SaulLM-7,54,141B
- A pioneering Large Language Model for Law -

Louis Lhotte | William Charveriat
Rayan Lebaddi | Augustin Cobena
CentraleSupélec

January 2024

Table of contents

01
Introduction

Introduction

<u>Idea</u>: Deep dive on SaulLM-7,54B's development, and how to build a specialized LLMs from acquisition of data to training intricacies for state-of-the-art legal proficiency

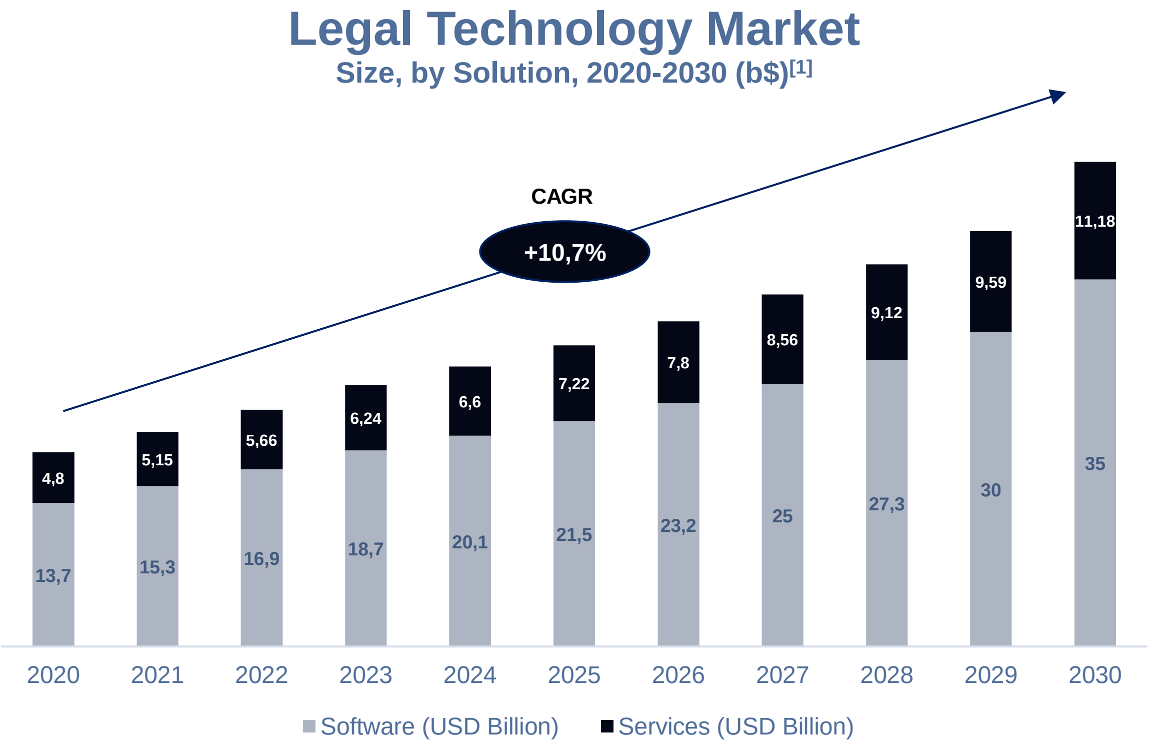
Objectives :

1. Introducing a new family of **Legal LLMs**
2. An improved **training & evaluation protocol** for legal LLMs
3. Model, code & Licensing for **innovation** in the sector

Innovations:

1. Introduction of SaulLM family of LLMs tailored to outperform its counterparts on Legal benchmarks.
2. Introduction of LegalBench-Instruct to better evaluate legal proficiency of LLMs (international & professional law, jurisprudence enriching)
3. MIT License to promote collaboration and adoption in the sector & research paper explicating each reasoning step for reproducibility or improvement projects.

Significant project in a rapidly growing market with double-digit CAGR



- **Rapid Growth** of Legal Technology Market
- Unmet demand for **legal automation** as document complexity increases
- LLMs could be a **key driver** of legal transformation

Source:
[1]: https://www.grandviewresearch.com/horizon/outlook/legal-technology-market-size/global

02

Extending the legal capabilities of Language Models

Proven Approach to Enhancing Mistral 7B: Overview and Key Steps[1][2]

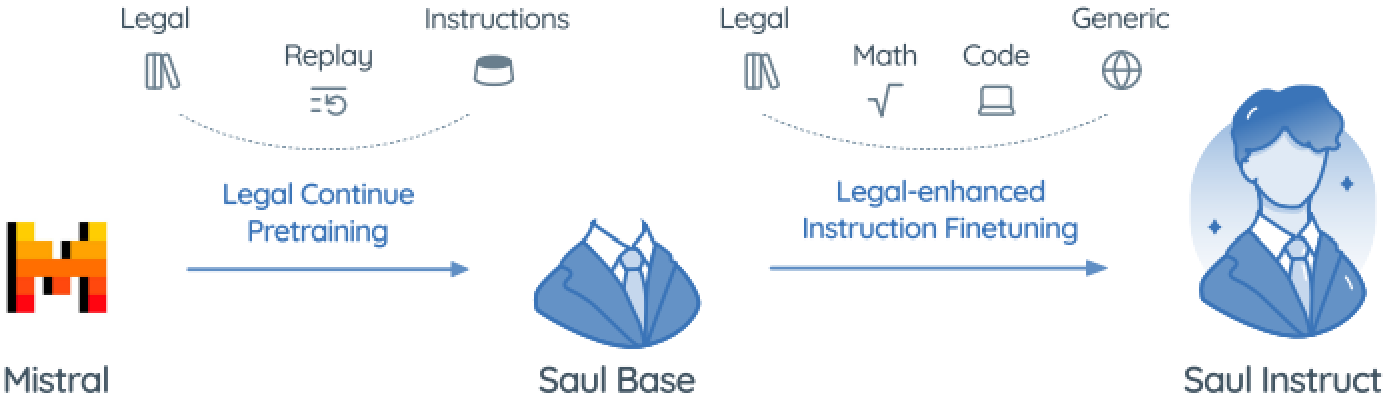
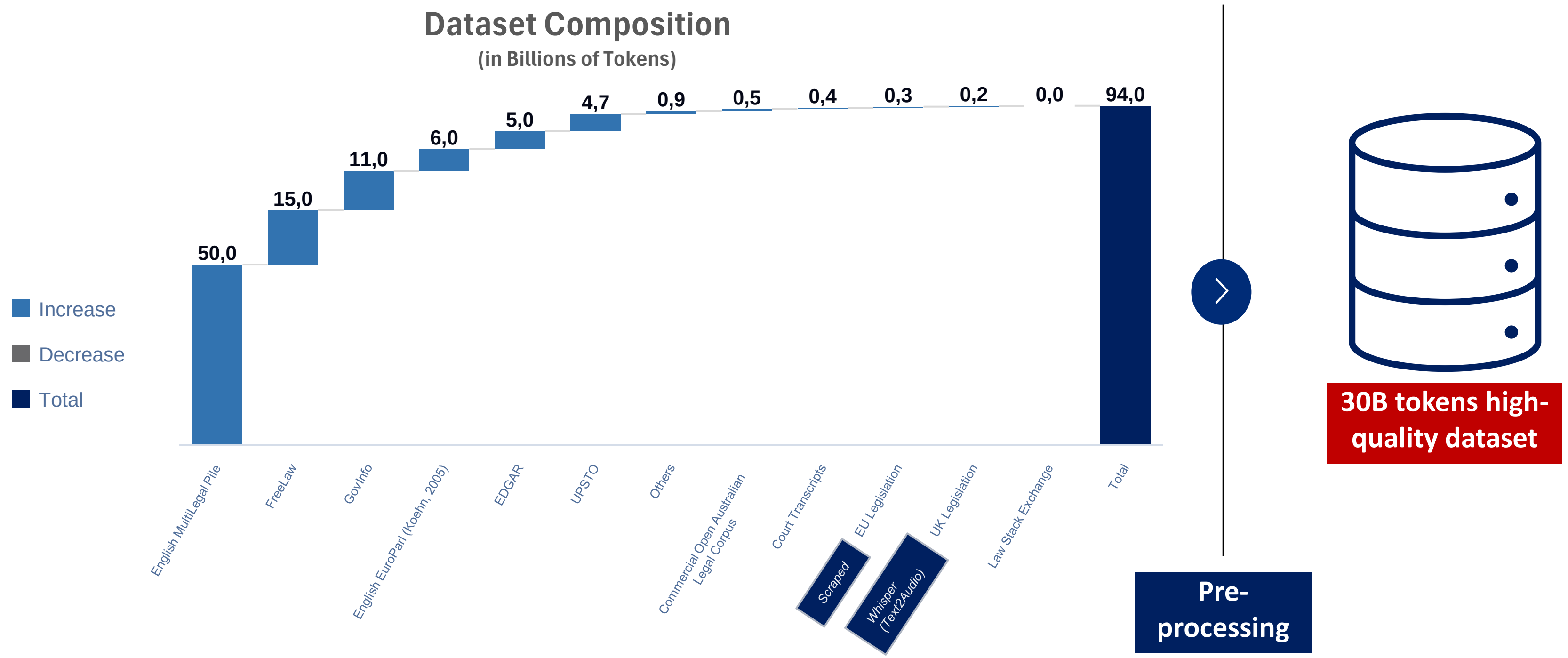


Figure 1: **Procedure for constructing `SaulLM-7B`.** We rely on legal datasets augmented with replay data, and instructions datasets. For fine-tuning we enrich our instruction finetuning dataset further with legal instructions.

Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. "SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain", Under review.

Legal Pretraining Corpora



First step: Legal Continue Pretraining

Goal :

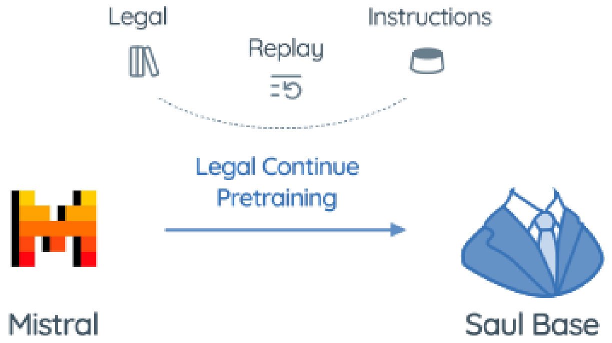
- Improve the model's **general knowledge** of legal matters
- Improve the model's **contextual understanding** of legal documents

Data :

- Massive **unlabelled dataset** of legal documents from multiple sources

Method :

- **Random token prediction** in the dataset (next token prediction, masked token prediction)



Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. "SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain", Under review.

Second step : Legal-enhanced Instruction Finetuning

Goal :

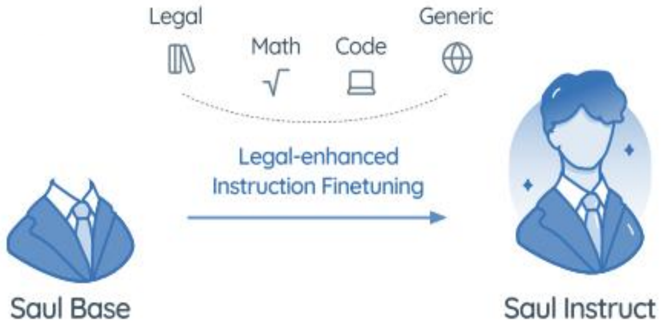
- Teach the model **how to follow human instructions effectively**, particularly in the legal context

Data :

- Generic **public datasets** for Instruction Finetuning
- Legal texts augmented with metadata for context

Method :

- First a **generic training** using existing methods and datasets
- Then, **transforming the legal texts into a coherent conversation** using a Mistral-7B-Instruct model, and train SaulLM to accurately answer the questions



Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. "SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain", Under review.

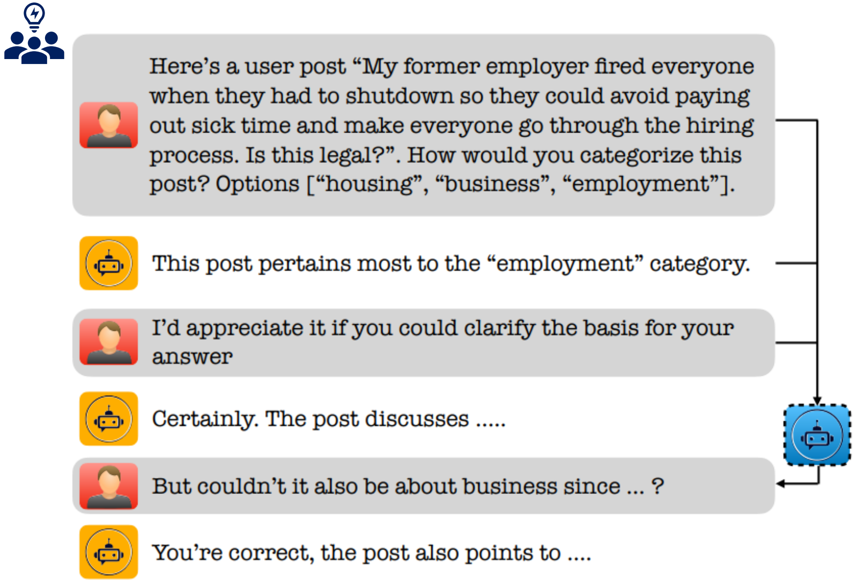
Instruction Finetuning for optimal legal expertise

General instructions

1. *SlimOrca*: Generic tasks (summarization, understanding)
2. *Meta Math QA*: Math inquiries (quantitative)
3. *UltraChat*: Conversational contexts (human-like dialogues)
4. *Glaive Code Assistant*: Code-based reasoning (queries)

Legal Instructions (through synthetic conversations):

1. Generated using Mistral-7B-instruct.
2. Covers legal document analysis via structured Q&A
3. Focused on reasoning and metadata interpretation



Example of turning a dataset with metadata into a conversation

Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. "SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain", Under review.

Obstacles & solutions

Obstacles:

1. Great variance across countries & jurisdictions

2. Catastrophic forgetting (i.e. losing Mistral's previously learned knowledge when trained on new data) + undisclosed Mistral datasets

3. Quality of data (artefacts from PDFs format such as page numbers, non-normalized Unicode chars, line breaks, repeated characters, …)

Solutions:

1. Focus on US, European & Australian jurisdiction across a diverse – but manageable range of legal systems. Especially, it avoids legal nuances that would lead to inconsistency

2. Retraining on commonly available data such as Wikipedia, github and StackExchange & inclusion of conversational data for robustness (FLAN collection, SPI…)

3. Data cleaning using Text Normalization (NFCK), Rule filters (removing top eight 10-grams e.g. "- - - - - - - - - -"), Deduplication, encoding issues deletion + HTML tags[2]

Trick: *Training of a KenLM model (Heafield[4], 2011, lightweight, domain-specific, open-source) on a small subset of curated legal data to filter any high-perplexity paragraph (i.e. surprising from the LLM perspective e.g. noisy data, weird characters etc…*)

Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[4]: Heafield, 2011, github: https://github.com/kpu/kenlm

03
SaulLM-54,141B

Advancing legal AI: saulLM-54B's innovations and enhanced capabilities over 7B

Motivation: While competitive, SaulLM-7B would show performance ceilings in handling more intricate reasoning or broader legal contexts, whereas a bigger model (SaulLM-54B) should show higher performance in tasks requiring legal understanding and generalization.

Difference with SaulLM-7B & LegalBERT & InCaseLawBERT:

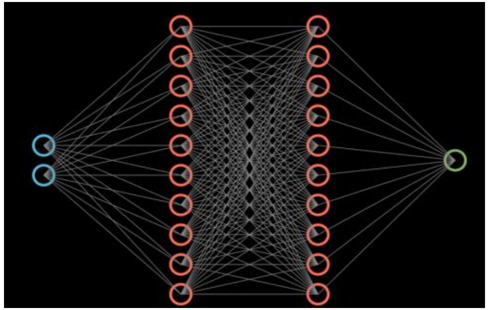
1. Pre-training dataset scale and scope
2. Instruction fine-tuning and specialization
3. Architectural enhancements

Innovations:

1. Larger 540B-token dataset with advanced cleaning and diversity (400 from web data)
2. SaulLM-54B incorporates richer (though synthetic still) legal-specific instructions through multi-turn interactions. Allows enhanced reasoning and domain adaptation
3. Mixtral-based **Mixture of Experts** (MoE) layers, supporting longer contexts (up to 32,768 tokens) and greater computational efficiency than SaulLM-7B.

Training Overview: LLM Architecture, GPU Design, and fine-tuning approach

Model Selection (Mixtral-54B resp. Mixtral-141B)



Model Dimension: 4096 (resp. 6144)

Hidden size: 14336 (resp. 16384)

Context Length: 32768 (resp. 65 536)

Pretraining: 8192 tokens

MoE layer: 8 experts

Infrastructure

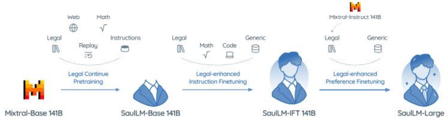
PyTorch

deepspeed

+

384 AMD MI250 GPUs (40% utilization), 64 AMD MI250 GPUs (fine-tuning), single AMD MI250 (evaluation), vLLM on NVIDIA A100 (synthetic data, library support)

Model Training process

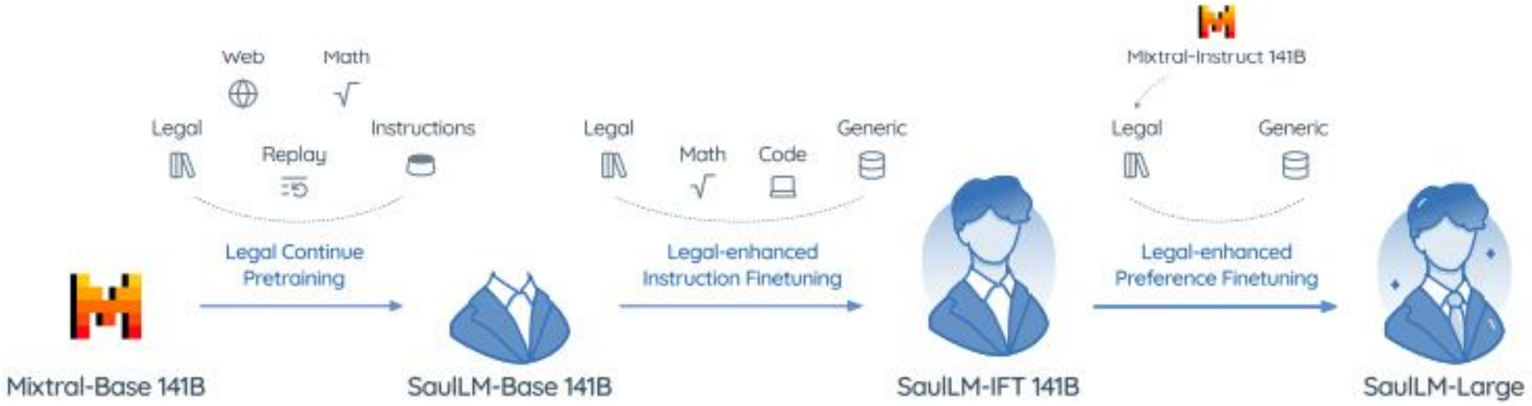


Optimiser: AdamW (β1 = 0.99, β2 = 0.90)

Lr = 2e-5 (resp. 1e-5 for IFT and 1e-6 during preference training)

Batch_size = 8 (resp. 4 for SaulLM-141B)

Progressive training: continued pretraining, instruction fine-tuning, and preference optimization



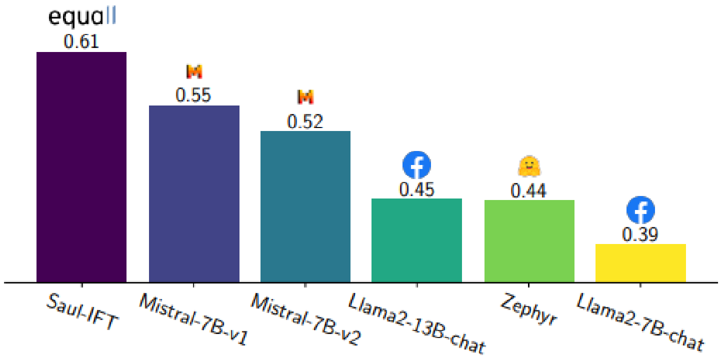
Generation process:

(1) initial legal-specific pretraining to enhance foundational knowledge
(2) instruction fine-tuning incorporating metadata like document type and issue date for more precise legal reasoning
(3) preference fine-tuning to progressively refine responses, enabling the assistant to better unpack legal reasoning and perform complex analyses

Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. “SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain”, Under review.

04
Results

SaulLM-7B exhibits stronger results on all 3 MMLU tasks



Performance of base models on LegalBench-Instruct

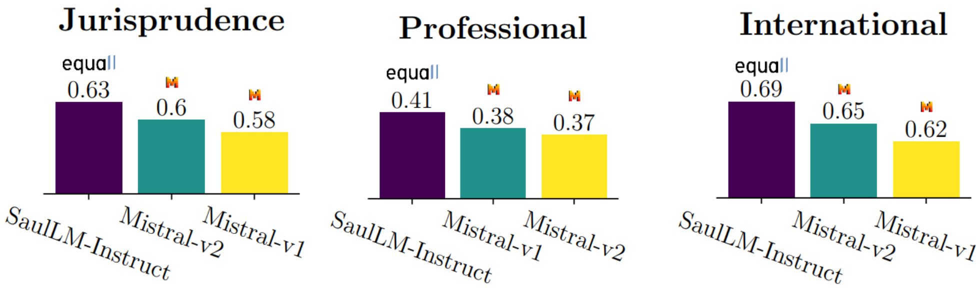


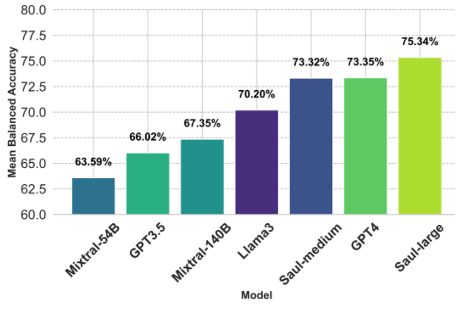
Figure 6: Instruct models on Legal-MMLU.

SaulLM-7B raises the bar for legal AI, achieving state-of-the-art performance through fine-tuning on tailored legal datasets

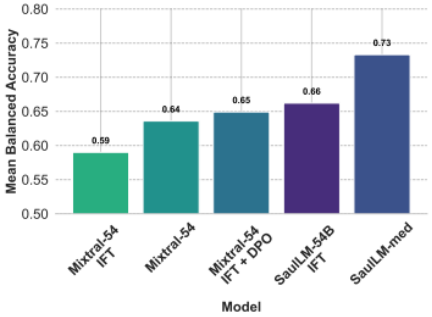
Echoing finding on LegalBench-Instruct, SaulLM-7B-Instruct displays superior performance on all three tasks of Legal-MMLU, with an average absolute improvement of 5 points with respect to Mistral-7B-Instruct-v0.1.

Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. “SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain”, Under review.

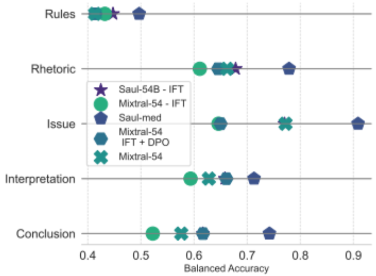
Continued pretraining allows SaulLM-54,141B to surpass its counterparts



Comparison of SaulLM-large and SaulLM-medium with existing models



Global Analysis. Role of continued pretraining.



Category Analysis: Role of continue pretraining.

SaulLM models outperform existing benchmarks, with SaulLM-large achieving the highest accuracy. Continued pretraining significantly enhances performance, particularly in domain-specific tasks like rhetoric and interpretation, underscoring its value in specialized fields such as law.

Source:

[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science

[3] Pierre Colombo et al. “SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain”, Under review.

05
Conclusions & Limits

Conclusions & Limitations

In these papers, the researchers introduced state-of-the-art models specifically **tailored for legal applications**, achieving **better performance** on specialized legal tasks compared to common multipurpose models such as Llama2, Mistral, and GPT.

Additionally, they contributed to the field by providing a cleaned version of the **LegalBench dataset** and new documents for perplexity **evaluation**, marking a significant advancement in the growing market of law-related software.

Unfortunately, the larger 54B and 141B models relied on proprietary datasets and required substantial computational resources for training, **posing challenges** for replication by other researchers and limiting **accessibility** within the broader academic and open-source communities.

Source:
[2] Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
[3] Pierre Colombo et al. “SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain”, Under review.

07
ANNEXE

Annex 1 - Sources

- **[2]** Pierre Colombo et al. "SaulLM-7B: A pioneering Large Language Model for Law," 2024. HAL ID: hal-04574874. Available at: HAL Open Science
- **[3]** Pierre Colombo et al. “SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain”, Under review.
- **[4]**: Heafield, 2011, github: https://github.com/kpu/kenlm
- **[5]** Guha, Neel, Daniel E. Ho, Julian Nyarko, and Christopher Ré. "LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models." 2023.
- **[6]** Jiang, Albert Q., Alexandre Sablayrolles, et al. "Mistral 7B." 2023.
- **[7]** Gao, Leo, et al. "The Pile: An 800GB dataset of diverse text for language modeling."
- **[8]** Hendrycks, Dan, et al. "Measuring massive multitask language understanding." 2020.

Questions

- How did you test that RLHF did not increase performances (as it would be greatly expensive and time consuming to do. Without testing it, it is difficult to see if it increases performance, and if you do it with small supervised corpus, it might not be efficient indeed but is not justification for no increased performance)?
- Deduplication in continued training → Does it not decrease colinearity (or its equivalent) and decreases generalization if you consider that there might be a distribution shift?
- How to determine quality of corpus regarding legal tasks?
- Sum of tokens Table 1 3,1,1 does not amount to total (1B tokens missing → from what sources?)
- Pre-processing → Majority of PDFs. If in image, how did you convert them in to strings? Did you use computer vision, and if yes, what worked best (Tesseract, OpenCV? Or discarded those that were in image formatting?)
- Why not putting gpt-4-o1 in the benchmarks, knowing that it is outperforming Mistral by a large margin? → Cost of tokens?
- More details on evaluation LegalBench to seek