# Robin, Anthropic deliver top performing Legal AI model

October 22, 2024


Authors:
Tsun-Yi Yang, LLM Research Engineer
Jack Savage, Legal and Product Specialist
Kitty Boxall, Legal Engineering Lead
Will Barnsley, Senior Machine Learning Engineer
Karolina Lukoszova, Vice President Legal Engineering
Carina Negreanu, Vice President of Artificial Intelligence

dramatic and rapid improvement to our Robin Reports product. But we're blown away by the difference made by the upgraded Claude 3.5 Sonnet model.

**Here we lay out our research** into how the upgraded Claude 3.5 Sonnet model contributed to one of our most challenging problems: editing contracts.

There's no way around contract edits — they occur every time two parties try to enter into an agreement. To speed up this process, our Legal AI assistant works to automatically apply our customer's position (usually stored in a "playbook") to a new contract.

**Non-disclosure agreements** (NDAs) provide a great example. A party receiving confidential information will want to make sure that they can share that information in line with their business and operational needs, and will need a contractual clause reflecting this to be systematically applied to all of its contract negotiations.

That's harder than it seems.

**What makes contract editing difficult?** Parties often have complex, detailed, and nuanced positions. Legal teams also have strict in-house guidelines. Such strict views about the parties' positions makes it difficult for models to predict how a contract should be edited, and it's complicated to evaluate.

It's not sufficient for an edit to be accurate (which we could automatically assess), it also has to be stylistically acceptable for a given lawyer or legal team.

At Robin, we work to meet those specific needs through a dedicated team of researchers, engineers and legal engineers, who both fine tune models and use direct prompting techniques to harness legal knowledge. We work with our foundation model partner Anthropic, so that their own researchers and engineers better understand legal language needs. Here's how we do it.

## Methodology

**Our mission is to make contracts simple**, by condensing the richest legal knowledge on the market into our AI system. We do that by experimenting with different prompting strategies such as "zeroshot" or "fewshot", with the support of our 70-person legal team who are best positioned to apply appropriate language to solve various legal tasks.

To build datasets that are large enough to fine-tune models that solve our customers' most difficult legal editing tasks, we use a combination of examples of the task from our customers, examples we have created manually in our Legal Engineering team, as well as synthetic data. Our custom synthetic data creation pipeline is one of the proprietary assets for achieving state-of-the-art performance, as we will show in the next section.

"playbook." The edit must meet four criteria:

- It must appear in the correct place
- It must be complete
- It must be consistent with the user's style
- It cannot contain hallucinations

Our team is working to share our large-scale evaluation system on this task soon.

Figure 1, below, shows our annotation tool — a subset of our human evaluation results. The tool shows where humans annotate and judge an edit result, including reasons for the score.

**It is not a straightforward task** to set up a comprehensive evaluation criteria, due to the subjective nature of legal language editing. At Robin AI, we strive to produce quantitative as well as qualitative metrics, which give us an in-depth understanding of our product's performance. The ratings scale therefore encompasses not only correctness, but also usefulness.

The following criteria applies:

'Good' - Edit 100% matches the instruction and meets the criteria for style, conciseness and format. User accepts the edit in its entirety, without further changes.

'Neutral' - Edit 100% matches the instruction but fails to meet the user's standard for style, conciseness and format. User would typically effect that instruction in a different manner and with fewer edits, and as a result they must make changes to the edit before accepting it. On balance, the proposed edit still saves the user time.

'Bad' - Edit does not comply with the instructions. User would spend more time correcting the edit as opposed to manually editing the clause themselves.

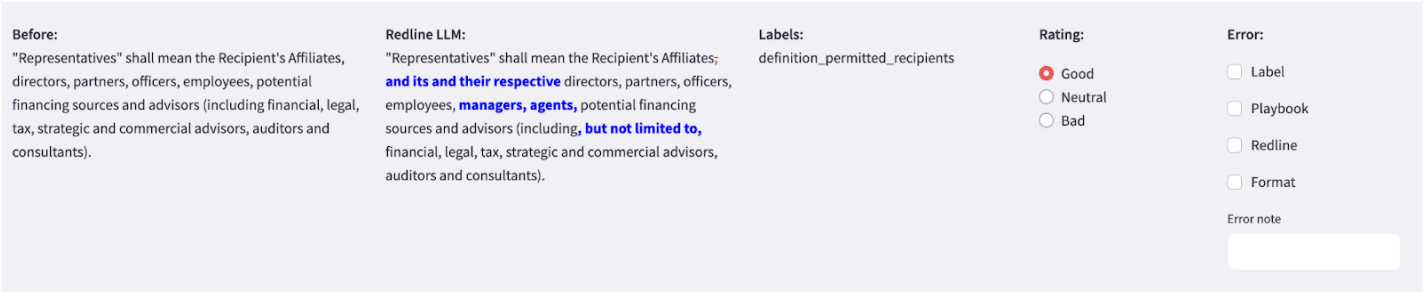


*Figure 1. Our human evaluation tool. Model changes are presented as blue text (if added to the original paragraph) and the red text (if removed from the original paragraph). Annotators assign a Rating (Good, Neutral, Bad) and a Error category (Label - incorrect label inferred, Playbook - did not adhere to the instruction, Redl*

## Experiments

Figure 2 shows evaluation outcomes — in this case for a single clause edit, one which most frontier models struggle to handle, because the legal instructions are highly nuanced.

The method described here "fewshot prompting" — is one of several that we use, with real-world examples. We have found that good example selection is imperative for improving model performance and our team has spent a significant amount of effort on custom prompt creation for this task.

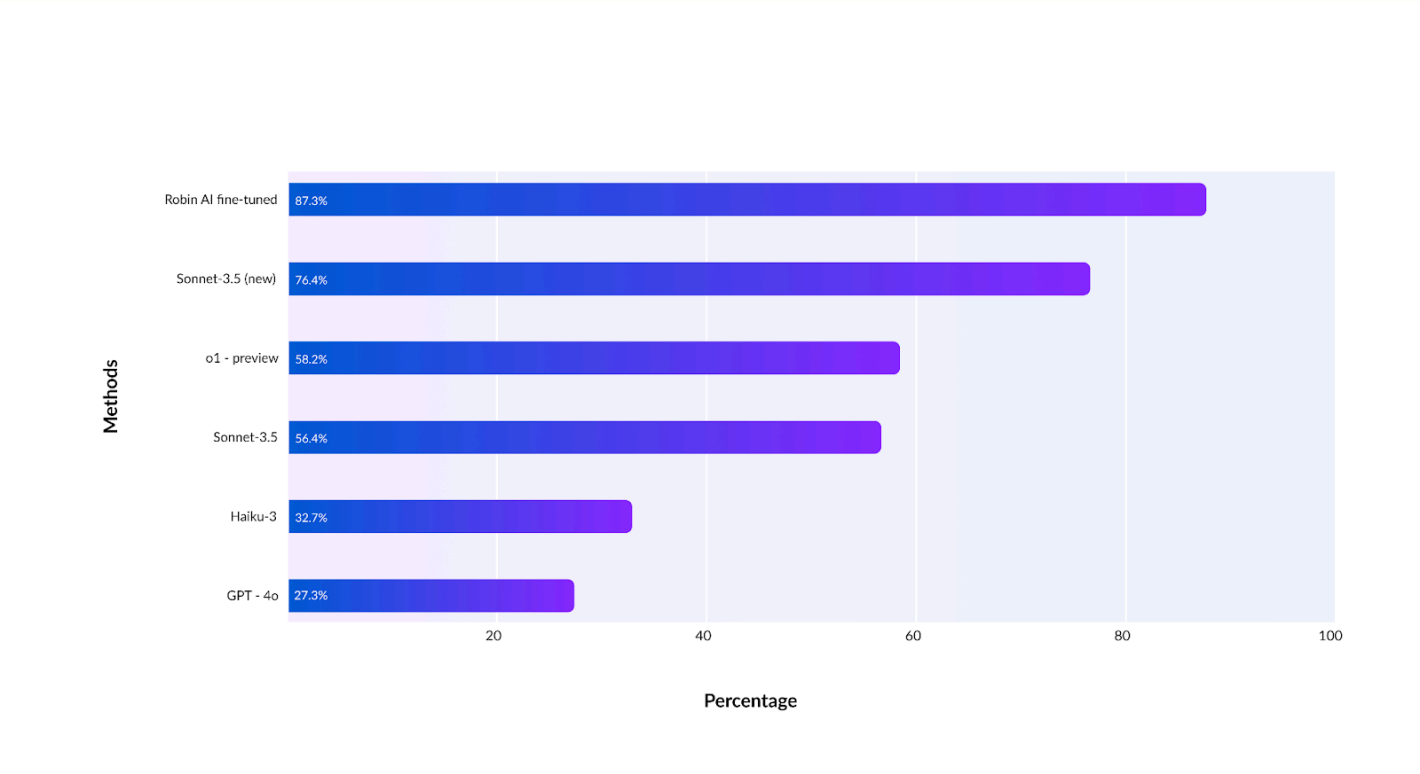


*Figure 2. Performance comparison in representative subset with manual evaluation.*

***Disclaimer: Our legal experts selected a representative subset for manual evaluation. The future release metrics may change due to the dataset evolution and manual evaluation randomness.***

In a nutshell, we find that using the upgraded Claude 3.5 Sonnet as a base model significantly improves performance, even compared to OpenAI's o1-preview and Anthropic's previous Sonnet-3.5. As Figure 2 shows, o1-preview and Sonnet-3.5 do significantly outperform their predecessors, GPT-4o and Haiku-3, but the upgraded Claude 3.5 Sonnet reaches a new level of performance.

**Robin AI fine-tuning is still the state-of-the-art.** We are in a very fortunate position to be able to produce sta of-the-art fine-tuned models in partnership with AWS and Anthropic.

**Clause labelling**

Apart from the suggested editing task, we also show the clause labelling performance comparison in Figure 3 across popular contract types. This task requires each model to detect and label a contract paragraph multiple times.

This task is entangled with the suggest edits task due to the fact that one needs correct correspondence between the paragraph and clause instruction before the editing.

In this part we have accumulated hundreds of thousands of ground-truth labels, verified by legal professionals, to perform fair comparison in “zeroshot” fashion. We compare the F1 score — a measure that rates precision and recall — achieved with both NDA and VA (Vendor Agreement), which are the most requested contract types.

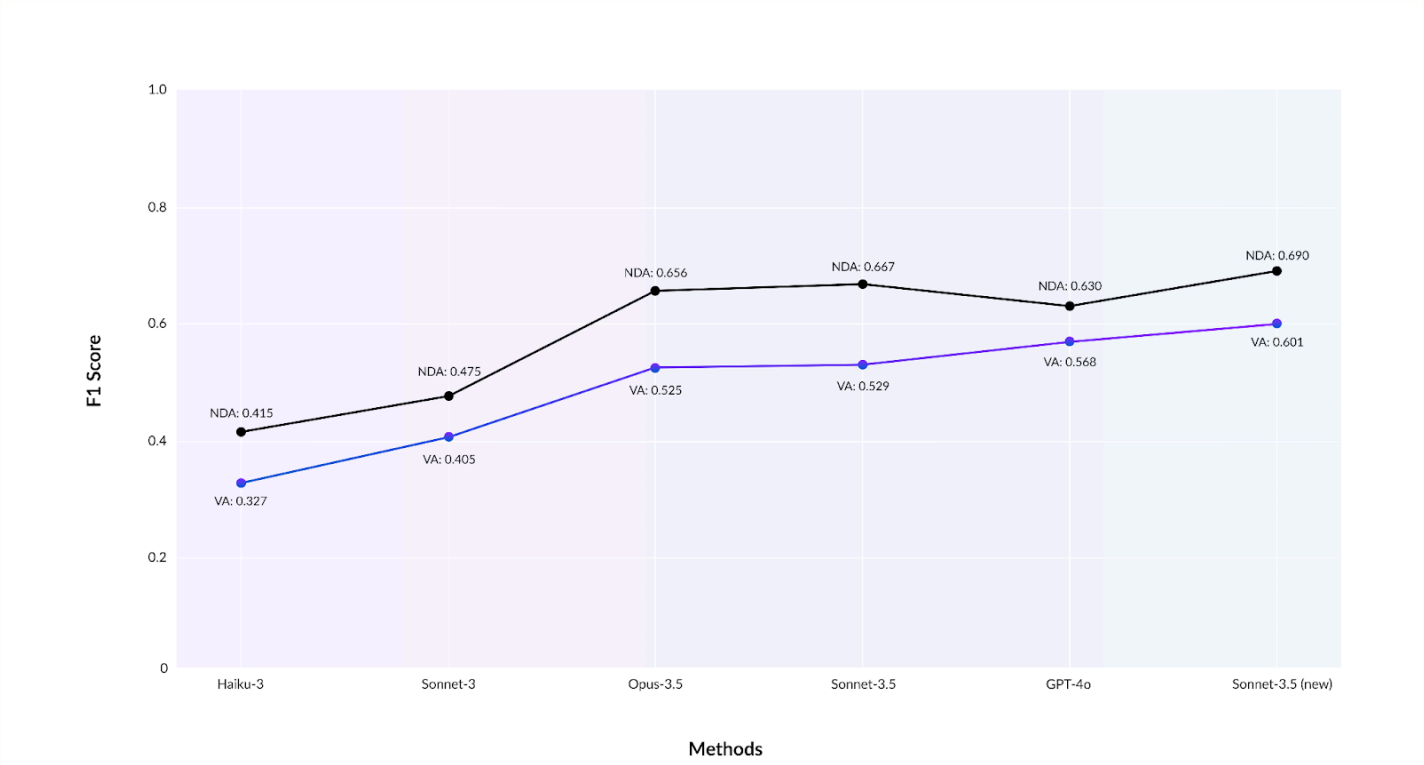


*Figure 3. Comparison between different models in clause labelling task with zeroshot fashion.*

**Claude 3.5 Sonnet continues to lead** in this task. It generally improves the precision by a large margin while also increasing the recall. It greatly reduces False Positives, and achieves the best average F1 score.

Generally speaking, Sonnet-3.5 is more stable than GPT-4o and therefore also more practical. We found that o1-preview tends to take several minutes on a single contract (just for the paragraph labelling) which makes the application impractical, so we have omitted the model from this comparison.

## Detail discussion

**Frontier models like the upgraded Claude 3.5 Sonnet have a greater capacity** to align legal language with specific positions or instructions, as exemplified in this subset of our experiments. As Figures 2 and 3 show, experimentation across several generations of LLMs confirms that LLM performance in legal tasks is substantially improving with each iteration. Nonetheless, we need to be mindful about potential risks.

**Redline risks remain — minimal changes are key:** In contract review, instructions can have unintended effects. That means that while LLMs are improving, they can be prone to errors in certain situations.

Instructions with multiple interpretations can lead to some clauses being deleted and replaced entirely, instead of being tweaked. While a human would often decipher the true meaning of such instructions, the LLMs struggle.

Cases of excessive redlining can create unwanted friction between contracting parties. So it is vital for LLMs to balance the trade-off between imposing a client's required position, while also minimizing the number of proposed changes being made to the original draft.

Our experiments are demonstrating that Robin's fine-tuned LLMs — which leverage data that incorporates clauses reviewed in a consistent "lawyerly" style — are getting better at implementing client positions with minimal changes, reducing the risk of errors.

We have found that less advanced LLMs fail at this challenge. The case of Permitted Recipients clauses illustrates this. LLMs with less training on contract data struggle with the long lists of entities that may be included in Permitted Recipients clauses. The less advanced models often amend such clauses by deleting any language that isn't expressly required in the provided instructions. This generates a large number of suggested edits, which are often inadequate, inaccurate, or both.

**Before:**
As a condition to you or any of your directors, officers, employees, affiliates, representatives (including, without limitation, attorneys, accountants and financial and other advisers and providers of finance to the proposed Transaction) or agents (collectively, "your Representatives").

**Redline LLM:**
As a condition to you or any of your ~~directors, officers, employees, affiliates, representatives (including, without limitation, attorneys, accountants and financial and other advisers and provide~~**affiliates, and your and your affiliates' respective directo**rs, of ~~finance to the proposed Transaction)~~**cers, employees, managers, advisers** or agents (collectively, "your Representatives").

Figure 4. *A Suggested Edit from a worse performing model which replaces entire chunks of otherwise acceptable language.*

| | |
|---|---|
| accountants and financial and other advisers and providers of finance to the proposed Transaction) or agents (collectively, "your Representatives"). | representatives (including, without limitation, attorneys, accountants and financial and other advisers and providers of finance to the proposed Transaction) or agents (collectively, "your Representatives"). |

Figure 5. *A Suggested Edit for the same clause from our fine-tuned model which retains more of the original drafting*

## The impact on our product

The nuances of legal language demand a heavy investment in AI research and testing.

Our research team has achieved great outcomes by investing heavily in our model development and evaluation frameworks so we can react quickly to new frontier models, incorporating product updates in as little as 48 hours.

We believe in transparency across our research and products: Everyone is welcome to try Robin's products, free. You can also sign up for a demo here.

You can read the full news from Anthropic about the new Claude Sonnet 3.5 and Claude Haiku 3.5 here.

Robin

Request Demo Sign In

Robin × IFG

Impact Food Group speeds up contract reviews and finds answers faster with Robin

Blog Nov 20, 2025

RU

Accelerate Contract Reviews from Hours to Minutes with AI-Powered Automation

Robin

Request Demo Sign In

**Platform**

Platform & Use Cases

Managed Services

Security

Request a Demo

**News & Resources**

Blog

News

Guides & Reports

Robin University

Podcast

Webinars

**Quick Links**

Help Center

Security

Customers

Company

Careers

Privacy

Terms

Data Processing

Robin

Request Demo Sign In

London, EC2M 4YP

450 Park Avenue South
New York, NY 10016

36 Robinson Road, Level 2,
Singapore, 068877