

Model Card: Seny Legal-Lite



Model overview

Seny legal-lite is a specialized base model built to enable the evaluation of AI agents' behavior within the law. It is the foundation for a new generation of judge models that check whether an AI system's output complies with a specific rule, across nine languages. Seny legal-lite is purpose-built to evaluate, control and monitor AI agent behavior against specific legal provisions and policies at runtime, across languages and jurisdictions.

Seny legal-lite was created by further training an open base model on a multilingual legal corpus curated by Alinia's legal specialists. Training compute was from the Barcelona Supercomputing Center, under a EuroHPC AI Factory grant. Seny is not instruction-tuned: it is post-trained into judges, each specialized for one rule.

 9 languages	English, Spanish, Catalan, French, Dutch, German, Mandarin Chinese, Italian, Portuguese.
 151 ms	API end-to-end median latency of judges built on Seny is 12x to 17x faster than frontier models.
 Frontier-model accuracy	Judges built on Seny reach and often exceed the accuracy of frontier and guard models.

Developed by Alinia AI	Parent model google/gemma-4-E4B	Trained at Barcelona Supercomputing Center (BSC), MareNostrum 5	Release October 2026
----------------------------------	---	---	--------------------------------

Intended downstream commercial use

 Compliance-specialized judges	Seny legal-lite is the backbone from which specialized classification heads are able to check whether AI interactions involve creditworthiness assessment or jailbreak attempts (auditing; guardrails).
 Fully customizable	Seny legal-lite can be further fine-tuned for jurisdiction-specific or sector-specific compliance (e.g., financial services, insurance, healthcare), or for additional languages.
 Integration into compliance workflows	By using retrieval-augmented generation to ground the model's outputs.

Training data

Seny legal-lite was trained on 3 billion tokens of legal text in nine languages, in two layers: 24.94% curated law (statutes, regulations, court judgments and official gazettes from official government and judicial portals) and 60.07% legal-filtered web text. Primary and normative sources were curated with per-source license verdicts; web-derived material is explicitly tiered as commentary and drawn from ODC-BY-licensed FineWeb-2.

No synthetic data was generated for training. Alinia’s regulatory engineering team (legal human experts) curated legal data sources composing the training corpus.

Evaluation

We share here a selection of Alinia judges post-trained on Seny legal-lite. The judges match or beat frontier models on accuracy while answering in about 150 milliseconds end to end.

Post-trained judge benchmarks

Each judge was evaluated on our own dedicated, out-of-distribution datasets and compared with leading models using the same prompt and evaluation setup. We present the results below. All datasets contain at least English examples. Multilingual datasets are indicated with a globe icon next to the judge (green for 9 languages and gray for 3). Seny legal-lite’s column is highlighted in green and the best result for each judge is in bold.

	Seny Legal-Lite	Claude Opus 5	GPT 6 Astra	Gemini 3.5 Flash Lite	GPT 5.4 Nano	gpt-oss- safeguard	Qwen 3.5 9B
Legal advice 🌐	0.834	<u>0.842</u>	✓ 0.864	0.770	0.730	0.794	0.753
Creditworthiness	✓ 0.792	<u>0.772</u>	0.767	0.642	0.759	0.712	0.743
Investment advice 🌐	<u>0.946</u>	✓ 0.948	0.816	0.888	0.902	0.900	0.907
Security 🌐	<u>0.866</u>	0.844	✓ 0.872	0.709	0.644	0.678	0.723
Tax (global) 🌐	<u>0.910</u>	✓ 0.925	<u>0.910</u>	0.843	0.886	0.861	0.667
Insurance (US) 🌐	<u>0.922</u>	0.907	✓ 0.929	0.875	0.886	0.902	0.766
Mean across judges	✓ 0.878	<u>0.873</u>	0.860	0.788	0.801	0.808	0.760

Quality is measured by the best F1 score.

✓ **0.000** Best 0.000 Second best 🌐 Multilingual

Latency benchmarks

Latency evaluations with API overhead indicate that Seny legal-lite is twelve to seventeen times faster than GPT 6 Astra, Gemini 3.8 Flash and Claude Opus 5.

	Seny Legal-Lite	Claude Opus 5	GPT 6 Astra	Gemini 3.8 Flash	GPT 5.4 Nano	gpt-oss- safeguard
Latency lower is better	✓ 151 ms	2576 ms	1879 ms	2029 ms	<u>785 ms</u>	836 ms

Base model benchmarks

We compare Seny legal-lite and Seny legal-lite’s base model, Gemma-4-E4B, on three open-source benchmarks: LegalBench (legal knowledge), MMLU (general knowledge), and MMMLU (multi-lingual general knowledge) to verify that there was no catastrophic forgetting. The inclusion of additional legal-related training data resulted in an uplift of 1.0 on the MMLU tests while showing a small loss in performance in both LegalBench and MMMLU.

Limitations

 **Disclaimer**

Model users decide whether to use the judge for auditing purposes (i.e. monitoring and traceability of AI Agents) or as guardrails (i.e. recommend the flagging or blocking of user-AI interactions; content redaction). Any other usage will be considered unintended use, such as, but not limited to:

- Legal advice or legal determinations without qualified human oversight. Outputs are not a substitute for a lawyer.
- Seny legal-lite is not trained as a safety classifier; task-specific guards should be used instead (when available).
- General-purpose open-domain assistant use.
- High-stakes decisions about individuals (e.g., eligibility, employment, or law-enforcement contexts).

 **Known limitations**

Authoritative statements of current law without retrieval. The model’s knowledge is frozen at training time, and it may hallucinate citations or misstate provisions. However, the intended use of this model as a base to post-train classifiers makes this limitation not applicable in practice.

Performance outside the nine supported languages is untested.

Responsible use

 Human oversight	Use Seny legal-lite as part of a broader moderation pipeline with human oversight.
 Guardrail use cases	Make sure to test the fine-tuned Alinia judges built on Seny legal-lite before deploying them in production. It is strongly recommended to deploy the Judge in an auditing (logging) mode first, for human experts to supervise Judge performance before switching to guardrail mode. The latter has a direct impact on end-user experience, such as blocking messages in conversational interfaces between user and AI agent.