

What does “good” look like?

At our Munich dinner with German Legal Tech Hub, lawyers discussed how to judge legal AI in practice: what benchmarks reveal, how to test tools on real matters and where professional judgement remains essential.

The Zeno Team • 26 October 2026 • 9 min read



IN BRIEF

- Legal AI quality depends on legal judgement, verification and usefulness, not accuracy alone.
- Benchmarks are credible only when their tasks, success criteria, sources and assessment methods are clear.
- Firms should evaluate the complete product on their own use cases, not rely on a foundation model score.
- Adoption works best when firms test recurring assignments, document corrections and support colleagues with practical examples.

At a recent dinner in Munich jointly organised by Zeno and German Legal Tech Hub, Matthijs Malessy, Benedikt Quarch and Daniella Domokos discussed how lawyers can assess the quality of legal AI. The conversation covered the role of benchmarking, how to test Legal AI tools and the difficulties firms encounter when putting these AI tools to work. A central question connects these issues: How do you set a credible standard for Legal AI? This article draws on the evening's discussions to examine how firms can establish that standard in practice.

Accuracy is only part of legal quality

An AI answer may state the law correctly yet miss the issue that decides the matter. It may also give a confident conclusion when the available facts do not justify one. A credible standard needs to account for these weaknesses.

The DELTA Benchmark uses the term “legal taste” to describe such a professional judgement in an AI answer and assesses substance, citation and form. It identifies the decisive issues, gives them appropriate weight and handles uncertainty carefully. The concept of legal taste helps explain why a fluent, factually accurate answer may still need substantial revision before a lawyer can use it.

[Read the full Delta Benchmark report.](#)

The elements that comprise legal taste do not exist in isolation: they come together to form a holistic view of the utility and quality of AI in legal practice. For firms, these standards should be defined before a firm evaluates a tool in order to make the right selection.

A research memorandum, for example, may need to identify the governing authorities, apply them to the supplied facts and address unresolved questions. Reviewers can then assess whether the answer fulfils the assignment, rather than relying on its fluency or apparent thoroughness.

The distinction matters because an omission can be more consequential than an incorrect statement. An answer that accurately describes the general rule but misses an applicable exception may leave the lawyer with a misleading account of the legal position.

Verification belongs in the evaluation

Establishing that a citation exists is an essential check, but it does not establish that the source supports the conclusion. A real judgment can be cited, alongside a claim it does not support. An answer may quote a passage but leave out a condition or exception that changes its meaning. An answer can also overlook a source that changes the analysis. These failures require different checks, as examined in Zeno's article, [The hallucination trap](#).

To combat this, firms should therefore assess how effectively a tool used for legal work enables verification. A lawyer needs to move from a claim to the relevant passage, read it in context and assess its application to the facts. The effort required deserves attention alongside the quality of the initial answer.

If verification involves repeating most of the research, the practical benefit may be limited. An evaluation should capture the corrections required, the questions left unresolved and the work needed before the output can be used. This is where Legal AI providers play a significant role: technology providers can build solutions that make verification simple. Clients often request this, as Matthijs Malessy mentioned:

“We focus on working together with our user, lawyers operate in Zeno day to day. We noticed the invisible burden of verification: spending the time saved by using AI on verifying the AI output instead. Therefore, we made this an integral part of the platform: ensuring legal sources and documents are always cited, down to the exact passage. Flagging when a claim is only partially supported by an underlying source, and allowing users to double check. Having Zeno ask a question to the user if essential details are missing, rather than assuming the right path. If a source is not available in the database, directly communicating this, rather than filling the gaps with fluff.”

In the conversations following the panel, German law firm leaders explicitly expressed their preference: They do not need the perfect-sounding answer; rather, they want honesty. Knowing the capabilities and limitations of a tool encourages them to keep experimenting, with the confidence that they can not only verify, but do so without spending excessive additional time.

Benchmark results need context

A benchmark compares performance on a defined set of tasks under specified conditions. Its relevance depends on how closely those tasks reflect the work a firm intends to undertake. For this reason, it is crucial to evaluate an AI tool on a firm's own use cases. Four questions help establish what a result means:

What was tested? The jurisdiction, practice area and type of assignment determine the relevance of the findings.

What counted as success? A vague score can conceal a failure on a decisive part of an answer.

What did the system have access to? Sources, tools and testing conditions affect output.

How were the answers assessed? Clear criteria, practitioner involvement and stated limitations allow readers to judge the strength of the findings.

Model performance and product performance are different

A foundation model's performance is one component of a legal AI product's quality. The application also determines which sources are available, how research is conducted and how the resulting work can be inspected and revised.

DELTA compares foundation models within a standardised legal research environment. A legal AI product adds its own sources, research tools and review features around the model. Firms therefore need to evaluate the product they will actually use; the underlying model's benchmark score cannot establish how well the complete product performs.

[See the benchmark report.](#)

Matthijs Malessy explained that his position is that model selection should be optimised based on evidence. This is centred on benchmarking outputs and evaluating which models perform the best on which specific legal task. An architecture that supports different providers allows those choices to be revisited as tools develop and use cases change.

Product evaluation must also examine how capabilities work together. In a contract review, retrieving the relevant law is only one step. The system must apply it to the document while following the supplied instructions and producing an output the lawyer can review. Testing isolated capabilities will not reveal every weakness in that process.

Adoption requires practical support

A firm can invest in legal AI and still struggle to establish where it is useful. Internal advocates may understand its potential but lack the time or materials to help colleagues apply it to their work.

Worked examples provide a practical starting point. They show the assignment, the information supplied and the output produced, including the corrections a lawyer needed to make. Colleagues can assess the process and decide whether it fits similar work of their own.

A recurring assignment is a useful basis for an initial trial. A small group can agree what a satisfactory result requires, test the workflow and document where further work is needed. Those findings can guide training and wider adoption.

A credible standard becomes useful when a firm can apply it consistently to its own work. A trial should establish whether the tool meets the agreed criteria, which errors recur and how much review remains necessary. Those findings give the firm a practical basis for deciding where to use the tool and what support colleagues will need.

Test Zeno for free

Explore how Zeno supports legal research, document review and drafting in your practice.

Test Zeno for free ↗

[Read this article online](#)