: Acknowledge the different groups involved, from vendors and law firms to regulators and consumers.

— Engagement: Map out how we'll gather input from the legal community, using surveys, interviews, and workshops to ensure everyone has a voice.

— Governance: Define who will oversee the process, ensure quality, and keep things moving forward.

— Key Questions: Dive into the big questions – what's working, what's not, and how can we make AI systems more transparent, reliable, and trusted?

Conclusion & Next Steps

Our inaugural event highlighted the many challenges (and exciting opportunities!) in building an AI benchmark system for legal AI. The discussions were lively and insightful, with participants eager to push this initiative forward.

What's next? We'll be hosting a virtual follow-up event to keep the conversation going, especially for those who couldn't attend in person. From there, a core working group will take the lead in drafting a consultation paper, with support from broader industry groups.

If you're passionate about shaping the future of legal AI, we want to hear from you! Stay connected by signing up for updates and joining our upcoming events. Together, we can create a transparent, trustworthy, and accountable AI ecosystem for the legal industry.