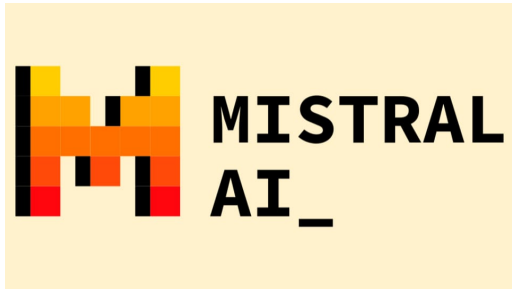
# Lextral

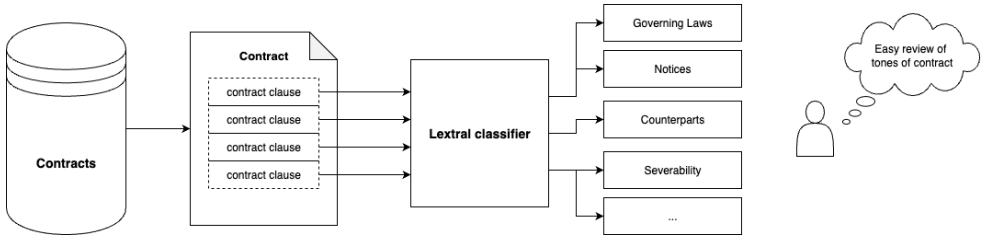
Contract Clause Classification

Julien Delavande
Mistral Use-case Take Home

13 August 2025



# The Use-Case in One Picture

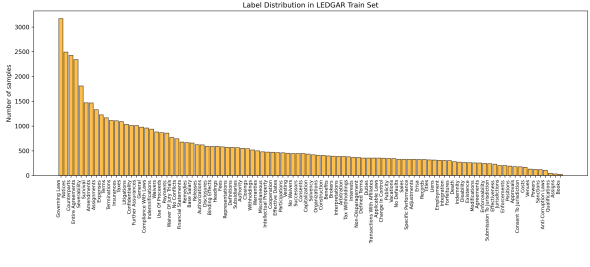



Reduce review time, increase consistency.

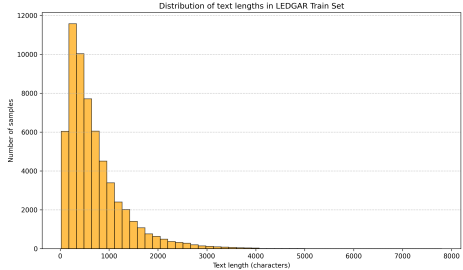
# Dataset Snapshot

Dataset: Lexglue/ledgar [Chalkidis et al. 2022]
60k clauses in the train set, 100 consolidated labels

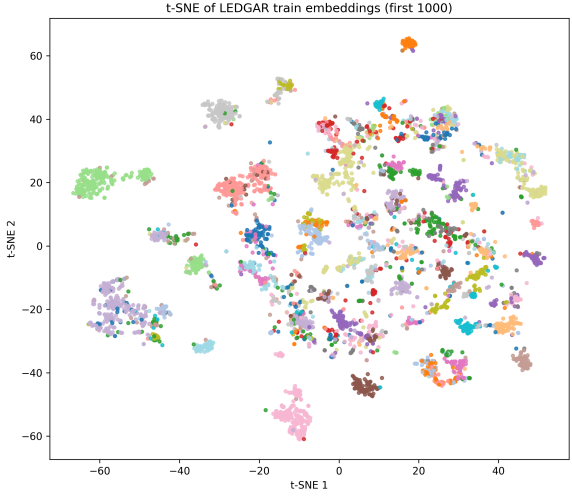
**Labels (Train)**




**Clause Lengths (Train)**

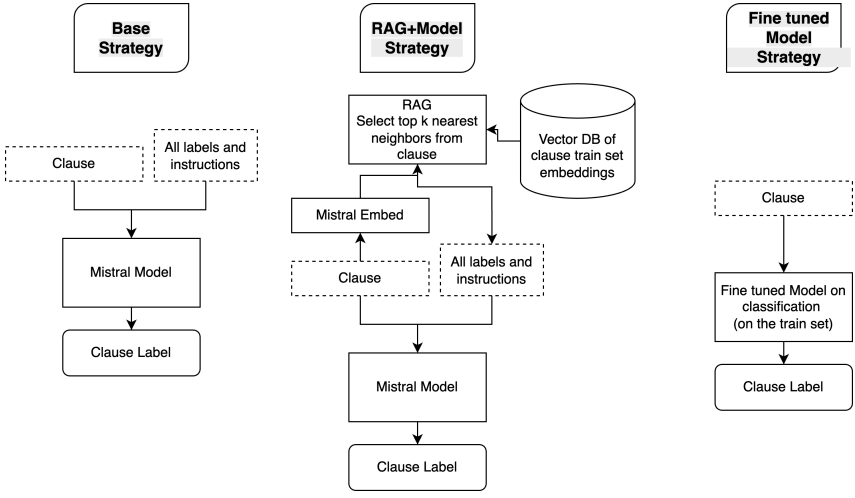


# Embedding Landscape (Train)



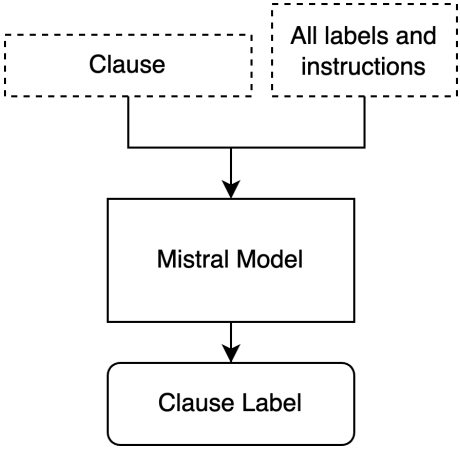

6k train embeddings; colors = labels. (t-SNE/UMAP)

# Strategies at a Glance




Zero-shot — RAG — Fine-tuned classifier

# Base Strategy




Clause + labels and instructions → Mistral model → Clause Label

**Model tested:** [Mistral AI team 2024; Mistral AI 2024]

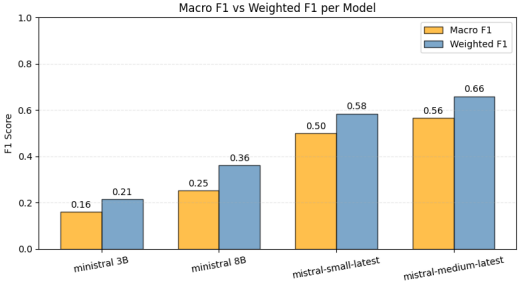
- Ministral-3B (API)
- Ministral-8B (API)
- Mistral-small-latest (API)
- Mistral-medium-latest (API)

**Test set: 1000 clauses**

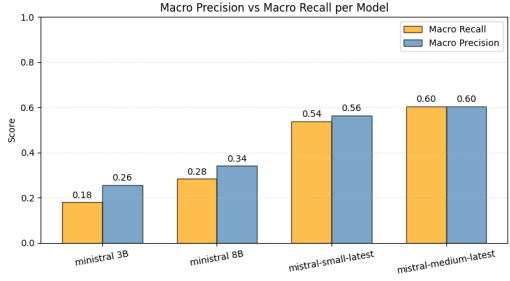
**Prompt:**

```
You are a contract clause classifier.
Classify the clause into one of:
{labels_str}
Clause: """{text}"""
Respond with only the category name.
```

# Base Strategy - Evaluation



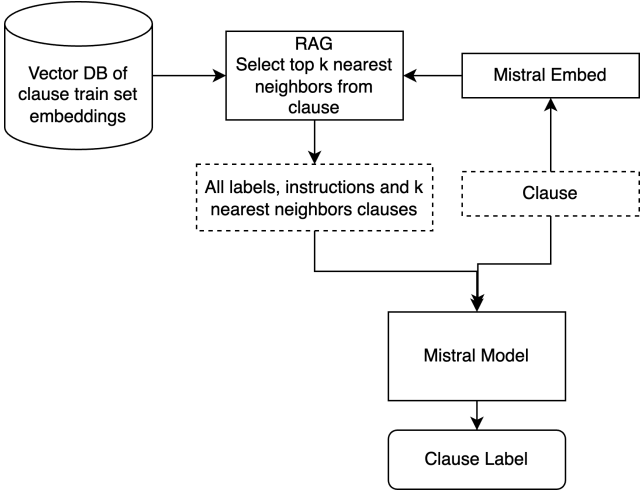

F1 scores per Model




Precision and recall per Model

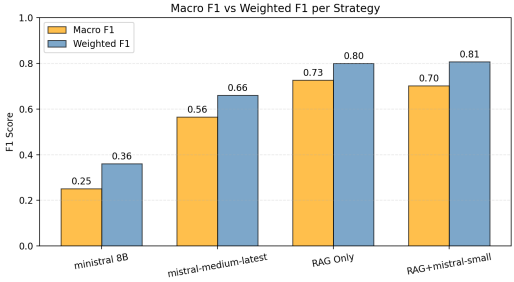
# RAG Strategy

[Lewis et al. 2020; Malkov and Yashunin 2016]

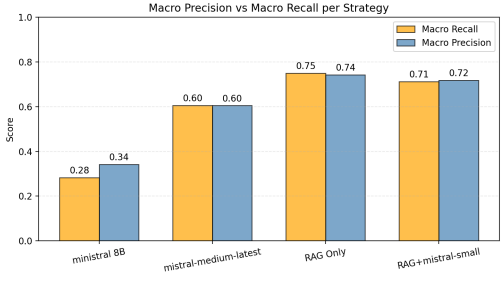



Embed query → pgvector (HNSW) → top-$k$ neighbors → few-shot prompt

# RAG Evaluation

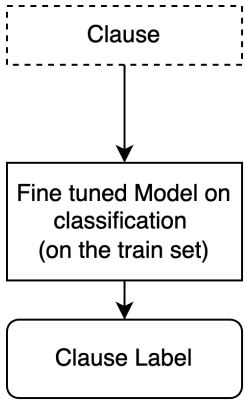



F1 scores per Model and RAG strategy




Precision and recall per Model and RAG strategy

# Finetuned Model Strategy

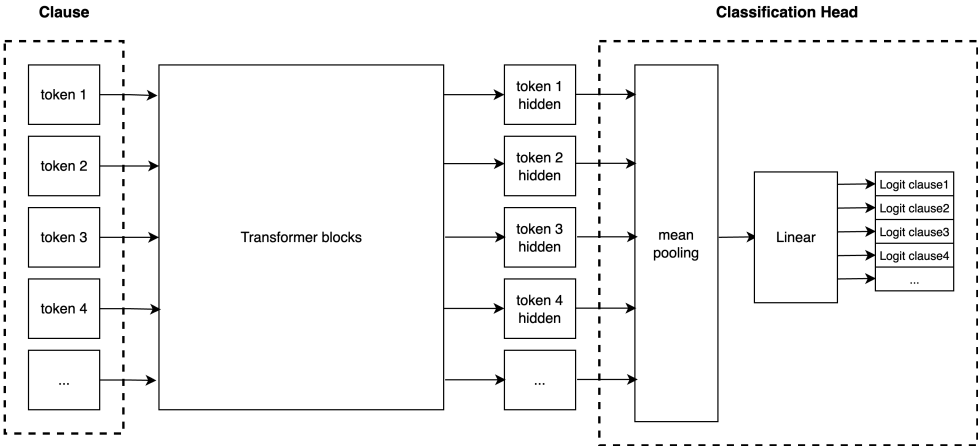



Clause → Ministral Classif finetuned model → Clause Label

**2 sub-strategies explored:**

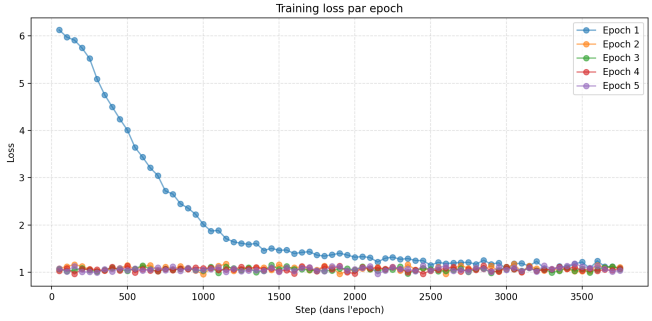
- Classification Head
- Classification Head + LoRA [Hu et al. 2021]

# Finetuned Model Strategy - Classification Head




Classification head architecture: only the head is trained

# Finetuned Model Strategy - Classification Head - Training

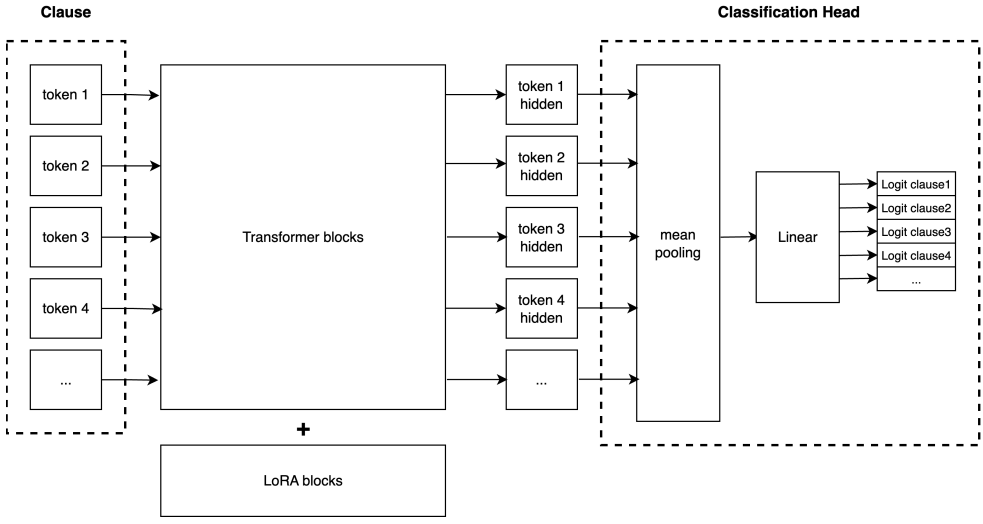



Evolution of the Loss of the classification head during training (Ministral 8B)

**Base Model: Ministral 8B**
**Training params:**

- train set len = 60e3
- BF16
- lr = 5e-5
- batch size = 4
- epochs = 5
- optimizer = ADAMW
- grad accum = 4
- scheduler = linear with warmup
- warmup ratio = 5e-2
- dropout = 0.1

# Finetuned Model Strategy - Classification Head + LoRA




Classification head + LoRA architecture: only the head is trained and LoRA matrices are trained

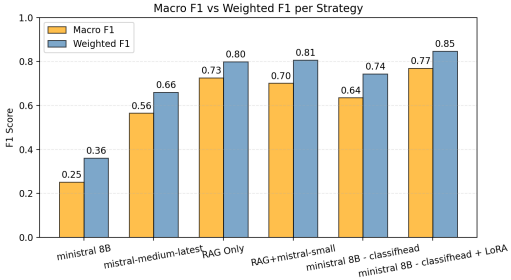
# Finetuned Model Strategy - Classification Head + LoRA - Training


Evolution of the Loss of the classification head + LoRA adapters during training (Ministral 8B)

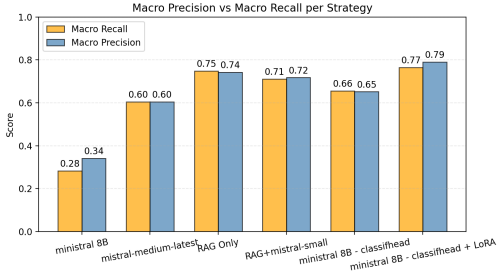
**Base Model: Ministral 8B**
**Training params:**

- Same as before
- target modules = qproj, kproj vproj, oproj, gateproj, upproj, downproj
- r = 4
- $\alpha = 16$
- LoRA dropout = 5e-2

# Finetuned Model Strategy - Evaluation




F1 scores per Model, RAG and finetuning strategy



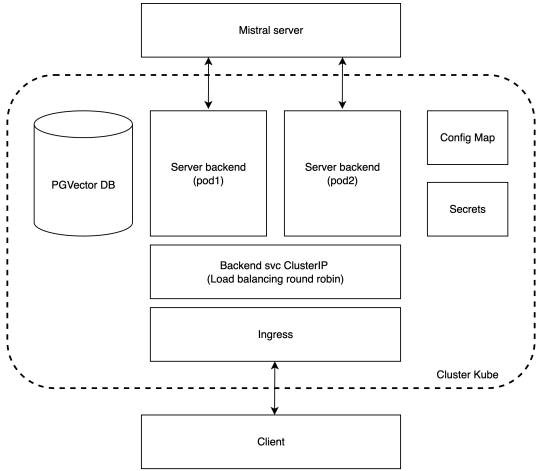

Precision and recall per Model, RAG and finetuning strategy

# Evaluation Overview

| Configuration | Latency (s) | Estimated Cost (USD) | Macro F1 |
|---|---|---|---|
| ministral-3B chatprompt | 0.5 | $0.43/m (10k user/day) | 0.1595 |
| ministral-8B chatprompt | 0.4 | $1.35/m (10k user/day) | 0.2517 |
| Mistral-small-latest chatprompt | 0.4 | $1.14/m (10k user/day) | 0.4989 |
| Mistral-medium-latest chatprompt | 0.8 | $4.80/m (10k user/day) | 0.5649 |
| RAG only (self-hosted cpu) | 5.5 | $390/m (10k user/day) | 0.7262 |
| Mistral-small-latest RAG (self-hosted cpu) | 5.2 | $1560/m (10k user/day) | 0.7013 |
| ministral-8B headclassifier (self-hosted A100) | 0.5 | $1800/m (10k user/day) | 0.6352 |
| ministral-3B headclassifierLoRA (self-hosted A100) | 0.5 | $1800/m (10k user/day) | 0.7690 |

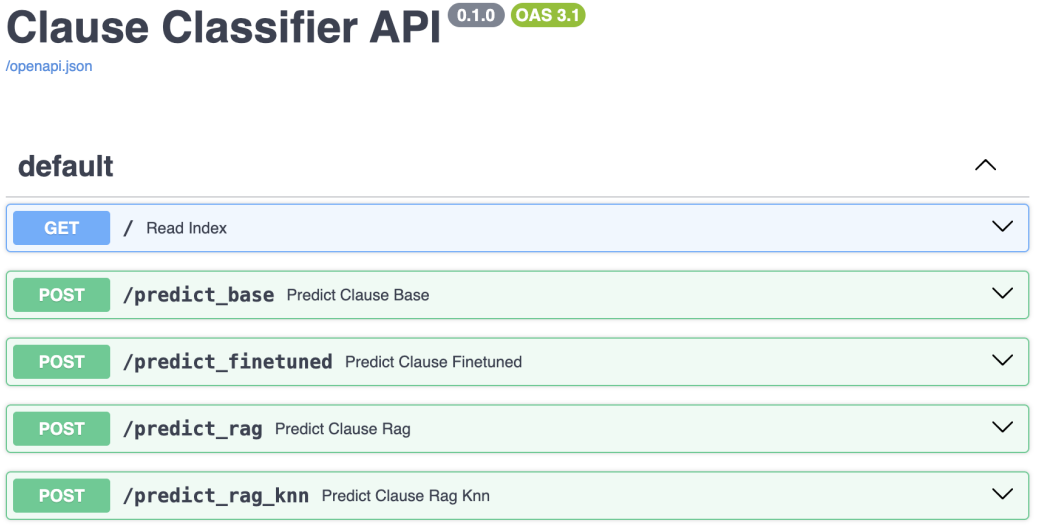
**Cost, Latency and macro F1 across strategies (m=month)**

# Infrastructure / Deployment




FastAPI, pgvector, Helm, Ingress, monitoring

# Infrastructure / Deployment - API



# Live Demo

**UI:** `https://lextral.delavande.fr`
**DOC:** `https://lextral.delavande.fr/docs`

**Thank you for listening**

juliendelavande@gmail.com

https://lextral.delavande.fr

https://github.com/juliendelavande/lextral

# Annex Cost Estimation Methodology

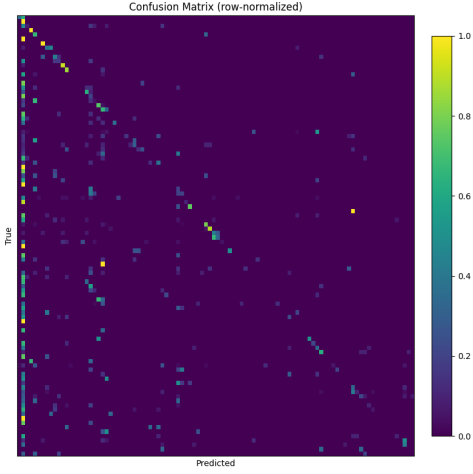
**Assumptions:**

- 10,000 requests/day $\Rightarrow$ 300k requests/month.
- API pricing from Mistral ($/M tokens), using 350 input + 10 output tokens/request:
  - Mistral Small: \$0.1/M*350 in, \$0.3/M*10 out $\Rightarrow$ \$0.00038/request.
  - Mistral Medium: \$0.4/M*350 in, \$2/M*10 out $\Rightarrow$ \$0.00104/request.
  - Ministral-3B: \$0.04/M*350 in, \$0.04/M*10 out $\Rightarrow$ \$0.00016/request.
  - Ministral-8B: \$0.1/M*350 in, \$1.00/M*10 out $\Rightarrow$ \$0.00016/request.
  - Mistral Embed: \$0.1/M*350 tokens $\Rightarrow$ \$0.000035/request.
- Machines rented 24/7 for one month:
  - CPU VM (16 vCPU / 64 GB RAM): \$0.50/h $\Rightarrow$ \$360/month.
  - L4 GPU: \$0.60/h $\Rightarrow$ \$432/month.
  - A100 80GB GPU: \$2.50/h $\Rightarrow$ \$1,800/month.
- RAG embedding cost only applied if computed at query time.

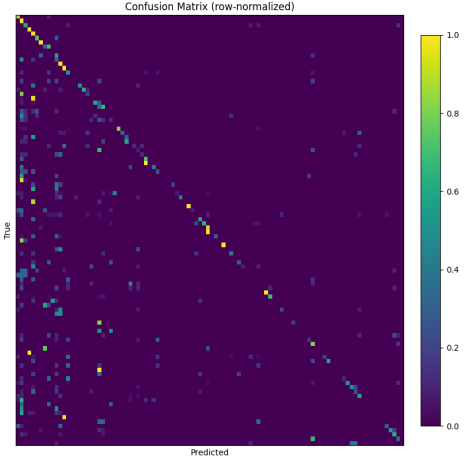
**Cost formula:**

$$\text{Monthly Cost} = \text{Infra/month} + \text{API cost/request} \times 300{,}000 + \text{Embed cost if applicable}$$

# Annex Confusion Matrix

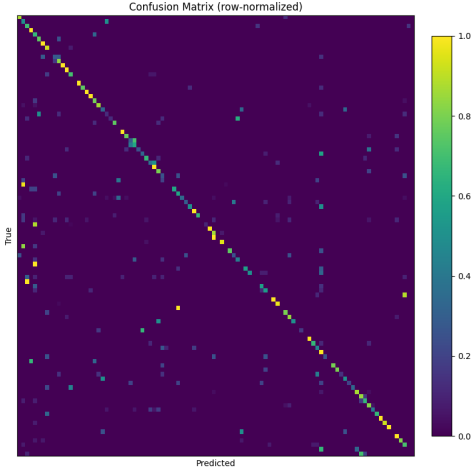



Confusion matrix for Ministral 3B per-class prediction distribution

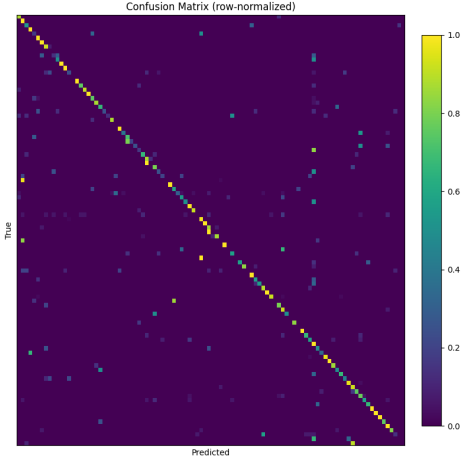



Confusion matrix for Ministral 8B per-class prediction distribution

# Annex Confusion Matrix

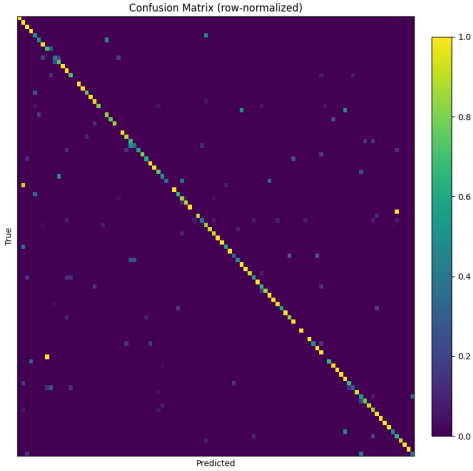



Confusion matrix for Mistral Small showing per-class prediction distribution

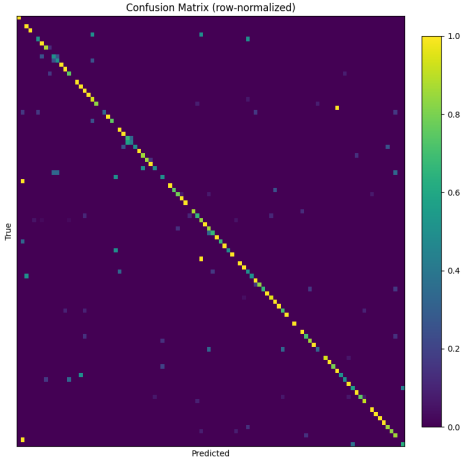



Confusion matrix for Mistral Medium showing per-class prediction distribution

# Annex Confusion Matrix

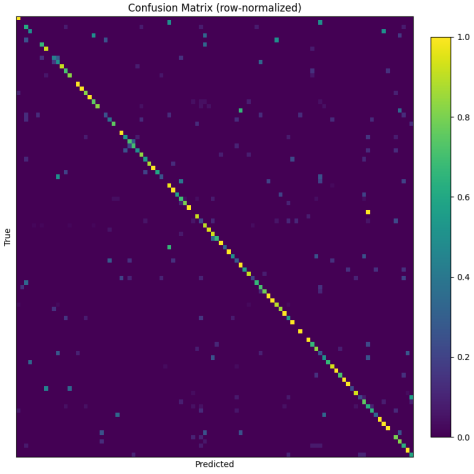



Confusion matrix for RAG baseline showing per-class prediction distribution

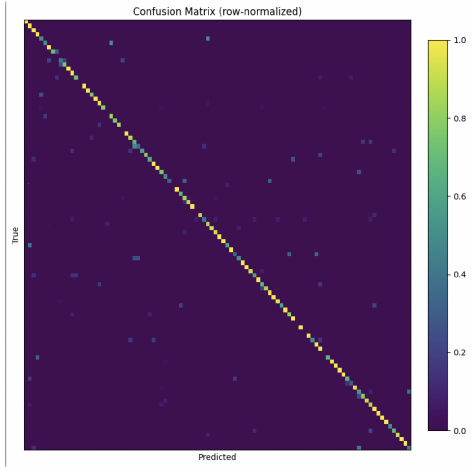



Confusion matrix for RAG with fine-tuned model showing per-class prediction distribution

# Annex Confusion Matrix




Confusion matrix for Ministral 8B with classification head showing per-class prediction distribution




Confusion matrix for classification head with LoRA showing per-class prediction distribution

# References I


- Chalkidis, Ilias et al. (2022). *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*. arXiv: 2110.00976 [cs.CL]. URL: https://arxiv.org/abs/2110.00976.
- Hu, Edward J. et al. (2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685. arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685.
- Lewis, Patrick et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *CoRR* abs/2005.11401. arXiv: 2005.11401. URL: https://arxiv.org/abs/2005.11401.
- Malkov, Yury A. and Dmitry A. Yashunin (2016). "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs". In: *CoRR* abs/1603.09320. arXiv: 1603.09320. URL: http://arxiv.org/abs/1603.09320.
- Mistral AI (2024). *Mistral API Documentation*. Online documentation. Accessed: 2025-08-11. URL: https://docs.mistral.ai.

# References II


Mistral AI team (2024). *Ministral-8B-Instruct-2410*. Hugging Face Model Card. Dense Transformer, 8B parameters, 128k context window, released under Mistral Research License. URL: https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.