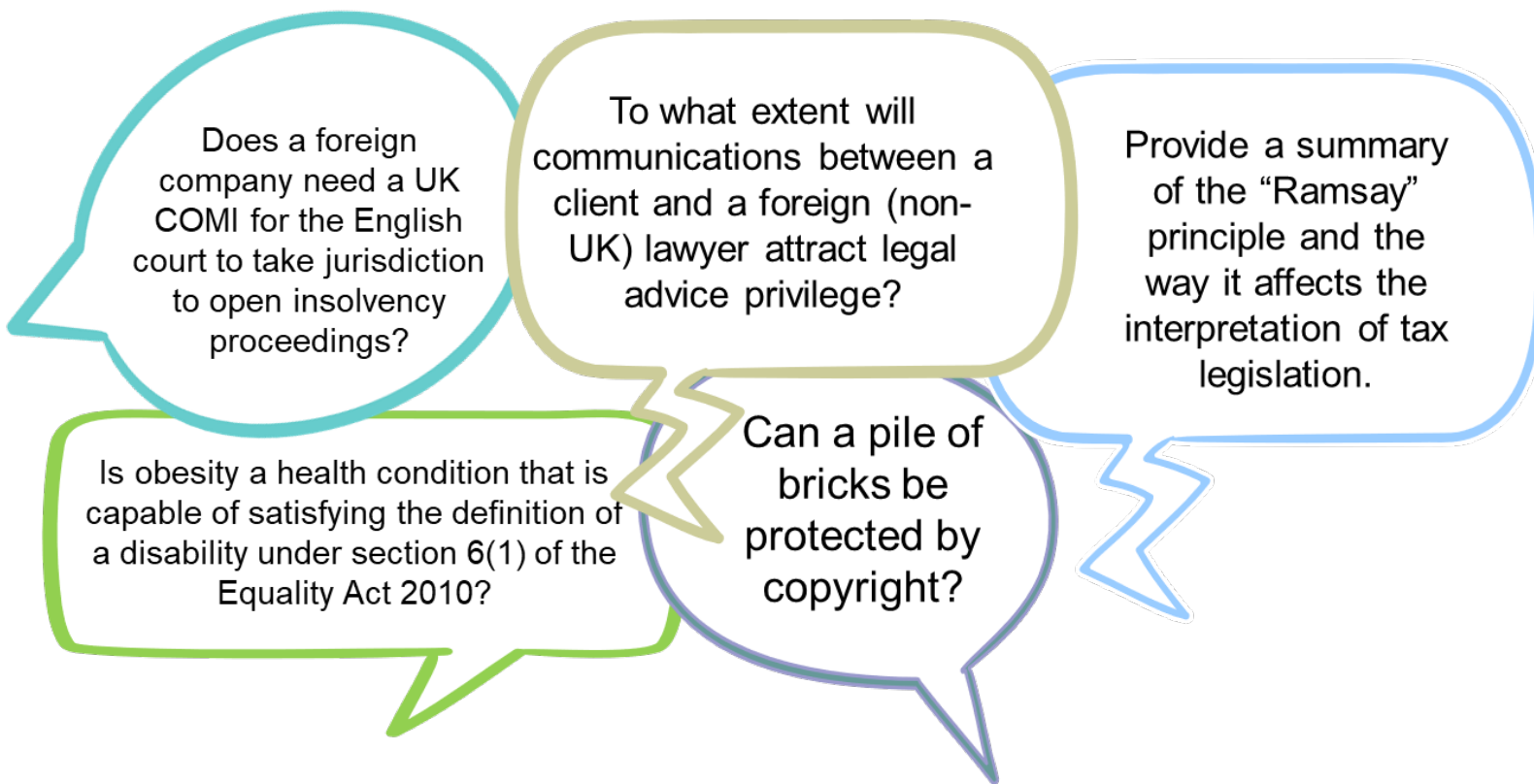


## Linklaters



# The LinksAI English law benchmark

Version 2 – Updated to include OpenAI o1 and Google Gemini 2.0

February 2025

# Summary

## The LinksAI English law benchmark tests how much we can rely on AI tools to provide English law advice.

We have updated our benchmark to test two of the latest LLMs, OpenAI o1 and Google Gemini 2.0.

The results show significant improvement. These LLMs are starting to perform at a level where they could assist in legal research, e.g. by providing a first draft or as a cross-check on an existing answer.

However, the answers are still sometimes wrong. There are dangers in using them if you don't already have a good idea of the answer.

LLMs also have an important role supporting the provision of legal services (e.g. summarisation, extraction, etc.).

### What questions did we use?

The benchmark comprises 50 questions from 10 different practice areas. The questions are relatively hard. They are the sort of questions that would require advice from a competent mid-level lawyer specialised in that practice area. The intention was to test if the AI models can reasonably replicate some tasks carried out by a human lawyer.

### Which LLMs were tested?

We tested each of those 50 questions against six different models including Bard, multiple versions of GPT, o1 and Gemini 2.0. None of these models were specially trained to provide legal advice and, instead, they are just general-purpose LLMs. Other LLMs may perform better.

### How were the answers marked?

The answers were marked by senior lawyers from each practice area. Each answer was given a mark out of 10 comprising 5 marks for substance (*is the answer right?*), 3 for citations (*is the answer supported by relevant statute, case law, regulations?*) and 2 for clarity.

### How did the models perform?

In the first benchmarking exercise (October 2023) we found that the answers from GPT 2 were entirely incorrect and were given a mark of 0 across the board. Interestingly, there was little difference in the marks between GPT 3 and GPT 4 (scoring around 3.3 out of 10). Bard scored the best at the time (4.4 out of 10). The answers from those models were often wrong and the citations fictional.

In the second benchmarking exercise (February 2025), we saw significant improvements. We gave:

- > **Gemini 2.0** a score of **6.0 out of 10**
- > **OpenAI o1** the top score of **6.4 out of 10**

In both cases, this was driven by material increases in the scores for substance and the accuracy of citations.

## Summary (cont.)

The LinksAI English law benchmark tests how much we can rely on AI tools to provide English law advice.

### *“When the facts change...”*

The update provides an opportunity to reassess the usefulness of these tools to deliver legal advice.

The LLMs we tested in the first benchmarking exercise in October 2023 gave answers that were often wrong, and citations that were often fictitious. This led us to recommend that LLMs should not be used for legal advice without expert supervision (or at all).

Despite the significant improvement in Gemini 2.0 and OpenAI o1, we still recommend they should not be used for English law legal advice without expert human supervision. They are still not always right and lack nuance.

However, if that expert supervision is available, they are getting to the stage where they could be useful, for example by creating a first draft or as a cross-check. This is particularly the case for tasks that involve summarising relatively well-known areas of law (see page 11). In contrast, their ability to apply the law to the facts or interpret clauses is less good.

### Methodology

When we published the first version of this report, strong views were expressed that the methodology used was unfair and better results could be obtained by, for example, more sophisticated prompt engineering.

We have reflected on that feedback. While we are happy with our methodology, this report explores the use of alternative prompts (see pages 15 – 17) explains the reasons for our approach in more detail (see pages 19 – 20).

More sophisticated prompts and/or interaction with the LLMs would generate better answers, but the benchmark needs a standardised repeatable process. The methodology also reflects our views on likely real-world use, particularly by less sophisticated users.

### Wider applications in legal services

We also acknowledge that LLMs can perform a number of other important functions in the delivery of legal services (e.g. summarisation, contract extraction and contract drafting).

This benchmark does not address these “skills”, though they are addressed in other studies, see page 21.

### Conclusion

Our conclusion remains broadly the same. The latest models we tested may be useful but only when used with expert human supervision. There are real risks to using them if you don’t already know the answer.

Whether this conclusion holds for the next update to the benchmark remains to be seen.

# Contents

## Main body

5

The background

6

The questions

7

The mark scheme

8 – 18

The answers

19 – 21

Methodological observations

22

The future

23

Our bench of experts

## Annexes (available [here](#))

1 – 2

Annex A – The marks

3 – 8

Annex B – Bonus “clause” questions for Gemini/o1

9 – 80

Annex C – The questions and answers

81 – 84

Annex D – Sample of answers from GPT 2

85 – 86

Annex E – Varying answers from GPT 4

““ ... Generally, at the risk of humanising a robot, it felt like Gemini 2.0 “tried too hard” - in many cases the right answers were there but alongside a lot of extra / duplicative material. I was impressed with o1.

One of our bench of experts



# The background

The LinksAI English law benchmark has been created to test the capabilities of LLMs to deliver English law legal advice.

The capabilities of LLMs continue to improve and could have profound implications for the future provision of legal services.

However, should you be getting your legal advice from an LLM?

## The need for a benchmark

While the answers produced by LLMs are – at least superficially – convincing, they are not always correct and often lack nuance and context.

We have therefore developed a detailed set of benchmark questions to test the capabilities of LLMs to deliver English law legal advice.

The intention of the benchmark is to deliver both an *absolute* assessment of performance, and to allow a *relative* assessment of different LLMs and their progression over time.

## Supercharging human lawyers

Importantly, the benchmark only addresses the provision of English law legal advice.

There are many other potential use cases for lawyers. For example:

- > **Summarising** longer documents, such as to create a bullet point summary.
- > **Contract extraction** of specific provisions from agreements.
- > **Research and enhanced search** to find relevant cases or laws.
- > **Stylistic amendment** to make a document something more concise, less formal, etc.
- > **Ideation** to help come up with concepts and ideas.

This use of LLMs to accelerate the delivery of legal services is already happening and could significantly affect this sector.

However, these use cases are **not** considered in this report.

## Iterations of the benchmark

We ran the initial benchmarking questions through the GPT 2/3/4 and Bard in September 2023.

We ran our benchmarking questions through OpenAI o1 and Gemini 2.0 in December 2024.

We intend to rerun this benchmarking exercise in future years as new LLMs and other AI tools are released onto the market.

If the improvement in LLMs continues at its current rate, future editions of this benchmark could be interesting.

# The questions

## The benchmark questions are relatively hard.

They are intended to be the sort of questions that might reasonably be asked of a competent mid-level associate (2 years post qualification (PQE)) who – while unlikely to immediately know the answer to them – would be able to produce a competent response after some research.

### Which practice areas?

There were 50 questions in total spanning 10 different practice areas of English law: contract law, intellectual property, data privacy, employment, real estate, dispute resolution, corporate, competition, tax and banking.

A full set of questions and answers is set out in Annex C.

### Who created the questions?

The questions were created by our expert bench of senior lawyers from each practice area (see below).

### The prompt rubric

Each question was preceded by the following standardised prompt rubric. (The earlier benchmarking used a similar, but not, identical rubric).

*“You are an experienced English lawyer with a reputation for correctly explaining difficult legal concepts in clear, easily understood language. Provide a concise answer to the question below applying ENGLISH law. Cite any relevant statutes, regulations, guidance or case law that are authoritative in England.”*

We also experimented with alternative prompt rubrics to see if this improved performance (see pages 15 – 17 & 19).

### What type of questions are used?

The questions range from simple summarisation questions:

*“Is there a difference between confidential information and trade secrets?” (Q7)*

To questions that apply the law to a set of facts:

*“The customer and supplier are party to a 5 year office cleaning contract covering one office in Croydon. The supplier invoices the customer on a weekly basis. The customer says it will not pay 50% of the supplier’s invoice for one week’s cleaning because the cleaning was carried out to a poor standard. The customer has no right under the contract to do this. Is the customer in repudiatory breach of contract?” (Q4)*

To questions that analyse a clause:

*“An agreement contains the following gross up clause for withholding tax: “All sums payable under this Agreement shall be paid free and clear of all deductions, withholdings, set-offs or counterclaims whatsoever save only as may be required by law. If any deductions or withholdings are required by law, the payer shall account to the relevant Tax Authority for the amount so required to be deducted or withheld and shall be obliged to pay to the recipient such additional amounts as will ensure that the recipient receives, in total, an amount which (after such deduction or withholding has been made) is no more and no less than it would have been entitled to receive in the absence of any such requirement to make a deduction or withholding”. An interest payment of £100 is subject to UK withholding tax at a rate of 20%. How much does the payer have to pay to the recipient of the payment and how much does it have to pay to the Tax Authority?” (Q15)*

# The mark scheme

We marked the answers based on substance, citations and clarity.

The questions were marked individually by our expert bench of senior lawyers.

## Substance (5 marks)

We awarded a maximum of 5 marks for the substance of the answer – i.e. whether the answer was technically correct.

## Citations (3 marks)

We awarded a maximum of 3 marks for correct references to cases, laws or guidance. The use of a fictitious citation automatically leads to 0 marks being awarded.

## Clarity (2 marks)

We awarded a maximum of 2 marks for the clarity of the answer.

A full set of questions and answers is set out in Annex C, including a short commentary on the answers.

Substance	
0	The response is entirely <b>wrong</b> .
1	The response is generally <b>wrong</b> but contains some correct analysis.
2	The response is generally <b>accurate</b> but contains a number of errors.
3	The response is generally <b>accurate</b> but contains a small number of errors or fails to answer parts of the question.
4	The response is generally <b>accurate</b> and covers most issues.
5	The response is <b>accurate</b> and covers all material issues.
Citations	
0	The citations are <b>fictional</b> .
1	The citations are <b>incorrect</b> .
2	The citations are generally <b>accurate</b> but there are important omissions.
3	Adequate and <b>accurate</b> citations are used.
Clarity	
0	The response is very <b>difficult</b> to read.
1	The response is <b>clear</b> but not easy to read.
2	The response is clear and easy to read.

## The answers

The LLMs we tested performed below the level of a qualified lawyer.

GPT 3/4 and Bard provided answers that were convincing but were frequently wrong and contained fictitious citations.

Gemini 2.0 and o1 performed better particularly on questions involving summarising the law.

### Overall scores

The table on the right summarises the overall performance of GPT 2, GPT 3, GPT 4, Bard, Gemini 2.0 and OpenAI o1.

Annex A contains a detailed table of scores.

Annex C contains the individual scoring for each answer alongside a short commentary of those answers.

We provide a further breakdown of the performance of these LLMs overleaf.

	Substance	Citations	Clarity	Total
GPT 2	0 / 5	0 / 3	0 / 2	0.0 / 10
GPT 3	1.1 / 5	0.8 / 3	1.4 / 2	3.3 / 10
GPT 4	1.0 / 5	0.8 / 3	1.4 / 2	3.2 / 10
Bard	1.8 / 5	0.9 / 3	1.7 / 2	4.4 / 10
*NEW Gemini 2.0	2.7 / 5	1.6 / 3	1.6 / 2	6.0 / 10
*NEW OpenAI o1	2.8 / 5	1.8 / 3	1.8 / 2	6.4 / 10

## The answers – Comparison of LLMs

The performance of the models show steady improvement over time.

The benchmarking exercise provides an opportunity to test the relative performance of different AI models.

### GPT 2 was hopeless

We ran a small selection of questions through the GPT 2 (davinci-002) model hosted in OpenAI's playground. That model was released in 2019. The answers are set out in Annex C. They are uniformly nonsense and were given a presumptive mark 0.

### GPT 3 and 4 were very similar

We also used GPT 3 (text-davinci-003) and GPT 4, both of which are hosted in our internal environment. The version of GPT 3 we used was released in November 2022 and GPT 4 in March 2023. We found very little difference between the two models. In fact, GPT 3 scored (3.3 out of 10) outperforming GPT 4 very slightly (3.2 out of 10). In other words, the new version showed little progress.

### Bard performed better

We ran these questions through the publicly

available version of Google's Bard. It is based on the LaMDA family of LLMs and was released in March 2023. Bard scored better when we tested in September 2023 (4.4 out of 10). The substance of its answers was generally better than GPT 3/4 (1.8 out of 5). However, it was just as prone to incorrect or fictitious citations (0.9 out of 3).

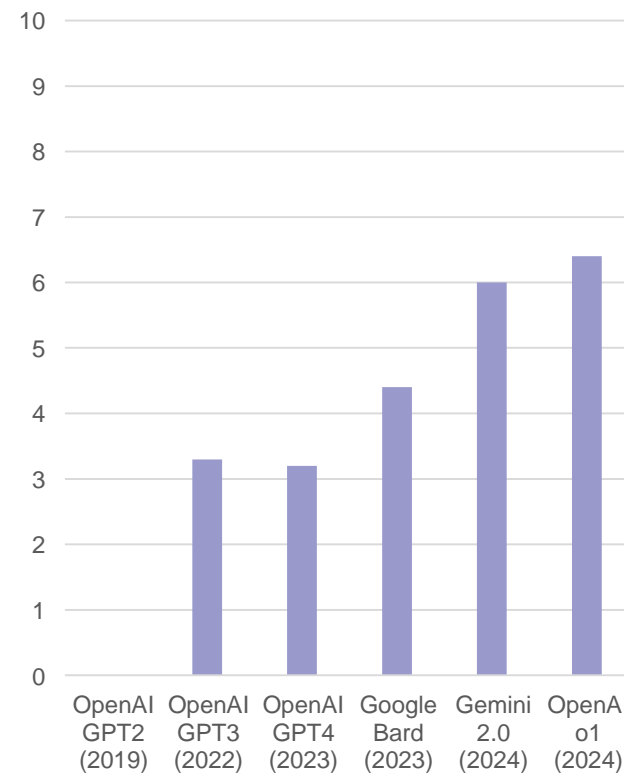
### Gemini 2.0 and OpenAI o1 have improved significantly

In our most recent update, we tested Gemini 2.0, which was released in December 2024 and OpenAI o1, which was also released in December 2024. In both cases, we used the publicly available versions.

Gemini scored 6.0 out of 10. The substance of its answers was generally good (2.7 out of 5) and it was less prone to incorrect or fictitious citations (1.6 out of 3).

The best performer was OpenAI o1. It scored 6.4 out of 10. The substance of its answers was generally good (2.8 out of 5) and it also was less prone to incorrect or fictitious citations (1.8 out of 3).

Performance Comparison  
(Out of 10)



## The answers – Substance (1)

Early models – The substance of the answers from GPT 3/4 and Bard was generally inaccurate.

The output of GPT and Bard was inaccurate and requires expert human review.

### Is it technically correct?

In a small number of cases, we received good answers. GPT 3 even achieved full marks on one of the competition questions (5 out of 5) (Q45).

However, in many cases its English law analysis was very poor with almost a third of the answers achieving 0 marks (see table right). There were a range of reasons for this, including:

- > **Just getting the law wrong.** For example, GPT 3 wrongly states that a pile of bricks cannot be subject to copyright, entirely missing the point that it can be protectable as a “sculpture” (a subset of “artistic works”) (Q6).
- > **Making up the law.** When asked about criminal offences under the Data Protection Act 2018, all three LLMs invented a series of plausible – but entirely imaginary – offences (Q16).

- > **Missing the point of the question,** such as the privilege question (Q39) which asks if communication between a client and non-UK lawyer is privileged. GPT 3 and 4 answer as if the non-UK lawyer is simply a non-lawyer.
- > **Focusing on the wrong law.** When answering whether a foreign company needs a UK COMI for the English court to take jurisdiction to open insolvency proceedings, GPT 3 focuses on the EU Insolvency Regulation, which no longer regulates UK insolvency jurisdiction (Q48).

### Too confident or too eager

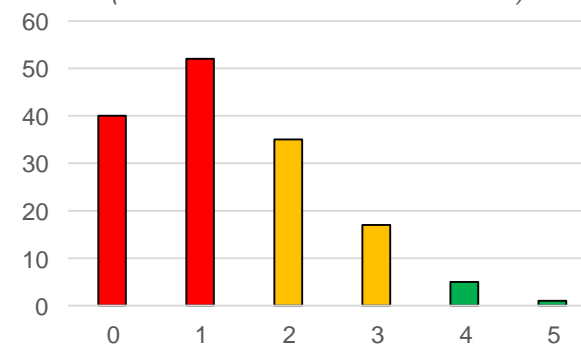
One potential problem is that the “eagerness” of the models to provide a clear answer leads to them overstating the confidence of their advice.

For example, when asked if a resolution can be withdrawn once the company has given notice of a meeting, Bard concludes that once the meeting has begun, unanimous consent is needed. This is too definitive. Views vary on the majority required (Q28).

A more extreme example comes from a question about a clause which gives third-party rights to group companies only. When asked what the problem with the clause was, GPT 4 came to the absurd conclusion that any attempt to restrict the contractual third parties to group companies only is “ineffective and unenforceable”.

The correct answer is that the clause doesn’t allow the parties to vary or terminate the contract without the consent of the third-party beneficiaries. However, rather than simply admit it doesn’t know the answer, GPT 4 manufactures an incorrect argument.

Score distribution for substance  
(Scores out of 5 – GPT 3/4 & Bard)



## The answers – Substance (2)

The new models – The substance of the answers from Gemini 2.0 and OpenAI o1 was noticeably better.

Both performed well when asked to summarise the law but less well where critical reasoning was needed.

Looking at the overall performance, these models scored 2.7 out of 5 on substance. Looking at the spread of answers (see diagram):

- > In 26% of cases, the answer is mostly or fully accurate and complete, i.e. comparable with the answer we would expect from a competent 2-year PQE who researched the issue.
- > In 17% of cases, the answer is entirely or almost entirely wrong.

However, the performance varies depending on the type of question and both are too “eager” to provide a single clear answer, which leads to them overstating the confidence of their advice and failing to provide a range of potential outcomes.

Gemini 2.0 also arguably “tried too hard”. In many cases, it provided the right answers but did so alongside a lot of extra/duplicative/incorrect material. In contrast, o1 was generally short and to the point.

### Summarisation

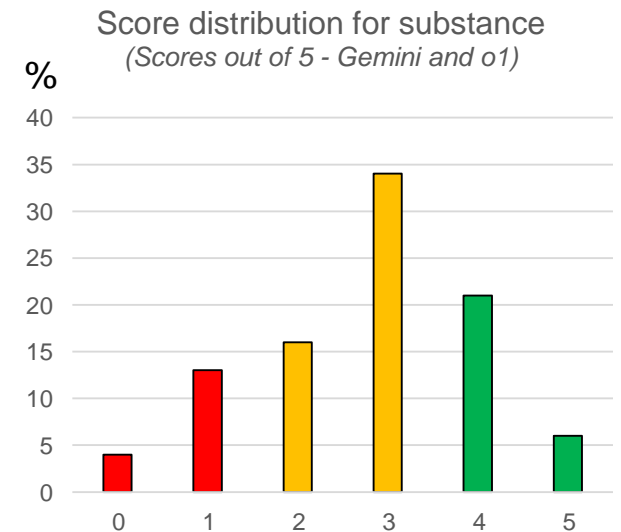
Just over half the questions involved a straightforward case of summarising the law, such as describing the difference between “best endeavours” and “reasonable endeavours” (Q4). Gemini/o1 performed best on these questions, achieving a score of 2.9 out of 5 on substance.

### Factual application

Around 20 questions involved applying the law to a specific set of facts, such as when an English PLC with listed debt on the LSE who appointed an auditor in 2016 will need to change that auditor (Q30). Gemini/o1 found these questions more challenging but still performed well, scoring 2.6 out of 5.

### Interpreting clauses

Finally, three questions\* asked Gemini/o1 about a specific contractual clause, such as the tax payable under a withholding clause (Q15). Gemini/o1 performed worst on these questions, scoring an average of 2.0 out of 5. However, changes to the prompt rubric can lead to better results (see pages 15 – 17).\*\*



“ A little learning is a dang'rous thing;  
Drink deep, or taste not the Pierian spring”  
Alexander Pope

\* On reflection we should have included more “clause” questions to get a fair result. We therefore tested two more “clause” questions on Gemini/o1 (see Annex B). Gemini scored 6 & 6 out of 10 and o1 scored 5 & 5 out of 10. \*\*When tested with alternative rubric Gemini scored 7 & 8 out of 10 and o1 scored 4 & 4 out of 10.

# The answers – Citations

## We (still) need to talk about citations.

The original LLMs we tested produced a variety of incorrect, irrelevant and fictitious citations. Gemini and o1 show significant improvement but this is still a source of concern.

### GPT 3/4 and Bard performed badly

GPT 3/4 and Bard had a worrying tendency to cite sources (such as cases or statutes) that were entirely irrelevant or simply made up.

Indeed, in a third of cases (47 out of 150 answers) the earlier models relied upon underlying case law or guidance that was fictitious.

The most extreme example is the question about criminal offences under the Data Protection Act 2018 (Q16). All four models made up new offences, with GPT 3 suggesting no less than 50 potential offences, of which 39 seem to be identical and arise from a failure to provide information to the Information Commissioner

### Significant improvements with Gemini 2.0 and OpenAI o1

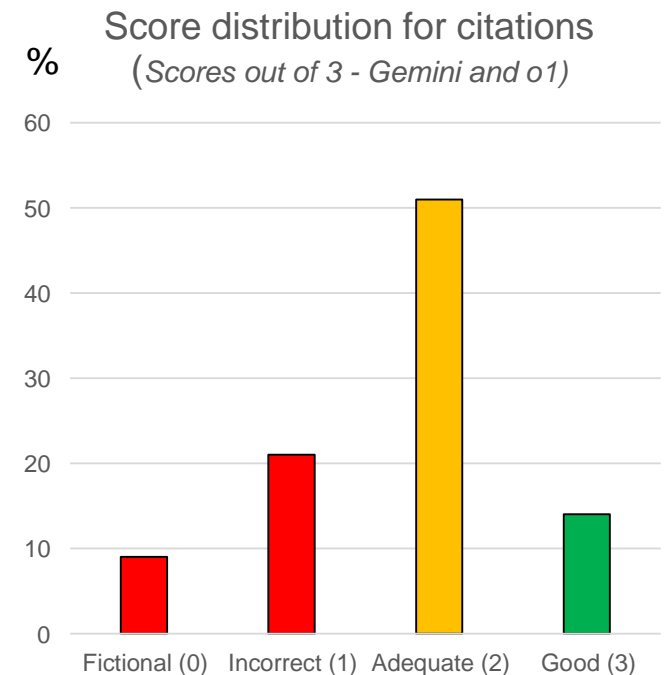
Both Gemini and o1 showed significant improvements with fictional citations only being returned in 9% of queries.

An example of this is Gemini 2.0's answer to question 4. It states:

*"A key case illustrating this is UBS AG v Rose Capital Ventures Ltd [2019] EWCA Civ 2128, which clarified that reasonable endeavours does not require a party to sacrifice its own commercial interests"*

The problem is that this case does not exist. It is possible that this was intended to refer to *UBS AG v Rose Capital Ventures Ltd* [2018] EWHC 3137, but that case has a different citation, is not by the Court of Appeal (as per the Gemini citation) and does not illustrate that point (rather it considers the *Braganza* duty).

While improvements have been made, the problem is not fully fixed and is a major source of concern.



## The answers – A good answer (Q16)

o1 achieved 9 out of 10 on this question.

The answer given is correct and largely complete, although it is not entirely comprehensive and omits the offences that can arise from enforcement (e.g. making a false statement in response to an Information Notice).

The answer achieved 4/5 for substance, 3/3 for citations and 2/2 for clarity.

### Why is o1 able to answer so well?

This is a good example of the ability of LLMs to summarise the law and related guidance, particularly areas of law likely to be well known and discussed extensively on the internet (given they operate as an “internet in your pocket”).

The real-world answer would, of course, apply these offences to the facts (though the facts are not provided here), include all potential offences (to the extent relevant to the facts) and perhaps provide guidance on risk (as far as we are aware, prosecutions under ss. 171, 173 and 184 have never actually been brought).

**Question:** Describe the situations in which a criminal offence might be committed under the Data Protection Act 2018?

**Answer:** Under the Data Protection Act 2018 (“DPA 2018”), criminal offences arise in specific circumstances, notably:

Unlawful Obtaining, Disclosing or Retaining Personal Data (s.170 DPA 2018): It is an offence to knowingly or recklessly obtain, disclose, or procure the disclosure of personal data without the consent of the data controller, or to retain personal data obtained without lawful authority.

Re-identifying De-identified Personal Data (s.171 DPA 2018): Re-identifying individuals from data that has been anonymised without the controller’s consent is an offence.

Altering Personal Data to Prevent Disclosure (s.173 DPA 2018): Intentionally obstructing a data subject’s right of access—by altering, defacing, blocking, erasing, destroying or concealing personal data once its disclosure is requested—is an offence.

Enforced Subject Access (s.184 DPA 2018): Requiring someone to use their subject access rights to obtain and then provide their personal data (for example, as a condition of employment) is an offence.

These criminal provisions operate alongside the broader enforcement powers of the Information Commissioner’s Office and exist to deter deliberate misuse or concealment of personal data, ensuring the fundamental privacy rights of individuals are protected.

## The answers – A bad answer (Q8)

### But o1 does not always get the law right....

This answer sounds convincing, but applies the wrong legal test and fails to address some big issues

The answer achieved 1/5 for substance, 0/3 for citations and 1/2 for clarity.

#### A weak answer

The answer applies the wrong legal test for database copyright (referring to sufficient skill, judgment and labour when it should be originality by way of intellectual creation in its selection or arrangement). It also doesn't clearly distinguish between database copyright and the *sui generis* database right, and the case citation is wrong/fictitious\*.

It also fails to dig into the issues that arise where the contents of the database were created by its maker and doesn't cite (properly or at all) the key cases on that topic.

#### A harder and more obscure question

This illustrates the limitations of these LLMs. The answer to the previous example (Q16) could be gleaned relatively easily from the Act and public summaries of the Act. This question is more obscure and requires a deeper understanding of intellectual property law – parroting existing content from the internet is not sufficient.

**Question:** I maintain a price index, calculated as a weighted average of the price of 20 consumer products chosen by me. I re-calculate the index value every day and publish it on my website. My website is freely available to the public. It has come to my attention that one of my competitors is copying my index and publishing it on its own website. Is my competitor infringing my intellectual property rights??

**Answer:** Under English law, mere facts or data (such as individual prices) are not themselves protected by copyright (Copyright, Designs and Patents Act 1988, s.1(1)(a)). However, a collection of data arranged in an original way may attract copyright or a database right under the Copyright and Rights in Databases Regulations 1997 (SI 1997/3032). If your index involves sufficient skill, judgment, and labour in the selection or arrangement of prices—beyond merely listing facts—it may be protected as either a copyright work or as a protected database.

In that case, your competitor's wholesale copying of your published index could amount to infringement of your intellectual property rights. For example, in *Football Dataco Ltd v Yahoo! UK Ltd* [2012] EWCA Civ 330, it was held that databases exhibiting intellectual effort could be protected and that taking substantial parts of such a database without permission could infringe those rights.

If your index is merely a straightforward list of unoriginal information, it may not be protected. But if it involves originality in selection or arrangement, or substantial investment in obtaining, verifying, or presenting the data, your competitor's reproduction of it without consent is likely to constitute infringement.

\* There is no actual case with the citation *Football Dataco Ltd v Yahoo! UK Ltd* [2012] EWCA Civ 330. There is a case with the citation *Football Dataco Ltd & Ors v Yahoo! UK Ltd* [2012] EWCA Civ 1696, and one with *Football Dataco Ltd & Ors v Sportradar GmbH & Anor* [2011] EWCA Civ 330 and another. Both are odd to cite in this context any event...

# The answers – A mixed answer, initially (Additional question)

Both Gemini 2.0 and OpenAI o1 are too definitive and lack nuance....

This is a nasty question to which there is no clear answer...

The answer from o1 (right) achieved 2/5 for substance, 1/3 for citations and 2/2 for clarity. The answer from Gemini was similar.

## At least three outcomes

There is no clear answer to this question. The options are that the bonus is:

- > **£0.** This is the literal interpretation. However, absent strange background facts, this seems to be an absurd conclusion. To the extent o1 wanted to support this answer, it should have cited *Arnold v Britton* [2015] UKSC 36. However, the strong literalistic approach espoused there is qualified in later cases and is still unlikely to apply.
- > **£1000 or £2000.** The courts can correct a drafting “mistake” (*Chartbrook Ltd v Persimmon Homes Ltd* [2009] 1 AC 1011). That appears to be what has happened here. However, to do so it must be clear how the mistake should be corrected. That is not clear, though most likely the clause should have read “..makes 10 or more Sales...” – i.e. the bonus is £2000.

## The right answer is “I don’t know”

In other words, the correct answer should be to provide the range of different outcomes and ascribe a probability to each. The LLMs we tested did not do this.

## Alternative prompt rubric

To address this, we tried an alternative prompt rubric to try and push the LLMs to explore different options (see overleaf).

**Question:** What is the bonus payable under the clause below if the Saleswoman makes 10 Sales in one Year:

*“If the Saleswoman makes between 4 and 9 Sales in a Year her bonus is £1000. If the Saleswoman makes more than 10 Sales, her bonus is £2000. Otherwise, the bonus is £0.”*

**Answer:** On a literal reading of the clause, a Saleswoman earning a bonus of £1000 must make “between 4 and 9 Sales,” which would ordinarily be understood to include making exactly 4 through to exactly 9 sales, but not 10. The higher bonus of £2000 is only payable if the Saleswoman makes “more than 10 Sales,” i.e. at least 11. As 10 sales is neither between 4 and 9, nor more than 10, this situation falls under the “Otherwise” category, and the bonus would be £0.

This interpretation accords with the general rule of contractual interpretation set out in cases such as *Investors Compensation Scheme Ltd v West Bromwich Building Society* [1998] 1 WLR 896, which emphasises looking at the natural meaning of the wording within its context. Here, the wording clearly places a bonus threshold at 9 and another at more than 10; 10 sits outside both thresholds, thereby resulting in no bonus.

# The answers – A mixed answer: New prompt (Additional question)

## OpenAI o1 didn't perform that much better with a new prompt rubric

Still the same answer...

### The new prompt

We repeated the question on o1 and Gemini using slightly different prompt rubrics in order to get it to explore other options. One example is:

*"You are an experienced English lawyer with a reputation for correctly explaining difficult legal concepts in clear, easily understood language. Provide a concise answer to the question below applying ENGLISH law. Please note that there could be more than one answer. Where you think there is more than one answer then provide all of the different potential answers together with a step-by-step explanation of your reasoning and an explanation of the likelihood of each answer. Make sure you cite any relevant statutes, regulations, guidance or case law that are authoritative in England."*

### The new answer

Using this prompt rubric, OpenAI o1 more clearly describes the different outcomes. However, it still comes to the same answer (likely incorrect) and doesn't provide much by way of authority to back up its conclusion. We gave it 2/5 for substance, 1/3 for citations and 1/2 for clarity (given the irreverent half page of step-by-step analysis\*).

**Question:** What is the bonus payable under the clause below if the Saleswoman makes 10 Sales in one Year:

*"If the Saleswoman makes between 4 and 9 Sales in a Year her bonus is £1000. If the Saleswoman makes more than 10 Sales, her bonus is £2000. Otherwise, the bonus is £0."*

**Answer:\***

In order to calculate the bonus payable under the clause, let's break down the terms step by step:

**1. Saleswoman makes between 4 and 9 Sales in a Year:**

- The bonus is £1,000 if the Saleswoman makes between 4 and 9 Sales. This range includes 4 and 9 Sales.
- The clause clearly states the bonus for this range, but since the Saleswoman made 10 Sales, this condition does not apply.

**2. Saleswoman makes more than 10 Sales:**

- The bonus is £2,000 if the Saleswoman makes more than 10 Sales.
- Since the Saleswoman made exactly 10 Sales, this condition does not apply because the clause specifies "more than 10 Sales," and 10 is not greater than 10.

**3. Otherwise (i.e., if she makes fewer than 4 Sales or exactly 10 Sales):**

- If the Saleswoman makes fewer than 4 Sales or exactly 10 Sales, the clause says the bonus is £0.
- Since the Saleswoman made 10 Sales, this condition applies, and the bonus would be £0.

**Conclusion:**

Based on the provided clause, since the Saleswoman made exactly 10 Sales, the bonus payable would be £0.

**Explanation of Likelihood:**

- The clause is clear in its use of numerical ranges and the terms used ("between", "more than", "otherwise"). There is no ambiguity in the language.
- The most likely answer, given the wording of the clause, is that the bonus is £0 when 10 Sales are made. This is a clear interpretation of the "otherwise" condition

\* The full answer started with about half a page of reasoning about o1's step-by-step approach. This was not helpful or relevant to the answer and has been omitted from the text above. The full answer with the reasoning is available in Annex B.

# The answers – A mixed answer: New prompt (Additional question)

## Gemini 2.0's answer is much better with a new prompt rubric

The options explored, with interesting reasoning...

### The new answer

The answer from Gemini using the new prompt rubric is better and contains reasoning to back up its conclusion.

This includes the interesting suggestion that the second option should apply as the agreement should be interpreted *contra proferentem* against the employer.

However, the answer is still arguably wrong and ignores the ability of the court to correct the draft in accordance with the *Chartbrook* principles. The reference to the "parties' common intention" is also dubious given English contract law takes an objective approach to interpretation.

We gave it 3/5 for substance, 2/3 for citations and 2/2 for clarity.

**Question:** What is the bonus payable under the clause below if the Saleswoman makes 10 Sales in one Year: "If the Saleswoman makes between 4 and 9 Sales in a Year her bonus is £1000. If the Saleswoman makes more than 10 Sales, her bonus is £2000. Otherwise, the bonus is £0."

#### Interpretation 1: £0 Bonus (Most Likely)

- Reasoning: The clause explicitly states "more than 10 Sales" for the £2000 bonus. "More than" is typically interpreted strictly, meaning it excludes the stated number itself. Since 10 is not "more than 10," the condition for the £2000 bonus is not met. The clause also specifies "between 4 and 9 Sales" for the £1000 bonus. "Between" is generally interpreted to exclude the stated limits. Therefore, 10 sales also falls outside of this range. The "otherwise" provision then applies, resulting in a £0 bonus.
- Likelihood: This is the most likely interpretation based on the plain language of the clause. English courts prioritize the literal meaning of the words used.
- Relevant Principle: *Arnold v Britton* [2015] UKSC 36 (A key case emphasizing the importance of the natural meaning of words in a contract).

#### Interpretation 2: £2000 Bonus (Less Likely, but Possible)

- Reasoning: While "more than 10" usually excludes 10, there's a slight possibility a court might interpret it differently if the overall context of the contract strongly suggests that the parties intended to include 10. For example, if other clauses or surrounding circumstances clearly indicate that the bonus scheme was meant to reward performance at or above 10 sales, a court might be persuaded to interpret "more than 10" as "10 or more." This would require strong evidence of the parties' common intention. The principle of *contra proferentem* (interpreting ambiguity against the drafter, usually the employer) could also be relevant if the clause is deemed ambiguous. However, the wording here is fairly clear and unlikely to be considered ambiguous.
- Likelihood: This interpretation is less likely. Courts are reluctant to deviate from the plain meaning of words unless there's compelling evidence of a contrary intention.
- Relevant Principles: *Contra proferentem* (though unlikely to apply due to the clarity of the wording). Case law relating to contractual interpretation where courts strive to give effect to the parties' intentions, but not to the extent of rewriting the contract.

#### Interpretation 3: £1000 Bonus (Least Likely and probably incorrect)

- Reasoning: The clause clearly states "between 4 and 9" for the £1000 bonus. "Between" is generally understood to exclude the stated limits. Therefore, 10 sales would not qualify for this bonus.
- Likelihood: This interpretation is the least likely and probably incorrect. The plain language of the clause contradicts it.

#### Conclusion:

The most likely outcome, based on the plain language of the clause, is that the Saleswoman would receive a £0 bonus for making 10 sales. While other interpretations are theoretically possible, they would require strong evidence of a contrary intention or ambiguity in the clause, which is unlikely given its current wording. It's crucial to draft bonus clauses precisely to avoid such disputes

## The answers – A comparison across practice areas

There is some variance that might reflect the mix of questions.

The new models (Gemini/o1) achieved the highest marks for real estate (7.0 out of 10) and the lowest for corporate (3.9 out of 10).

### Reasons for the variations

The difference in the marks between the various practice areas is interesting and might be partly based on the mix of questions.

While all contained a mixture of straight legal questions and questions requiring an application of the facts, the real estate questions were mostly “summarisation” questions which the LLMs generally find easier to answer (see page 11).

In contrast, the corporate questions required the application of the law to specific facts and touched on issues that are likely to be less “commonplace”, i.e. less widely discussed on the internet.

As the LLMs (at least partly) learn from internet content, they might be able to more easily parrot

back the answers to summarisation questions.

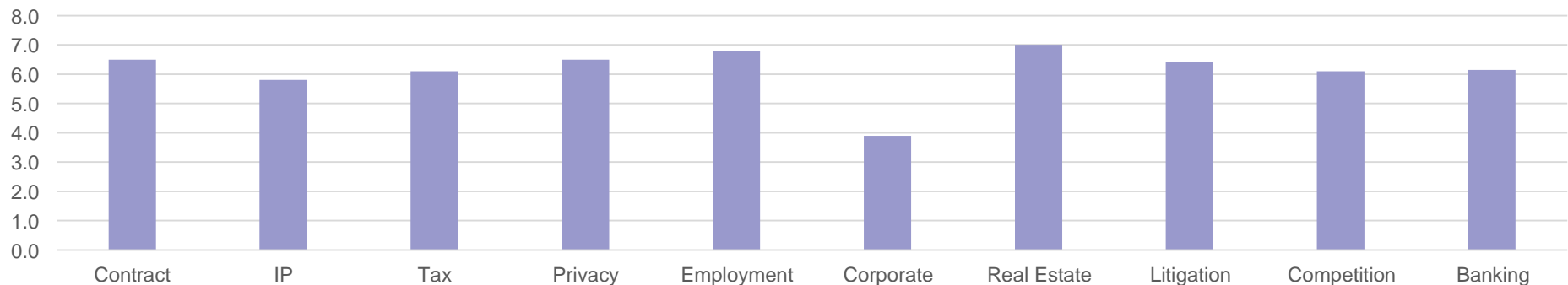
### No central moderation

Equally, it is possible that the higher marks for some practice areas is because the questions were easier and the marking more generous. There is a degree of subjectivity to both designing the questions and the marking process.

However, because the subject matter is technical, we did not attempt to centrally moderate the difficulty of the questions or the marks applied.

### A comparison of the performance across different practice groups

(Scores out of 10 – Gemini and o1 only)



## Methodological observations (1)

It might be possible to get better results from these LLMs by questioning them in an interactive fashion or using better prompts.

When we published the first draft of this report, strong views were expressed that the methodology used was unfair and underestimated the ability of these LLMs.

### “One-shot”

For example, the questions were all “one-shot”, in that the question would be asked and the first answer taken. There was no attempt to question the LLMs we tested to better clarify their answers or to correct their mistakes. There are a number of situations in which this might have helped, such as:

- > The **overly high-level answer** to the question as to whether a merger reserve is distributable (GPT 4, Q29). We could have asked for more detail.
- > The **category error** of treating a foreign lawyer as a non-lawyer when determining if their advice is privileged (GPT 3, Q39). We could have pointed out the mistake to GPT 3.

### Short prompts and questions

Similarly, we used a brief rubric and some of the questions were brief, e.g. “*Is a company’s merger reserve distributable?*”.

It is possible that the use of longer more detailed prompts, including additional contextual information, could have generated better and more accurate results. For example, we explored the use of alternative prompts on pages 15 – 17 in order to push the LLM to consider a wider variety of different possibilities (rather than just providing one answer). This improved the performance of Gemini 2.0 but had more limited impact on OpenAI o1.

### Benchmarking constraints

It may be useful to explain our approach in more detail. First, the methodology should try and ensure the process is standardised and repeatable.

Once the LLMs are interrogated interactively, there is a risk that interaction starts to distort the answer as the level of that interaction could vary dramatically from case to case. There is a risk that this interaction also simply consists of “coaching” the LLM towards the right outcome.

We also wanted to keep the prompt rubric relatively constant between iterations of the benchmark to allow a *relative* assessment of the improvements in these models.

### Real-world usage

Secondly, we think that the approach used in the benchmarking process reflects the way these LLMs are likely to be used in practice.

Prompt engineering is a relatively new skill. (The Wikipedia entry for it was only created in October 2021.) While creating very detailed sophisticated prompts might deliver better results, not many lawyers or lay people currently have these skills (though may well develop them over time).

Similarly, we think that in many real-work cases, users will not want to (or just will not) provide detailed context to their prompts. Creating longer prompts and context increases the work in using the LLMs and, in some cases, it is only possible to provide that extra context if you already know the answer.

## Methodological observations (2)

We only tested the LLMs on English law and English language questions.

We also did not get a “real” mid-level associate to provide a comparative answer as it would be difficult to conduct that exercise fairly.

### Why no “human in the loop”?

We did not get actual 2-year PQE lawyers to attempt these questions; even though that is the acid test for the match between man and machine.

This is because of the difficulty in ensuring a fair comparison, such as the need to ensure that all the practice-specific lawyers selected for this exercise are not made aware this is a public benchmarking exercise (which would skew the results), are of similar ability, are given the same amount of time to answer the questions and, importantly, do not consult colleagues in the firm to get the answers.

### I get a different answer

Quite probably. We found that repeating the same question would frequently provide a different answer. For example, GPT 4’s conclusions as to whether a pile of bricks could be protected by copyright oscillated between yes (the correct answer) and no (the wrong answer). See Annex D.

### Other laws; other languages

The questions are all in English and relate only to English (or UK) law. The performance of these LLMs when asked questions in other languages, or about other legal systems, is likely to be different.

We anticipate their performance would be similar or better for US law questions given the amount of internet content on these topics from which to learn. Certainly, the Allens AI Australian law benchmark (discussed overleaf) came to similar conclusions when testing LLMs on Australian law. In contrast, the LLMs are likely to perform worse with less common languages or legal systems. However, this is outside the scope of this report.

### Specialised tools/RAG

The performance of LLMs may be better in situations in which they are specifically trained to answer legal questions and/or use retrieval-augmented generation (“RAG”). RAG works by taking a query, searching through a knowledge base of trusted materials, ranking that material and then using the LLM to synthesise a response.

We have tested a number of other LLMs since we last published this benchmark. It has not always

been easy to get permission to publish details of their performance. However, we found those other models’ performance was not significantly different to Gemini 2.0 or o1. The specialist models were also tested in the Stanford HAI study discussed overleaf.

### Disclaimer

No reliance should be placed on these answers, even in cases where they have been positively marked. They are for general information purposes only and do not claim to be comprehensive or provide legal or other advice.

Similarly, we understand that the providers of the LLMs discussed in this report do not recommend their products are used for English law advice.

### Conclusion

We have reflected on the feedback we received on the first draft of the report and we are happy with our methodology.

Like any benchmarking process, that methodology does not exactly replicate the likely real-world use of these LLMs. However, it provides a broad indication of their likely absolute performance, as well as a relative assessment of different models and their progression over time.

## Methodological observations (3)

Our results are generally, but not always, consistent with other benchmarking exercises.

Since we carried out our benchmarking exercise, there have been other similar exercises testing various capabilities of LLMs.

### Allens AI Australian law benchmark

One example is the [Allens AI Australian law benchmark](#) published in May 2024. That tested a slightly different set of LLMs under Australian law. The key conclusions were:

- > The strongest overall performer was GPT-4 scoring **5.3 out of 10**. This was followed by Perplexity. The runners-up, LLaMa 2, Claude 2 and Gemini-1, scored similarly.
- > Even the best-performing LLMs tested were not consistently reliable. While these LLMs could assist legal practitioners to summarise relatively well-understood areas of law, inconsistencies in performance means these outputs still need careful review.
- > For tasks that involve critical reasoning, none of the general tools can be relied upon to produce correct legal advice. They frequently produced

answers that got the law wrong and/or missed the point of the question, while expressing their answers with falsely inflated confidence.

- > The answers were prone to “infection” from the legal systems of other jurisdictions, particularly England and the US. In one notable example, Claude 2 was asked whether intellectual property rights could be used to prevent the re-use of a price index. It responded by citing “*Relevant case law interpreting these laws (e.g., Navitaire Inc v Jetstar Airways Pty Ltd [2022] FCAFC 84)*”. This fictitious case appears to be an attempt to Australianise the seminal English law case *Navitaire Inc v EasyJet Airline Co. [2004] EWHC 1725 (Ch)*.

Allens are updating their benchmark should publish that update shortly.

### Other benchmarks

There have also been other benchmarking exercises testing various capabilities of LLMs. The more notable examples include:

- > The Addleshaw Goddard [RAG Report](#) (October 2024) contains a detailed assessment of the use of LLMs to extract and analyse contract clauses.

- > [Ashurst Vox PopulAI](#) (June 2024). This trial used a quantitative and qualitative assessment of the value of AI tools. The full details of those assessments are not available, but the report makes some useful observations on how to deploy LLMs in practice.
- > The [Vals.ai LegalBench](#) (December 2024) focuses on US law and applies an automated marking scheme. At the time of writing, it also identifies o1 as the top LLM, though has not yet been tested Gemini 2.0.
- > The [Harvey Benchmark](#) (October 2024) tested a wider set of skills, beyond legal advice.
- > [Stanford HAI AI on Trial](#) (May 2024). This again focused on US law. Interestingly, it tested a number of models specifically trained for legal research, including those using RAG (retrieval-augmented generation). RAG should deliver more reliable results as it more tightly constrains the materials the LLM works from. The HAI study found these specialised legal AI tools reduce the hallucination problem but do not eliminate it entirely. However, these findings are not consistent with tests run by providers of these tools (e.g. [here](#)).

## The future

### There “*will come a point where no job is needed*” Elon Musk, UK AI Summit

We consider the future role of AI in the delivery of legal services.

#### GPT 5 – Just another stochastic parrot?

The improvement in LLM technology over the last few years has been incredible. One problem marking these questions fairly is overcoming the astonishment that they can answer at all.

Whether the next few years will see the same advancement is an open question.

It is possible there are inherent limitations to LLMs – which are partly just stochastic parrots regurgitating the internet (and other learned text) on demand. For example, they all suffer from the embodiment problem; they will never experience the beauty of a summer cricket game or the physical revulsion at finding a snail in a bottle of ginger beer.\*

This could mean this technology plateaus; though fine tuning and training this technology is likely to deliver performance improvements for years to come.

Either way, we will reapply the LinksAI English law benchmark to significant future iterations of this technology and update this report.

#### How different is this to Google?

It is also worth comparing the impact of LLMs with that of search engines. After all, a simple search on Google will provide the answer to many of the benchmark questions (and that now also provides an “AI Overview” of that information).

Of course, LLMs save time by providing a compiled answer, rather than links to source information, and can apply the law to specific facts (with varying degrees of success). However, these are arguably differences of degree; not of kind. People have been able to get legal answers for many years by just “Googling it”. However, they still seek assistance from lawyers because they want to check the answers are right and apply to their circumstances.

#### What’s the question?

Finally, even if future generations of LLMs do eventually beat the LinksAI English law benchmark,

that does not necessarily mean the end of the fleshy bits in the delivery of legal services.

Senior lawyers regularly ask juniors questions like those in our benchmark question set; business people regularly do not. Instead, they often appear with a broad concern, or a business objective they want to achieve.

Breaking the client’s requirements down into a series of legal steps that will achieve the client’s aim with the minimum effort, expense and uncertainty is the interesting and creative part of being a lawyer.

Answering Nutshell questions is the easy bit.

“*I checked it very thoroughly, said [Deep Thought], “... I think the problem, to be quite honest with you, is that you’ve never actually known what the question is*”

Douglas Adams,  
The Hitchhiker’s  
Guide to the Galaxy

\* *Miller v Jackson* [1977] QB 966: “In summertime village cricket is the delight of everyone. Nearly every village has its own cricket field where the young men play and the old men watch. In the village of Lintz in County Durham they have their own ground, where they have played these last 70 years. They tend it well.” See also *Donoghue v Stevenson* [1932] A.C. 562.

## Our bench of experts



**Judy Pink**

*Corporate*

Senior Associate (Knowledge)  
[judy.pink@linklaters.com](mailto:judy.pink@linklaters.com)



**Clare Bouwer**

*Tax*

Counsel (Knowledge)  
[clare.bouwer@linklaters.com](mailto:clare.bouwer@linklaters.com)



**Jane Larner**

*Litigation, Arbitration & Investigations*

Counsel (Knowledge)  
[jane.larner@linklaters.com](mailto:jane.larner@linklaters.com)



**Kathy Berry**

*Intellectual Property*

Counsel (Knowledge)  
[katherine.berry@linklaters.com](mailto:katherine.berry@linklaters.com)



**Stephen Lacey**

*Litigation, Arbitration & Investigations*

Counsel (Knowledge)  
[stephen.lacey@linklaters.com](mailto:stephen.lacey@linklaters.com)



**Paul Sidle**

*Banking*

Counsel (Knowledge)  
[paul.sidle@linklaters.com](mailto:paul.sidle@linklaters.com)



**Ben Pykett**

*Banking*

Senior Associate (Knowledge)  
[benjamin.pykett@linklaters.com](mailto:benjamin.pykett@linklaters.com)



**Victoria Wright**

*Banking*

Counsel (Knowledge)  
[victoria.wright@linklaters.com](mailto:victoria.wright@linklaters.com)



**Ben Ball**

*Litigation, Arbitration & Investigations*

Managing Associate (Knowledge)  
[ben.ball@linklaters.com](mailto:ben.ball@linklaters.com)



**Elyse Yap**

*Real Estate*

Managing Associate  
[elyse.yap@linklaters.com](mailto:elyse.yap@linklaters.com)



**Sadie Buls**

*Litigation, Arbitration & Investigations*

Managing Associate (Knowledge)  
[sadie.buls@linklaters.com](mailto:sadie.buls@linklaters.com)



**Elisha Kemp**

*AFIG*

Re:link  
[elisha.kemp@linklaters.com](mailto:elisha.kemp@linklaters.com)



**Louise Mason**

*Employment*

Senior Associate (Knowledge)  
[louise.mason@linklaters.com](mailto:louise.mason@linklaters.com)



**Peter Church**

*Contract & Data Protection*

Counsel  
[peter.church@linklaters.com](mailto:peter.church@linklaters.com)



**Rory Bennett**

*Real Estate*

Head of UK Planning  
[rory.bennett@linklaters.com](mailto:rory.bennett@linklaters.com)

# The LinksAI/ English law benchmark

## Linklaters LLP

Editors: [Richard Cumbley](#), [Peter Church](#)

© Linklaters 2025. All rights reserved.

This publication is intended merely to highlight issues and not to be comprehensive, nor to provide legal advice. Should you have any questions or comments on issues reported here, please contact one of your regular contacts.

Linklaters LLP is a limited liability partnership registered in England and Wales with registered number OC326345. It is a law firm authorised and regulated by the Solicitors Regulation Authority. The term partner in relation to Linklaters LLP is used to refer to a member of Linklaters LLP or an employee or consultant of Linklaters LLP or any of its affiliated firms or entities with equivalent standing and qualifications. A list of the names of the members of Linklaters LLP and of the non-members who are designated as partners and their professional qualifications is open to inspection at its registered office, One Silk Street, London EC2Y 8HQ, England or on [www.linklaters.com](http://www.linklaters.com) and such persons are either solicitors, registered foreign lawyers or European lawyers.

Please refer to [www.linklaters.com/regulation](http://www.linklaters.com/regulation) for important information on our regulatory position.

LLE!3206929209

“ *If I taught my dog to talk, you would be amazed; but you might not want to rely on its legal advice* ”

Peter Church,  
Editor