CASE STUDY: CASE-BASED RESEARCH

Intelligent document processing brings speed and objectivity to research

Llama 3 8B helps uncover the truth in the most challenging research cases

Meta

At a glance

Case-Based Research (CBR) provides automated research and assessment of documents for over 100,000 customers. CBR uses Llama to classify and evaluate cases that are too difficult for exact-match and sci-kit learn methods to classify, eliminating manual review for thousands of cases a week.

Industry
Technology (SaaS)

Use case
Automating document processing

Goal
Classify and assess the most complex information quickly and accurately

Llama version
Llama 3 8B Instruct (fine-tuned)

Deployment
Predibase

Company size
Over 100,000 global customers, millions of research cases per month

Results*

97%+
accuracy on the easiest 98% of research cases

90%+
accuracy on the most difficult 2% of research cases

5x
less expensive than the leading 4th gen LLM in tests

30x
faster response times in production versus the leading 4th gen LLM

*All results are self-reported and not identifiably repeatable. Generally, expected individual results will differ.

The challenge

Assessing information requires meticulous, unbiased review

Reviewing and assessing information is a challenging task for human experts. Documents and records often contradict each other, making it difficult to create an objective evaluation. Performing a thorough manual review of a single research case can take hours of an expert's time.

Case-Based Research (CBR) offers an automated platform for research that reduces manual review by 95%. The system uses a tuned logistic regression model that successfully processes 98% of the millions of research cases CBR completes monthly.

CBR

Case-Based Research (CBR) is a pseudonym for a successful SaaS company that uses AI and machine learning (ML) to extract salient details from multiple document types and create objective research reports. Due to the delicate nature of their work, we have omitted specific workloads and withheld the identity of the company and their employees.

The most difficult cases were the hardest to automate

The remaining 2% of research cases have more than 230 distinct data categories and often noisy documentation that even expert professionals struggle to review. CBR tried processing these edge cases with a deep neural network (DNN) model, but it could only process half of the outliers with acceptable accuracy. Tens of thousands of research cases still had to be processed by hand each month.

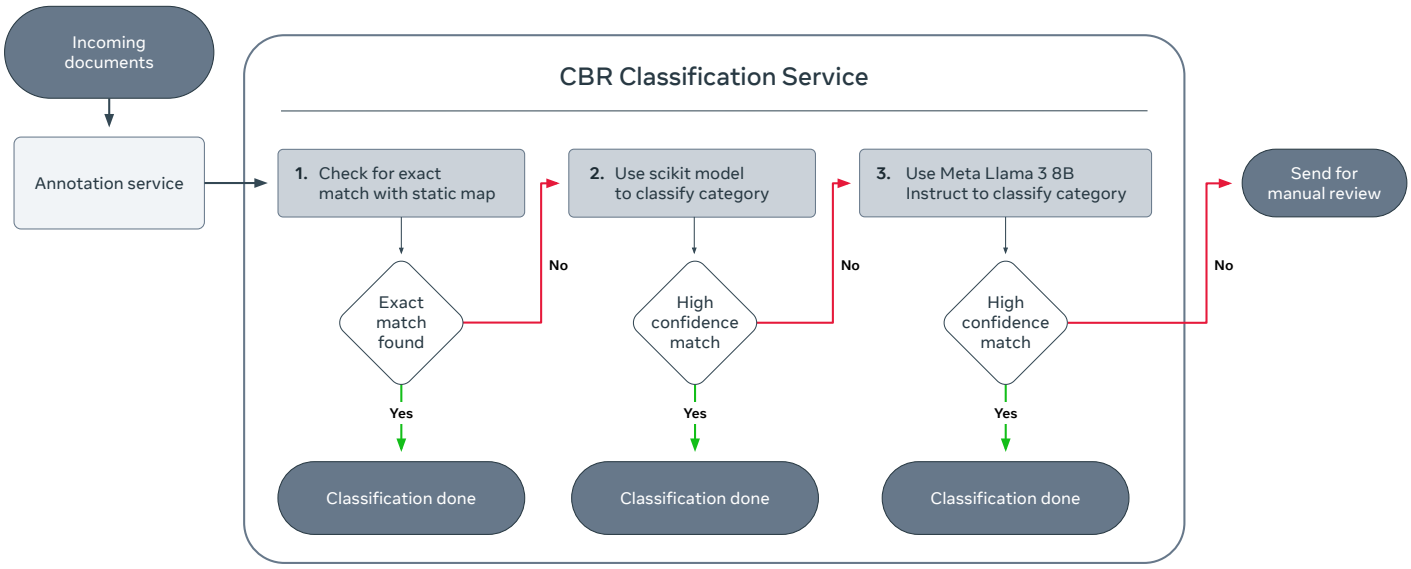
> "The DNN was able to classify only half of the most difficult 2% with decent accuracy, leaving the rest unclassified, resulting in more work for our customers. We needed a system with better performance, so we decided to explore different techniques using LLMs."
>
> Staff ML Engineer
> CBR

The solution

Llama cracked the hardest research cases

CBR used the Predibase platform to fine-tune a Llama 3 8B Instruct model and develop a classification and assessment application that reached 90% accuracy on the most difficult research cases while maintaining sub-second response times. By using a lightweight, small language model (SLM), CBR met their platform's low-latency SLAs with far less computing overhead and far lower costs than a commercial model.

> "By using Predibase to fine-tune and serve a Llama 3 8B Instruct model, we achieved better accuracy, faster response times and 5x reduction in costs compared to traditional generative AI approaches utilizing commercial LLMs."
>
> Staff ML Engineer
> CBR



CBR uses increasingly sophisticated classification services to process information efficiently.

Fine-tuned Llama 2 7B beat the leading 4th generation LLM in head-to-head experiments

The CBR team explored multiple techniques to find the best possible combination of accuracy, speed and price. Trials included 4th gen LLM, 4th gen LLM with training set via retrieval-augmented generation (RAG), fine-tuned Llama 2 7B and fine-tuned Llama plus 4th gen LLM. The results demonstrated the overwhelming value of task-specific fine-tuning.

Technique	Model	Prompt	Accuracy: easy research cases (98%)	Accuracy: difficult research cases (2%)	Round trip time (RTT)	Cost
Expert LLM	4th gen LLM	Workload, instructions, 230 categories	87.8%	81.8%	15 seconds	~$12k
Expert + RAG	Extend: 4th gen LLM + training set	Workload, instructions, 6 examples	95.8%	79.3%	7 seconds	~$7k
Fine-tuned LLM	Llama 2 7B	Workload	97.2%	85.0%	.5 seconds	<$800
Fine-tuned + Expert	Llama 2 Chat + 4th gen LLM		No gain	No gain	15 seconds	

Predibase fine-tuning and production environments contribute major gains

Throughout their experiments, the CBR team worked with different training and hosting environments. One platform stood out for its accuracy, consistency and performance: Predibase. Built on open-source LoRAX and Ludwig, Predibase is optimized for training and serving smaller, fine-tuned models.

Using Predibase, the CBR team fine-tuned a Llama 3 8B Instruct model on a 150,000-row dataset. The team compared full fine-tuning with Predibase's Parameter Efficient Fine-Tuning (PEFT), which uses Low-Rank Adaptation (LoRA) to reduce the number of parameters needed for fine-tuning. The team found that PEFT/LoRA delivered comparable performance to full fine-tuning in less time with lower training costs.

After fine-tuning, Llama 3 8B Instruct reached 90% accuracy on the most difficult 2% of research cases. In production on the Predibase platform, the model maintained 90% accuracy and 0.15 sec response times while processing full-scale traffic loads.

The outcome

Faster, fairer assessments for complex research cases

With Llama 3 8B Instruct and the Predibase platform, CBR has been able to automate processing for its most difficult research cases. This new capability saves their clients hours of expert labor and speeds up research processes. It also helps ensure that complex research cases receive thorough, unbiased reviews — a mission-critical goal for CBR.

97%+
accuracy on the easiest 98% of research cases

90%+
accuracy on the most difficult 2% of research cases

5x
less expensive than the leading 4th gen LLM in tests

30x
faster response times in production versus the leading 4th gen LLM

> "Using Llama and Predibase, we've achieved lightning-fast inference and reduced costs by 5x compared to the leading 4th generation LLM. Most importantly, we've been able to build a better product for our customers, leading to more objective, transparent and efficient research."
>
> Staff ML Engineer
> CBR

How can Llama enable your business?

Discover more about developing generative AI applications with open-source Llama large language models that bring unmatched control, customization and flexibility.

Learn More

Related Stories ▸

Meta