Benchmarking the Invecorum Agent against frontier AI agents and local LLM setups

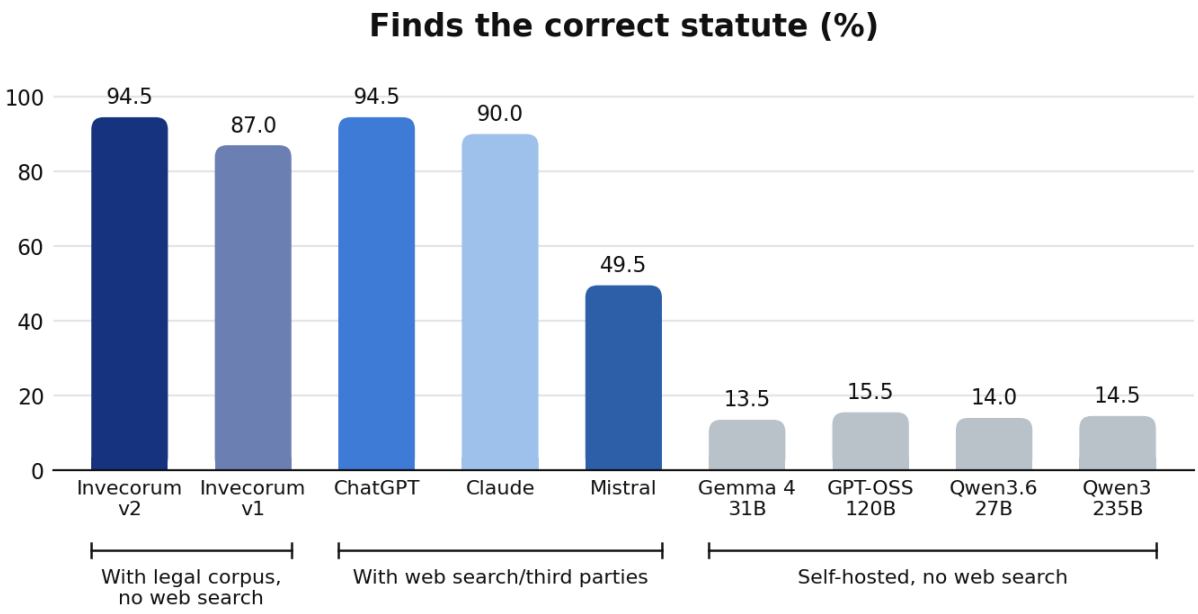
As of 1 August 2026

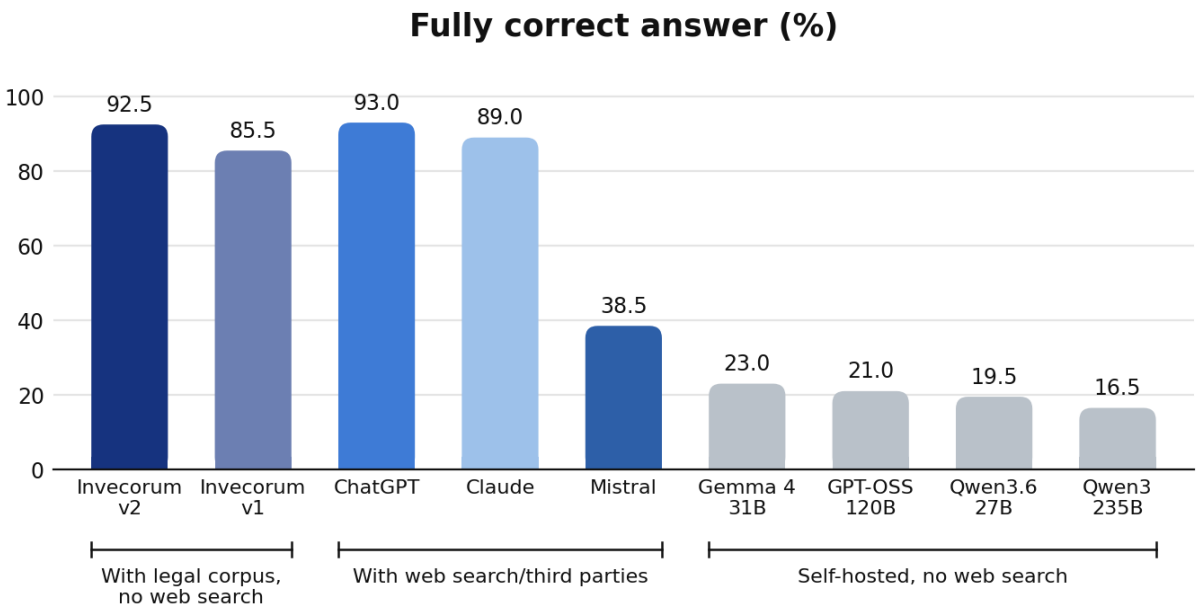
The tax-tech market is growing by the day. The large AI models deliver high quality, but for German tax practices they run into professional secrecy constraints. Until now, tax advisors therefore often had to choose between quality and compliance.

The Invecorum Agent delivers both. In tax-law fact-pattern research we compared it with seven other setups, including ChatGPT (GPT-5.5) and Claude (Opus 4.8), each with web search. The result: the Invecorum Agent is on par with ChatGPT. All other setups rank behind. And it runs entirely on rented GPUs in German data centres.

Reliable answers with the exact citation

The basis is a benchmark of 200 fact-pattern questions drawn from ongoing advisory practice (methodology at the end of this document). We tested nine setups: the Invecorum Agent in the current version v2 and the previous version v1, three AI agents with web search (ChatGPT, Claude, Mistral Le Chat) and four self-hosted open-source models. We measure whether a system finds the correct statute and whether it answers the question fully correctly:





The Invecorum Agent v2 matches ChatGPT on finding the correct statute. On fully correct answers, half a point separates the two. Compared with the previous version v1, v2 gained seven points there. Claude ranks behind on both markers. ChatGPT reaches its scores with on average more than six web searches per question and the longest response time of the four strongest systems. In addition, every case is routed through the provider's search infrastructure. The Invecorum Agent works without web search and without external search services.

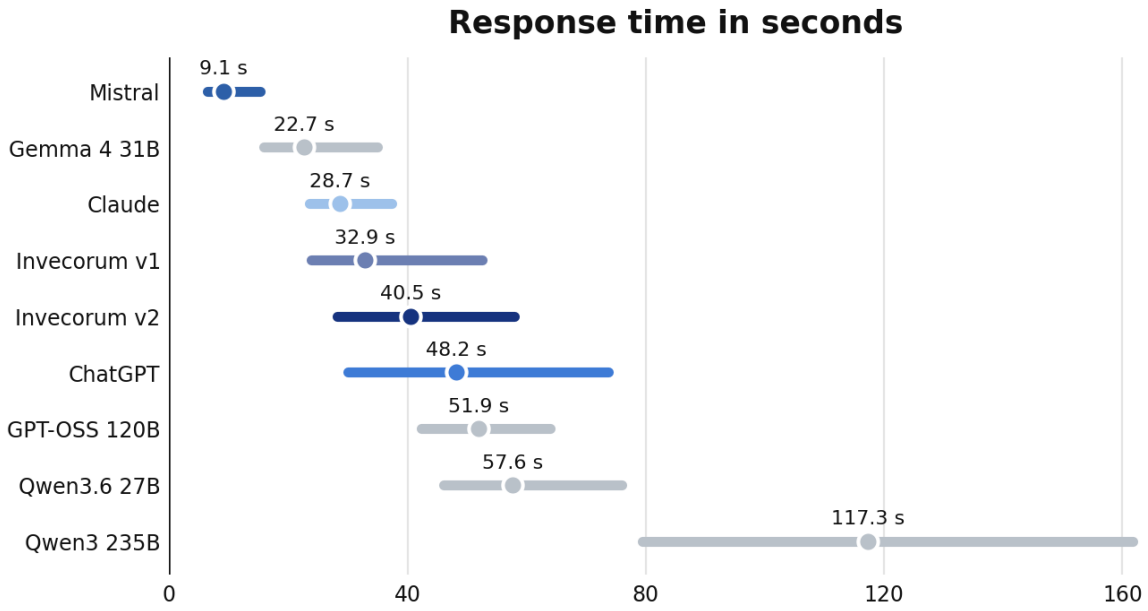
Mistral answers only 38.5 % of questions correctly despite using web search. The self-hosted models range from 16.5 % to 23.0 %. This narrow band holds across all model sizes from 27B to 235B parameters. Simply self-hosting a standard model is not enough for this task. What matters is access to the statutes in the correct version. Access alone is not enough, though: even with web search, results range from 38.5 % to 93.0 % depending on the model driving it. Source and model have to work as one system.

Equally important in practice: in none of its failure cases did the Invecorum Agent invent a citation. Even there, every cited statute had actually been opened in the legal corpus first. Errors stay traceable.

Response times compared

Mistral answers fastest overall. Among the strongest systems, Claude is the fastest and ChatGPT, because of its many web searches, the slowest. The two Invecorum versions sit in between. Version v2 is slightly slower than v1 but remains clearly faster than ChatGPT. Three of the four self-hosted models were slower than every agent in our test. The largest took almost two minutes at the median.

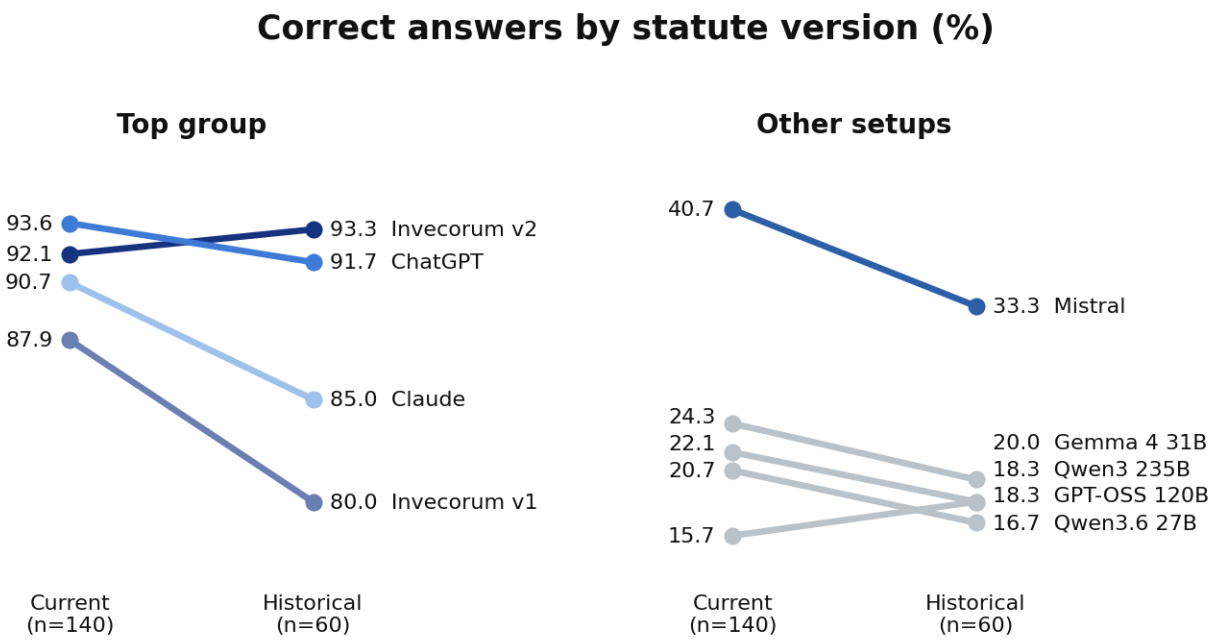
The dot shows the typical response time (median), the bar the range containing half of all answers:



Historical statute versions

Questions about earlier assessment periods are the most demanding part of the benchmark. Almost all systems lose accuracy there. Answers then easily rely on today's legal position instead of the version valid at the relevant date. This type of error is particularly problematic because the answers sound convincing.

The Invecorum Agent looks up statutes in the legal corpus as of the relevant date. Version v2 keeps its accuracy stable on legacy cases. At 93.3 % it reaches the best value of all setups tested there. Historical search was a focus of the development from v1 to v2:



Usage limits compared

ChatGPT Plus (23 euros per month in Germany) caps GPT-5.5 at around 160 messages per 3 hours. After that, it automatically switches to a smaller model. Claude Pro (around 20 euros per month) works with 5-hour windows plus additional weekly limits.

Limits remain even in the business tiers: ChatGPT Business (from 29 euros per user) caps the Thinking variant of GPT-5.5 at 3,000 messages per week. Unlimited usage is only available in ChatGPT Enterprise, starting at 150 users. With Claude, higher limits are available in the Team Premium tier for roughly 110 euros per user. Enterprise is priced individually (all provider information as of July 2026).

The Invecorum Agent costs 69 euros per month on the solo licence and has no usage limits. It runs on rented GPUs from German providers. The software stack on top is built in-house. Cost per query is therefore low and fully predictable. That is why the subscription includes unlimited usage with no throttling.

Consistent quality

The limits lead to a second difference: control. ChatGPT automatically switches to a smaller model at the usage limit. Claude does the same at certain limits. Which model produces an answer is decided by the provider in that moment. In day-to-day work this is often not visible.

The Invecorum Agent always runs in the same configuration. Every request is answered by the same system at any time of day.

Client data stays in Germany

With ChatGPT and Claude, the facts of the case, i.e. client data, are processed with every web search through the provider's systems. Third-party search providers are partly involved as well. Processing can then also take place outside the EU. Even with EU data residency booked, OpenAI explicitly excludes web search from it. Retrofitted research databases for these environments do not change this either: the facts of the case are still processed by the AI provider itself. The Invecorum Agent runs exclusively in German data centres whose operators are bound to confidentiality. Case facts never leave this environment. This matters for professional secrecy obligations (sec. 203 German Criminal Code) and for the requirements for service providers under sec. 62a of the German Tax Advisory Act (StBerG).

Bottom line

In its current version v2, the Invecorum Agent is on par with ChatGPT and ahead of all other setups. On historical statute versions it reaches the best value of all systems tested. Of the offerings compared, it is the only one with unlimited usage at consistent quality. Client data stays in confidentiality-bound German data centres throughout. Self-hosted standard models without a legal corpus are not an alternative: on the same benchmark they reach at most 23.0 % correct answers.

Methodology

The benchmark consists of 200 fact-pattern questions drawn from ongoing advisory practice. Each question is answered by exactly one time-versioned statute. 140 questions concern the current version, 60 a historical one. Such legacy cases come up daily in assessments, tax audits and appeals.

Scoring is strict: the citation only counts for the exact statute in the correct version. The answer only counts if it contains every key point of the answer key. Scoring was done by the same, unchanged judge model for all systems. All systems ran in their default configuration.

Systems tested: the Invecorum Chat Agent in the current version v2 and the previous version v1. References in the text refer to v2 unless stated otherwise. ChatGPT with GPT-5.5 and Claude with Opus 4.8, each with web search enabled and unrestricted. Mistral Le Chat with Mistral Medium 3.5 and web search. In addition, four self-hosted open-source models without web search and without a legal corpus: Gemma 4 31B, GPT-OSS 120B, Qwen3.6 27B and Qwen3 235B Thinking. The response times of these four models depend on the hardware used and should be read as indicative. The Invecorum v2 and Claude runs were scored against a corrected version of the answer key. This adjusted 4 of 200 cases. The effect on comparability is at most 2 points.

An example from the benchmark:

Question: “A client asks for a tax assessment of his situation as of 6 July 2026. He carries out a sustained activity that regularly generates income. However, he argues that he is not an entrepreneur because he has no intention of making a profit. How is his activity to be classified legally?”

Target citation: sec. 2 German VAT Act (UStG), version effective 1 January 2023

Expected answer: He is an entrepreneur; the activity qualifies as commercial or professional; the absence of a profit motive is irrelevant.