[klara]case

BENCHMARK · EDITION 2026-11

# How often does an AI model cite a case number that is not in our collection?

We asked four AI models which case law applies: ten questions from six areas of law. Each question ran three times: with no tools, with web search and with klaracase. We then compared every case number they cited with our collection.

| 4 | 6 | 3 | 902 |
|---|---|---|---|
| AI models | Areas of law | Setups per question | Cited case numbers |

| | |
|---|---|
| EDITION | 2026-11 |
| DATE | Published on 3 November 2026 |
| QUESTION SET | Question set 1 |
| MODELS | Models tested, measured on 3 November 2026: GPT-6.1 Terra · GPT-6 Astra · Claude Opus 5.5 · Claude Sonnet 5.5 |

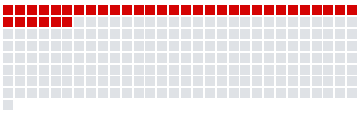
01 · RESULT

## How often does an AI model cite a case number that is not in our collection?

No tools

**36** of 241

cited case numbers are not in our collection



Share: 15% (36 of 241)
Wrong date: 5 of 241

With web search

**27** of 306

cited case numbers are not in our collection

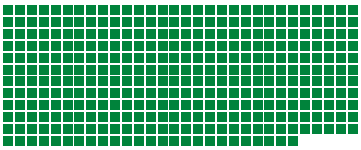


Share: 9% (27 of 306)
Wrong date: 7 of 306

With klaracase

**355** of 355

cited case numbers came from a tool result (traceable)



Share: 100% (355 of 355)
Not in our collection: 0 of 355
Wrong date: 1 of 355

02 · AREAS OF LAW

## Results by area of law

Per area of law and setup: case numbers that are not in our collection, out of all cited.

| AREA OF LAW | No tools | With web search | With klaracase |
|---|---|---|---|
| Criminal law | **5 of 16** | **8 of 85** | 0 of 92 |
| Business crime | **6 of 59** | **1 of 47** | 0 of 59 |
| Tenancy law | **4 of 77** | 0 of 68 | 0 of 85 |
| Employment law | **8 of 50** | **9 of 53** | 0 of 64 |
| Tax law | **5 of 28** | **1 of 32** | 0 of 32 |
| Social security law | **8 of 11** | **8 of 21** | 0 of 23 |

# Results by model

All six areas of law combined, per model and setup. The names are the exact model versions tested, measured on 3 November 2026.

| SETUP | CITED | CONFIRMED | WRONG DATE | WRONG COURT | NOT IN OUR COLLECTION | FOUND ELSEWHERE | OPEN | TRACEABLE |
|---|---|---|---|---|---|---|---|---|
| **GPT-6.1 Terra** | | | | | | | | |
| ■ No tools | 23 | 21 | 0 | 0 | **2** | 1 | 1 | – |
| ◆ With web search | 39 | 36 | 2 | 0 | **1** | 1 | 0 | – |
| ● With klaracase | 60 | 60 | 0 | 0 | 0 | 0 | 0 | 60 |
| **GPT-6 Astra** | | | | | | | | |
| ■ No tools | 46 | 43 | 1 | 0 | **2** | 0 | 2 | – |
| ◆ With web search | 62 | 61 | 0 | 0 | **1** | 0 | 1 | – |
| ● With klaracase | 65 | 64 | 1 | 0 | 0 | 0 | 0 | 65 |
| **Claude Opus 5.5** | | | | | | | | |
| ■ No tools | 69 | 59 | 0 | 0 | **10** | 7 | 3 | – |
| ◆ With web search | 97 | 78 | 2 | 0 | **17** | 5 | 12 | – |
| ● With klaracase | 123 | 123 | 0 | 0 | 0 | 0 | 0 | 123 |
| **Claude Sonnet 5.5** | | | | | | | | |
| ■ No tools | 103 | 77 | 4 | 0 | **22** | 10 | 12 | – |
| ◆ With web search | 108 | 97 | 3 | 0 | **8** | 4 | 4 | – |
| ● With klaracase | 107 | 107 | 0 | 0 | 0 | 0 | 0 | 107 |

The columns “Confirmed”, “Wrong date”, “Wrong court” and “Not in our collection” add up to the number under “Cited”. “Confirmed” means: same court and same date as in our collection. “Found elsewhere” and “Open” split the column “Not in our collection”. “Traceable” counts the cited case numbers that came from a tool result.

# How we measured

### ■ No tools

1 Question
2 AI model
3 Answer with case numbers

The model answers the question on its own. No search, no tools, no internet.

COMPARED WITH
Our collection of court decisions on the day of the measurement

### ◆ With web search

1 Question
**2 AI model with web search**
3 Answer with case numbers

The model may search the internet before it answers.

COMPARED WITH
Our collection of court decisions on the day of the measurement

### ● With klaracase

1 Question
**2 AI model with access to our database**
3 Answer with case numbers

The model reads our database through tools. We count how many cited case numbers came from a tool result. We call that traceable.

COMPARED WITH
Our collection on the day of the measurement. Also: did the case number come from a tool result?

## What counts as one?

One case number cited in one answer.

## How large is the sample?

Four models, ten questions in six areas of law, each question in three setups. The models cited 241 case numbers with no tools, 306 with web search and 355 with klaracase. The totals differ, because each model decides for itself how many decisions it cites. That is why every figure sits beside its own total.

## How do we check a case number?

We count every cited case number and compare it with our collection as it stood on the day of the measurement. It is confirmed (same court, same date), has a wrong date, has a wrong court, or is not in our collection.

## How often did we measure each question?

Once per model and setup. We take no average and claim no repeat runs.

TRAP QUESTIONS

**20** of 24

Two more questions name sections that do not exist. They are not part of the ten questions. They drew 24 answers. An automatic check flagged 20 of them as saying the section does not exist. A person still has to review the other 4.

## Why the questions stay unpublished

We do not publish the wording of the questions. Published questions end up in the training data of the models. A model then already knows the questions, and the test loses its value.

# What this benchmark does not show

### What “not in our collection” means

It only means: we did not find the case number in our collection. The decision can still exist. We looked these case numbers up on dejure.org. Some exist there (“found elsewhere”). The rest is still open.

| SETUP | NOT IN OUR COLLECTION | FOUND ELSEWHERE | OPEN |
|---|---|---|---|
| No tools | 36 | 18 | 18 |
| With web search | 27 | 10 | 17 |
| With klaracase | 0 | 0 | 0 |

“Open” means: a person has not checked these case numbers yet. We claim nothing about them.

01 We did not measure whether an answer is legally right. Nobody reviewed the legal content of the answers. We only count cited case numbers.

02 klaracase built the tool and graded the run itself. We are the vendor and the checker at once.

03 We count a wrong date for each case number, from the text of the answer. An automatic reader assigns the dates. It can be wrong.

04 The figures apply to the four model versions tested on the day of the measurement. There were ten questions in six areas of law. The figures say nothing about later versions or other questions.

05 All figures come from one recorded measurement on 3 November 2026. They do not change as our collection grows.

06 · EDITIONS

## All editions

A new edition appears about every three months. The counts are comparable only between editions with the same question set.

| EDITION | DATE | QUESTION SET | MODELS | ■ No tools NOT IN OUR COLLECTION | ◆ With web search NOT IN OUR COLLECTION | ● With klaracase TRACEABLE |
|---|---|---|---|---|---|---|
| 2026-11 | 3 November 2026 | 1 | 4 | 36 of 241 | 27 of 306 | 355 of 355 |

07 · CONTACT

## Talk to us

### Research and press

Questions about the method or the data behind this benchmark? Write to us. We share the measurement data with researchers on request.

pr@klaracase.de

### Law firms and companies

Want to know whether klaracase fits your work? Write to us. We will arrange a call.

info@klaracase.de

### This edition online

klaracase.de/en/benchmark/edition/2026-11

**Try it yourself**

Create a free account and try klaracase today.

klaracase.de/register