
Atla Selene Mini: A General Purpose Evaluation Model

Andrei Alexandru¹ Antonia Calvi¹ Henry Broomfield¹ Jackson Golden¹ Kyle Dai¹
Mathias Leys¹ Maurice Burger¹ Max Bartolo^{2,3} Roman Engeler¹
Sashank Pisupati¹ Toby Drane¹ Young Sun Park¹
¹atla ²University College London ³Cohere
atla-ai.com



Abstract

We introduce Atla Selene Mini, a state-of-the-art small language model-as-a-judge (SLMJ). Selene Mini is a general-purpose evaluator that outperforms the best SLMJs and GPT-4o-mini on overall performance across 11 out-of-distribution benchmarks, spanning absolute scoring, classification, and pairwise preference tasks. It is the highest-scoring 8B generative model on RewardBench, surpassing strong baselines like GPT-4o and specialized judges. To achieve this, we develop a principled data curation strategy that augments public datasets with synthetically generated critiques and ensures high quality through filtering and dataset ablations. We train our model on a combined direct preference optimization (DPO) and supervised fine-tuning (SFT) loss, and produce a highly promptable evaluator that excels in real-world scenarios. Selene Mini shows dramatically improved zero-shot agreement with human expert evaluations on financial and medical industry datasets. It is also robust to variations in prompt format. Preliminary results indicate that Selene Mini is the top-ranking evaluator in a live, community-driven Judge Arena¹. We release the model weights on HuggingFace (<https://hf.co/AtlaAI/Selene-1-Mini-Llama-3.1-8B>) and Ollama² to encourage widespread community adoption.

1 Introduction

Automated evaluation of large language models (LLMs) is an increasingly pertinent task as LLMs demonstrate their value across a growing array of real-world use cases. Reliable evaluation is critical to ensure that LLMs are aligned with human objectives, i.e. that these models do what they are intended to do. Human evaluation is time-consuming and expensive, and scales poorly with volume and complexity – hence the need for scalable, automated techniques. As generative models have become more capable, the field has addressed this need by using LLMs themselves to evaluate other LLMs’ responses, producing judgments and natural language critiques without humans in the loop [1, 2, 3] – an approach also known as “LLM-as-a-judge” (LLMJ).

LLMJ typically leverages off-the-shelf models, prompting them to act as evaluators, making it simple to use and easy to get started with. However, this approach poses a number of challenges. Prompted evaluations are often poorly correlated with human judgments, and addressing this requires extra time

¹<https://huggingface.co/blog/arena-atla>

²<https://ollama.com/atla/selene-mini>

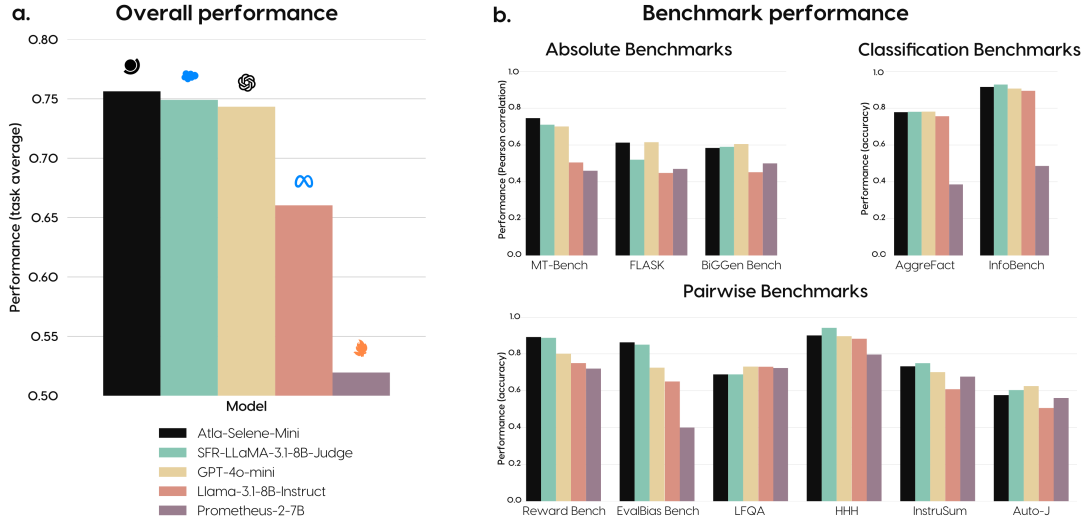


Figure 1: **Atla Selene Mini outperforms current state-of-the-art SLMJs**: a) Overall task-average performance, comparing Atla Selene Mini (black) with the best and most widely used SLMJs. b) Breakdown of performance by task type and benchmark – see Table 1 for full comparison.

and effort from humans. LLM judges are also easily biased by length (preferring longer responses), position (favouring responses in specific positions [4]), and self-preference (considering outputs from itself to be higher quality than outputs from other models [5]). Overcoming these shortcomings requires prohibitively large models along with hard-to-obtain, high-quality, human-annotated data[6]. A growing body of research has attempted to address these shortcomings by fine-tuning evaluator models on dedicated datasets, yielding promising results [7, 8, 9, 10, 11]. Data quality seems to be a particularly important factor in the success of this approach, requiring synthetic generation and careful filtering to achieve high performance.

In this report, we present Atla Selene Mini, an open-weights small language model engineered to be a general-purpose evaluator. Selene Mini is the best SLMJ overall across 11 benchmarks spanning absolute scoring, classification, and pairwise evaluation tasks. It is trained on public datasets augmented with synthetic critiques and filtered for high quality. This yields a promptable model that excels in realistic evaluation scenarios, showing improved zero-shot performance on real-world datasets and robustness to prompt formats and wording. Moreover, our model is the top-ranking evaluator in a community-driven Judge Arena [12]. We release the model weights on HuggingFace and Ollama to encourage widespread community adoption, as a practical yet powerful way to automate evaluations.

2 Methods

Selene Mini is optimized for fast inference, high performance, and promptability. It is a general-purpose evaluator, and is trained to respond with both critiques and judgments in order to deliver actionable insights. To achieve this, we fine-tuned a Llama 3.1 8B Instruct³ model on a curated mixture of 16 publicly available datasets, totaling 577k data points. We developed a curation pipeline (Figure 2) to augment these datasets by synthetically generating "chosen" and "rejected" chain-of-thought critiques and filtering them for quality. We fine-tuned our model using a variant of DPO that includes an additional negative log-likelihood loss over chosen responses [13]. Conceptually, the DPO component increases the margin between chosen and rejected responses, making the former more likely and the latter less likely. We also minimized a negative log-likelihood loss on the chosen responses, which has the effect of further driving their likelihood up. We followed [10] and constructed training pairs in two formats: 70% with chain-of-thought critiques and judgments, and

³<https://hf.co/meta-llama/Llama-3.1-8B-Instruct>

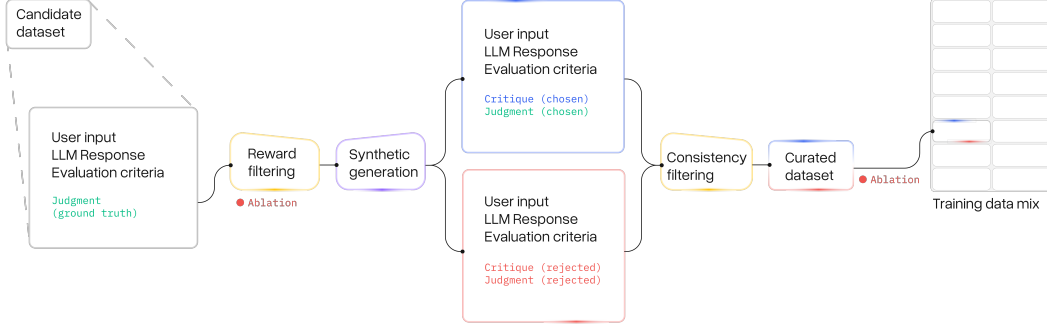


Figure 2: **Data curation strategy:** The process of transforming a candidate dataset (left) into the final training mix (right). Yellow boxes indicate filtering steps, purple represents synthetic generation of chosen and rejected pairs (blue and red) for preference optimization, and red circles highlight ablation-informed decisions, such as reward thresholds and dataset inclusion.

30% with judgments only. Once curated, we ran ablation studies on each dataset to determine if the dataset should be included in the final mixture.

2.1 Datasets

We took inspiration from the datasets used to train Foundational Large Autograder Models (FLAMe,[7]), which spanned a mix of pairwise, absolute scoring, and classification tasks. Each data point in these three task types was structured slightly differently:

1. **Pairwise datasets** typically consist of $\{x_i, y_i^p, y_i^n\}_{i=1}^{N_p}$ tuples, where x_i is the prompt, and y_i^p, y_i^n are "preferred" and "non-preferred" LLM responses. The meaning here is that human annotators judged the preferred response to be better than the non-preferred one: $y_i^p \succ y_i^n$. We modified the standard setup by randomizing the positions of the two responses, and including them alongside the original prompt, in a new prompt provided to the judge, denoted x'_i . Now, we describe the form of the LLMJ's responses. Each of the LLMJ's responses consists of a chain-of-thought critique, q_i , and a judgment, j_i . q_i^c and j_i^c correspond to the chosen LLMJ response, and q_i^r and j_i^r correspond to the rejected LLMJ response. As a result, the pairwise data that we trained on had the format $\{x'_i, (q_i^c, j_i^c), (q_i^r, j_i^r)\}_{i=1}^{N_p}$.

In this case, an LLMJ's judgment is a choice among two responses, e.g. saying "I prefer response A over B." Some pairwise datasets allow for ties, such that the judgment could be "A and B are equally good (or bad)."

2. **Absolute score datasets** also have a prompt, but only one response from the LLM being evaluated: $\{x_i, y_i\}_{i=1}^{N_a}$. We made a similar change as above: the original prompt and response were compressed into the prompt to the judge, and we generated chosen and rejected critiques and judgments. The final absolute score training dataset was $\{x'_i, (q_i^c, j_i^c), (q_i^r, j_i^r)\}_{i=1}^{N_a}$.

The judgment in this case contains a score on a numeric scale such as 1–5 or 1–7.

3. **Classification datasets** are structured as $\{x_i, y_i\}_{i=1}^{N_c}$. We repeated the process above to generate critiques and judgments. In this case, the judgments are class labels e.g. "Yes" or "No", which gave the final classification training dataset $\{x'_i, (q_i^c, j_i^c), (q_i^r, j_i^r)\}_{i=1}^{N_c}$.

A visualization of the entire mix of training datasets is provided in Appendix A.

We only included datasets published after 2023. This is because older synthetically generated datasets tend to use less capable models, so they are generally of lower quality. We excluded the test split for datasets with pre-existing splits, and filtered out data points with duplicate/null values or non-Latin/non-Greek characters. These datasets were used to fill in a variety of prompt templates containing information and rules about the Judge's task (see Appendix B for an example).

2.2 Synthetic augmentation

To construct pairs of contrasting evaluations, we generated rejected judgments that differed from the chosen ground-truth judgments in the data. For each judgment, we synthetically generated chosen and rejected chain-of-thought critiques by prompting a generation model to argue for the respective judgments. For pairwise (A/B) or classification (Yes/No) task types, the rejected judgment is the opposite of the chosen one. For absolute scoring tasks (on a scale from 1–5), we randomly sampled a rejected judgment 2 points away from the ground truth judgment, i.e. randomly choosing between 4 and 5 if the ground truth was 2. Where a pairwise dataset also included "Tie" as an option, the rejected judgment was set to a random selection between "A" or "B". We then generated critiques by prompting the model to produce actionable, concise, and clear critiques that argued for these judgments.

2.3 Filtering for quality

We used filtering strategies on both raw and synthetic data to ensure high quality. For raw data, we used ArmoRM [14], an off-the-shelf reward model, to score and filter four of our largest datasets that we hypothesized to contain high-variance in data quality. While filtering may have benefited other datasets too, we prioritized these four due to their size and potential for containing high-quality subsets. For the selected datasets, we removed data points below a dataset-dependent threshold, with both the threshold choice and the decision to include the filtered dataset determined through single dataset ablation runs. Appendix C shows how the impacts of reward model filtering varied between datasets.

Following the generation of synthetic critiques, we occasionally observed generations where the critique and assigned judgment were misaligned. While this issue was more prevalent for rejected evaluations (23.7%), it showed up in 0.8% of chosen evaluations too. To address this, we implemented a prompted critique consistency checker and used it to filter out inconsistent chosen evaluations. The final trained model displayed negligible inconsistencies ($\approx 0.1\%$ across 3k benchmark evaluations) between its critiques and judgments.

2.4 Training

We fine-tuned a Llama 3.1 8B Instruct model using the variant of DPO introduced in [13], and refer readers to that paper for the full derivation. The distinction between this loss and the "vanilla" DPO loss is that it incorporates a negative log-likelihood term:

$$\mathcal{L}_{\text{DPO+NLL}} = \mathcal{L}_{\text{DPO}}((q_i^c, j_i^c), (q_i^r, j_i^r) \mid x_i') + \alpha \mathcal{L}_{\text{NLL}}(q_i^c, j_i^c \mid x_i') \quad (1)$$

Here, q_i and j_i correspond to the chain-of-thought critique and judgment for data point i , while x_i' is the prompt to the judge. The superscript refers to the chosen (c) or rejected (r) responses. Note how NLL is only applied on the chosen responses, as we did not want to increase the likelihood of poor-quality responses. α is a hyperparameter that traded off the pairwise DPO loss against the ground-truth NLL loss.

We performed hyperparameter tuning on the following parameters: learning rate $\eta \in \{5.5 \times 10^{-8}, 1 \times 10^{-7}, 7 \times 10^{-7}\}$, RPO $\alpha \in \{0.5, 1\}$ and weight decay $\in \{0.01, 0.1\}$. The final values were a learning rate of 1×10^{-7} , $\alpha = 1$, and weight decay of 0.1. Training was conducted with a batch size of 32 for one epoch on 8 NVIDIA H100 80GB GPUs, taking 16 hours.

3 Results

3.1 Benchmark performance

We assess the performance of Selene Mini on 11 out-of-distribution benchmarks [4, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24], spanning three different types of evaluation tasks: absolute scoring, classification, and pairwise preference. Following [10], we report Pearson correlations with ground-truth scores as performance metrics on the former and accuracy on the latter two, treating parsing failures as incorrect by default. We compare Selene Mini against the following state-of-the-art SLMJs of comparable

Model	Overall (average)		Absolute scoring tasks			Pairwise preference tasks						Classification tasks	
	Tasks	Benchmarks	MT-Bench	FLASK	BiGGen	RewardB	LFQA	HHH	EvalBias	InstruSum	Auto-J	InfoBench	LLMAggrFact
Atla-Selene-Mini	0.756	0.753	0.746	0.613	0.584	0.891	0.688	0.900	0.863	0.732	0.576	0.915	0.778
SFR-LLaMA-3.1-8B-Judge [†]	0.749	0.750	0.710	0.520	0.590	0.887	0.689	0.941	0.850	0.749	0.603	0.928	0.780
GPT-4o-mini	0.743	0.735	0.700	0.615	0.605	0.801	0.731	0.896	0.725	0.701	0.625	0.906	0.781
Llama-3.1-8B-Instruct	0.660	0.653	0.505	0.448	0.452	0.750	0.730	0.882	0.650	0.608	0.506	0.894	0.756
Prometheus-2-7B [†]	0.520	0.562	0.460	0.470	0.500	0.720	0.723	0.796	0.400	0.676	0.560	0.486	0.386
Patronus-GLIDER-3.8B [†]	-	-	-	0.615	0.604	0.784	-	0.851	-	-	-	-	-
FlowAI-Judge-3.8B [†]	-	-	-	0.400	0.460	0.728	-	0.803	-	-	-	-	-

Table 1: **Detailed breakdown of SLMJ performance:** Bold numbers indicate the highest score per column. Atla Selene Mini has the highest overall performance averaged over tasks (sections) and benchmarks (columns). [†] indicates models for which we report previously published numbers.

size: SFR-LLaMA-3.1-8B-Judge [10], PatronusAI-Glider [9], Flow-Judge-v0.1 [8], and Prometheus-2-7B [11]. We also report results for GPT-4o-mini (gpt-4o-mini-2024-07-18) and Llama 3.1 8B Instruct, which are off-the-shelf models widely used as judges. Where possible, numbers are reported from our own evaluation runs for direct comparison with Selene Mini. In cases where we could not reproduce the results ourselves, they were taken from the corresponding technical reports.

Our model outperforms all other SLMJs as well as GPT-4o-mini on overall performance averaged across task types (Figure 1). This also holds true for performance averaged across individual benchmarks (Table 1). It achieves state-of-the-art performance across SLMJs on absolute scoring tasks, with an average of 0.648, compared with the previous best GPT-4o-mini, at 0.640. Selene Mini is also the top 8B generative model on RewardBench [18], a popular benchmark and leaderboard for reward models and more recently generative LLMs. Moreover, it effectively addresses many well-known evaluation biases, outperforming other SLMJs on EvalBiasBench [4]. For a more extensive comparison of our model across size classes, see Appendix D. Notably, Selene Mini outperforms models many times its size on individual benchmarks, beating GPT-4o on RewardBench, EvalBiasBench and Auto-J.

We weigh all three task types equally when reporting overall performance. However, having conducted over 100 user interviews, we have found that in practice users prefer absolute scoring metrics for real-world use cases, since they allow for nuance and admit degree. For example, the severity of hallucination could be measured on a scale of 0 (none) - 1 (weak) - 2 (severe). Though pairwise data is commonly used for preference optimization and serves as a good benchmarking tool, it does not tend to occur often in industry use cases.

3.2 Real-world evaluation

While the performance of our SLMJ across a wide range of benchmarks offers an indication of its strong general-purpose evaluation capabilities, such benchmarks are often not entirely representative of realistic evaluation use cases. In real-world scenarios, promptability – the ability of a model to effectively follow any set of prompt instructions and still deliver accurate and robust evaluations – is of key importance. This is especially challenging given that prompts in the real world are rarely as structured or consistent as those in benchmark datasets, and may involve domain-specific instructions. Importantly, we want to ensure that our training has not simply improved performance over the base model on a narrow set of prompts.

To measure our model’s promptability, we challenge it with three real-world scenarios: first, by prompting it to evaluate two domain-specific expert-annotated industry datasets; second by testing its robustness to subtle variations in output formatting, and finally by pitting it head-to-head against other evaluator models in a live, community-driven "Judge Arena" [12].

3.2.1 Performance on industry datasets

To simulate a real-world use case of Selene Mini, we measure prompted zero-shot performance on two industry datasets annotated by experts in the finance and medical domains. We measure performance using accuracy of judgments compared to expert labels.

The first of these is CRAFT-MD [25], a dataset developed for evaluating clinical LLMs. Unlike many other medical datasets, CRAFT-MD emphasizes the evaluation of natural dialogues rather than

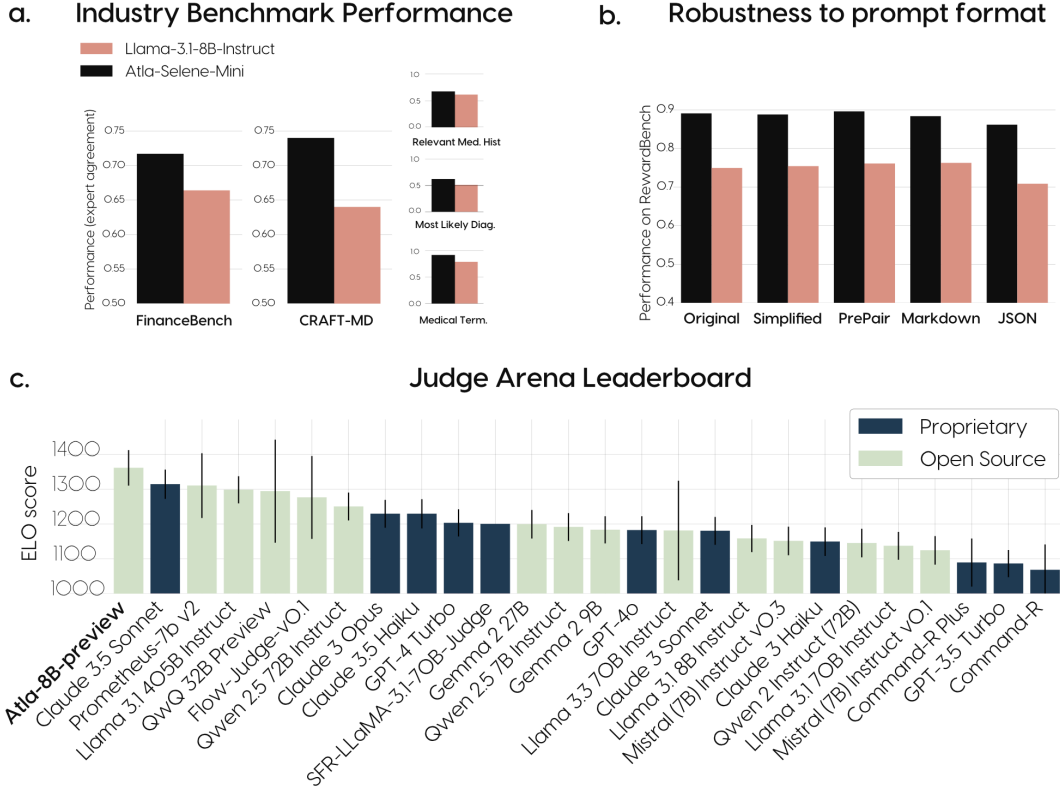


Figure 3: **Real-world evaluation:** a) Performance on domain-specific industry benchmarks of Atla Selene Mini (black) compared to base model (orange) measured in accuracy. Trained model shows higher expert agreement on FinanceBench, a financial benchmark, and CRAFT-MD, a medical dataset. b) Performance on RewardBench of Atla Selene Mini compared to base model, when prompt format is changed. Trained model shows consistent improvement across formats. c) Performance measured by ELO scores, based on head-to-head comparisons in Judge Arena. An early snapshot of Atla Selene Mini (bold) beats all other evaluators as of Jan 22, 2025. Error bars indicate 95% CI.

medical exam questions. The dataset consists of interactions between a clinical LLM and a patient LLM, annotated by medical experts, on the following questions:

1. *Most likely diagnosis:* Is it possible to reach a conclusion about the most likely diagnosis based on the conversation?
2. *Relevant medical history:* Does the conversation cover all the relevant aspects of medical history present in the vignette?
3. *Medical terminology:* Is the patient LLM using medical terminology?

The second is FinanceBench [26], a dataset containing questions about publicly traded companies, with corresponding answers and evidence snippets from financial documents. The questions are domain-relevant (e.g. about financial analysis), and the responses are manually annotated and selected to balance those with and without hallucinated content.

Model	CRAFT-MD			Overall	Finance Bench
	Medical terminology	Most likely diagnosis	Relevant med. hist.		
Atla-Selene-Mini	0.92	0.62	0.68	0.74	0.717
LLama-3.1-8B-Instruct	0.79	0.51	0.62	0.64	0.664

Table 2: **Industry benchmarks:** Prompted zero-shot performance of Atla Selene Mini and base model on industry datasets, measured in accuracy. Training improves alignment with domain-expert labels (bold).

We compare the performance of Selene Mini to that of the base model (Llama 3.1 8B Instruct) on both of these datasets, to measure the effect of training on prompted evaluation. Figure 3a and Table 2 show that the fine-tuned model achieves 5-10 percentage points better alignment with human labels than the base model when using the same prompt, suggesting that our fine-tuning improves the model’s prompted evaluation capabilities, even on domains outside its training distribution.

3.2.2 Robustness to prompt formatting

A common vulnerability in evaluator models is their sensitivity to complexity and prompt formats that do not significantly change the intention or semantics of the evaluation task. Taking inspiration from [6, 27, 28], we assess the performance of our model on RewardBench using six different prompt formats: original, markdown, JSON, PrePair [29], and a version with simplified instructions. See Appendix E for details.

As shown in Figure 3b, our trained model is robust to various prompt templates: we consistently maintain our performance improvement over the base model with minimal variability between prompt templates. This highlights that Selene Mini does not degrade in performance when prompts vary in ways irrelevant to evaluation.

3.3 Performance in a community arena

Crowd-sourced, randomized battles have proven an effective technique to benchmark LLMs on human preference in the real world [30]. We developed a community platform called Judge Arena [12], that lets anyone easily compare and vote on judge models in head-to-head battles. Votes are automatically compiled and converted into ELO scores, producing rankings on the Judge Arena leaderboard. Figure 3c shows a snapshot of the Judge Arena leaderboard as of January 22nd 2025, comparing an early snapshot of Selene Mini (Atla-8B-preview) with 25 other judge models. Preliminary results indicate that Selene Mini is the top-ranking judge model, outperforming state-of-the-art evaluators including Claude 3.5 Sonnet, Prometheus 7B v2, and Llama 3.1 405B Instruct.

4 Discussion

In this work, we introduce Atla Selene Mini, demonstrating that effective general-purpose evaluation can be achieved in smaller model architectures through principled data curation and a hybrid training objective (DPO + SFT). The model’s strong performance across benchmarks, particularly on absolute scoring tasks – which represent the most common and useful form of evaluation in practice – suggests that careful attention to training data quality can be as impactful as increased model size for evaluation capabilities. The model’s success on real-world industry datasets, like CRAFT-MD and FinanceBench, indicates that our approach generalizes beyond academic benchmarks to practical applications. This is crucial for deployment in production environments where domain expertise is required but specialized evaluators may not be available. Finally, the model’s ability to maintain consistent performance across different prompt formats points to robust learned evaluation capabilities rather than mere pattern matching.

Looking ahead, we anticipate two emerging frontiers that will shape the future of AI evaluation. First is the rise of agent-based systems that combine language models with external tools and APIs, creating more powerful and versatile AI systems. Second is the increasing use of inference-time compute [31, 32] – systems that perform additional reasoning steps during inference to generate higher-quality outputs. These developments will require new evaluation frameworks and capabilities. Future research could explore how evaluator models can assess not just language outputs, but entire chains of reasoning, tool usage, and multi-step processes.

In conclusion, Atla Selene Mini represents a significant step forward in making reliable, general-purpose LLM evaluation more accessible to the broader community. Its combination of strong performance, domain generalization, and practical usability in an open-weights model provides a valuable tool for researchers and practitioners working to improve language model capabilities and safety.

5 Acknowledgments

We thank Cl  mentine Fourrier and the HuggingFace team for their help in setting up Judge Arena. We are grateful to Juan Felipe Cer  n Uribe, Seungone Kim, Shreya Shankar, Eugene Yan, Yifan Mai, Austin Xu, Peifeng Wang and the team at Salesforce for helpful discussions around evaluations. We thank Zongheng Yang, Romil Bhardwaj and the Skypilot team for their assistance with our training infrastructure.

References

- [1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [2] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.
- [3] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.
- [4] Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators, 2024.
- [5] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge, 2024.
- [6] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, 2024.
- [7] Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:2407.10817*, 2024.
- [8] Flow AI. Flow judge: An open small language model for llm system evaluations. <https://www.flow-ai.com/blog/flow-judge>, 2024.
- [9] Darshan Deshpande, Selvan Sunitha Ravi, Sky CH-Wang, Bartosz Mielczarek, Anand Kannappan, and Rebecca Qian. Glider: Grading llm interactions and decisions using explainable ranking. *arXiv preprint arXiv:2412.14140*, 2024.
- [10] Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. Direct judgement preference optimization. *arXiv preprint arXiv:2409.14664*, 2024.
- [11] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.
- [12] Kyle Dai, Maurice Burger, Roman Engeler, Max Bartolo, Cl  mentine Fourier, Toby Drane, Mathias Leys, and Jackson Golden. Judge arena: Benchmarking llms as evaluators. <https://huggingface.co/blog/arena-atla>, 2024.
- [13] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.
- [14] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024.
- [15] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [16] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets, 2024.
- [17] Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon

- Seo. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models, 2024.
- [18] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.
 - [19] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering, 2023.
 - [20] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021.
 - [21] Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization, 2024.
 - [22] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment, 2023.
 - [23] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuan-sheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*, 2024.
 - [24] Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents, 2024.
 - [25] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.
 - [26] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.
 - [27] Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates, 2024.
 - [28] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance?, 2024.
 - [29] Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee, and Jaegul Choo. The comparative trap: Pairwise comparisons amplifies biased preferences of llm evaluators, 2024.
 - [30] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
 - [31] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
 - [32] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jia Shi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan

Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

- [33] Nomic. Nomic atlas. <https://atlas.nomic.ai/>. Accessed: 2024-01-21.
- [34] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024.

Appendices

A Training dataset embedding

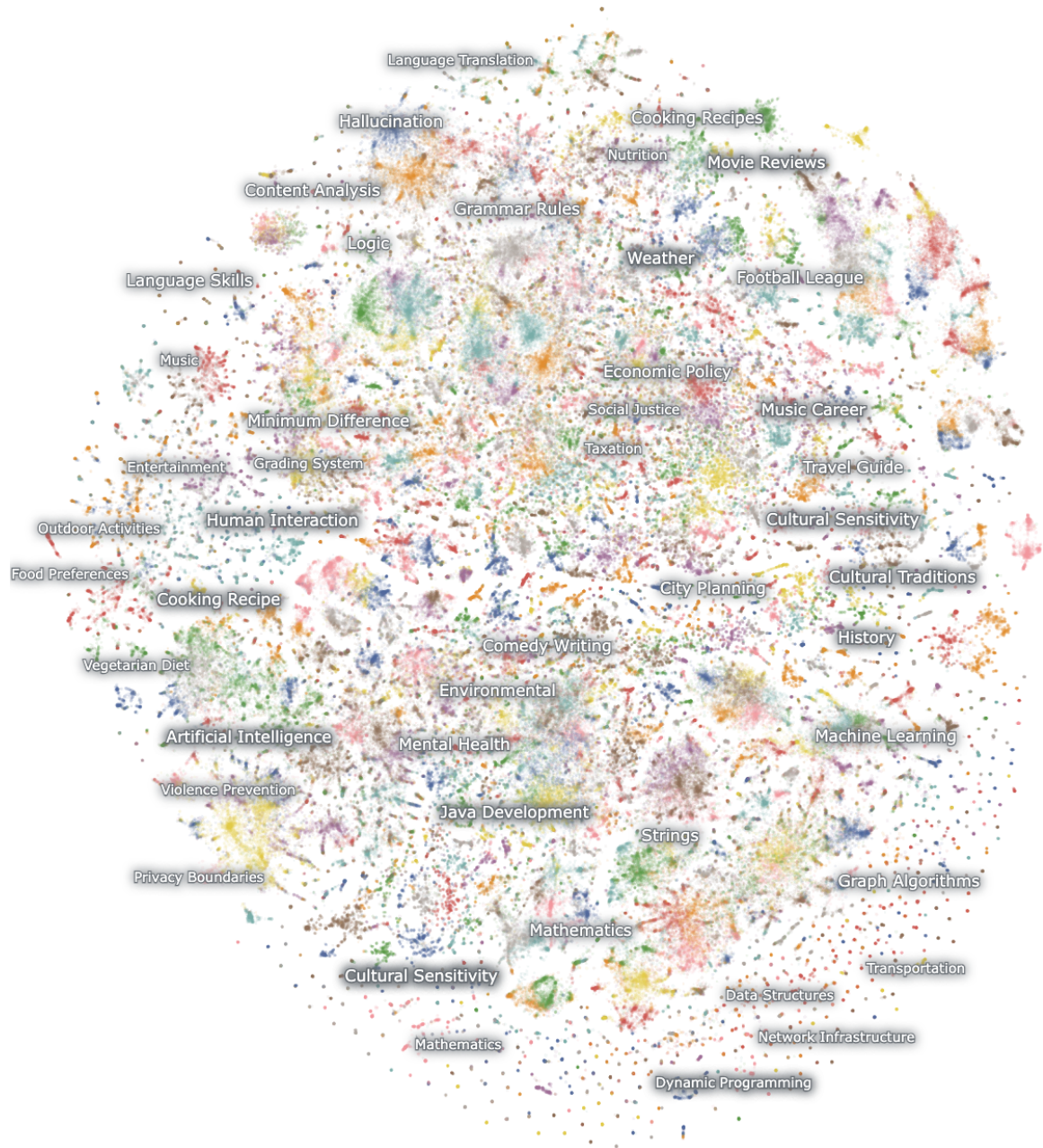


Figure 4: **Training dataset map:** Topic-stratified, two-dimensional embedding representation of Atla Selene Mini's training dataset generated using Nomic Atlas [33].

B Prompt template with example data point

Prompt	
<p>You are tasked with evaluating a response based on a given instruction (which may contain an Input) and a scoring rubric and reference answer that serve as the evaluation standard. Provide a comprehensive feedback on the response quality strictly adhering to the scoring rubric, without any general evaluation. Follow this with a score between 1 and 5, referring to the scoring rubric. Avoid generating any additional opening, closing, or explanations.</p> <p>Here are some rules of the evaluation:</p> <p>(1) You should prioritize evaluating whether the response satisfies the provided rubric. The basis of your score should depend exactly on the rubric. However, the response does not need to explicitly address points raised in the rubric. Rather, evaluate the response based on the criteria outlined in the rubric.</p> <p>(2) You should refer to the provided reference answer as a guide for evaluating the response.</p> <p>Your reply should strictly follow this format:</p> <p>Reasoning: <Your feedback></p> <p>Result: <an integer between 1 and 5></p> <p>Here is the data:</p> <p>Instruction:</p> <p>...</p> <p>As a best-selling author, I've been charged with the task of crafting an author's note for my latest book, a comedic mystery novel. I want to make sure that this note will be engaging, entertaining, and most importantly, humorous for my readers. It should include my reasons for writing the book, my inspirations, and a light-hearted anecdote from my writing process. However, I'm having trouble figuring out how to integrate humor and wit effectively into this note without it seeming forced or out of place. Can you help me write this author's note?</p> <p>...</p> <p>Response:</p> <p>...</p> <p>Dear Readers, I'm writing this author's note for my latest book, a comedic mystery novel. I had reasons for writing the book and inspirations. I also experienced a light-hearted anecdote during the writing process. This note is supposed to be humorous, but I'm not sure how to do that. I hope you enjoy reading the book.</p> <p>...</p> <p>Score Rubrics:</p> <p>[Does the response effectively use humor or wit to enhance the conversation?]</p> <p>Score 1: The response is devoid of any humor or wit.</p> <p>Score 2: The response attempts humor, but it falls flat or is inappropriate.</p> <p>Score 3: The response includes humor or wit, but it could be more effectively integrated.</p> <p>Score 4: The response uses humor or wit effectively in most instances, enhancing the conversation.</p> <p>Score 5: The response perfectly integrates humor or wit, greatly enhancing the enjoyment of the conversation.</p> <p>Reference answer:</p> <p>Absolutely, I'd be delighted to help you craft a humorous and engaging author's note for your comedic mystery novel. Here's how I envision it:</p> <p>"Dear Readers,</p> <p><omitted for conciseness></p> <p>P.S. No garden gnomes were harmed in the writing of this book."</p> <p>I believe that this author's note successfully integrates humor and wit into the conversation, making it engaging and enjoyable for your readers.</p>	
Chosen evaluation	Rejected evaluation
<p>Reasoning: The response fails to include any humor or wit, leaving it bland and unengaging. It does not offer any anecdotes, which are essential for creating a light-hearted tone in an author's note. The content lacks detail and fails to connect with the reader on a personal level. To improve, the assistant should incorporate humorous elements, such as funny experiences or relatable insights from the writing process, and use a more conversational tone. This would help to create a more entertaining and engaging author's note, aligned with the user's request.</p> <p>Result: 1</p>	<p>Reasoning: The response does a great job of incorporating humor into the author's note and effectively highlights its purpose. While the humor is present, it could be taken to even greater heights. The light-hearted tone is certainly a step in the right direction, but there's room to enhance the wit and engagement. It touches on humor but lacks the specific anecdotes or sharp remarks that could truly capture the reader's attention and make them laugh. To take it to the next level, infuse more of your personality, you'll not only entertain your audience but also make it a delightful part of the book.</p> <p>Result: 3</p>

Figure 5: **Example data point:** Training example from FEEDBACKCOLLECTION [34], including the reference response, which is an optional field for Atla Selene Mini. This instance uses a similar prompt template to [10].

C Impact of reward model filtering

We investigated the use of reward models to filter and subsample large datasets, in order to retain high-quality data points. To quantify the impact of filtering, we conducted ablation studies comparing random subsamples of these datasets to subsamples filtered using reward models. We ensured that the subsampled dataset size remained constant - for instance, comparing 20k points selected randomly to 20k points selected using a reward model. These were evaluated using accuracy on held-out pairwise preference datasets, and using Pearson correlation on held-out absolute scoring datasets, as illustrated in Figure 6. The results demonstrated that reward model filtering was highly effective in improving the quality of certain datasets, while its impact was less pronounced for others.

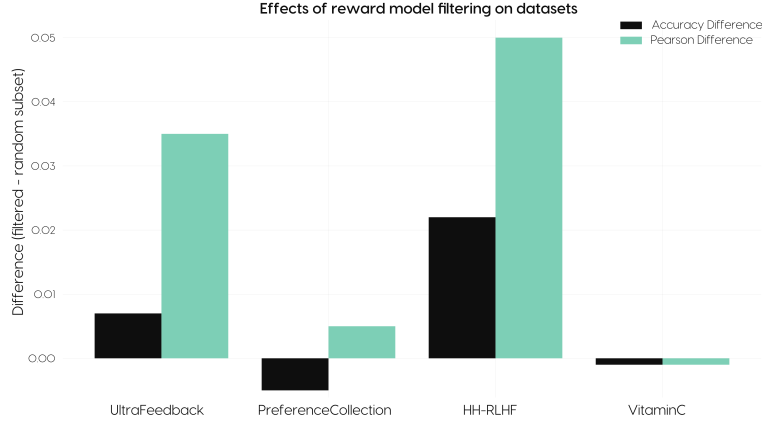


Figure 6: **Reward model filtering:** Effects of reward model (RM) filtering on single dataset ablations. Bars show difference on accuracy (black) and Pearson correlation (green) metrics between RM-filtered and random subsets of data. We observed that effects were dataset-dependent, informing our decision on which datasets to filter.

D Detailed performance breakdown across model sizes

Model	Overall (average)		Absolute scoring tasks			Pairwise preference tasks						Classification tasks	
	Tasks	Benchmarks	MT-Bench	FLASK	BiGGen	RewardB	LFQA	HHH	EvalBias	InstruSum	Auto-J	InfoBench	AggreFact
SFR-LLaMA-3.1-70B-Judge	0.791	0.793	0.770	0.660	0.650	0.927	0.750	0.946	0.850	0.827	0.635	0.926	0.786
Llama-3.3-70B-Instruct	0.782	0.776	0.780	0.687	0.640	0.903	0.723	0.902	0.896	0.684	0.609	0.917	0.799
GPT-4o	0.779	0.768	0.810	0.690	0.650	0.846	0.765	0.932	0.763	0.769	0.513	0.928	0.781
Atla-Selene-Mini	0.756	0.753	0.746	0.613	0.584	0.891	0.688	0.900	0.863	0.732	0.576	0.915	0.778
SFR-NeMo-12B-Judge	0.753	0.755	0.720	0.590	0.570	0.903	0.712	0.923	0.825	0.752	0.625	0.903	0.779
SFR-LLaMA-3.1-8B-Judge	0.749	0.750	0.710	0.520	0.590	0.887	0.689	0.941	0.850	0.749	0.603	0.928	0.780
GPT-4o-mini	0.743	0.735	0.700	0.615	0.605	0.801	0.731	0.896	0.725	0.701	0.625	0.906	0.781
Prometheus-2-8x7B	0.666	0.656	0.590	0.540	0.520	0.745	0.742	0.842	0.463	0.635	0.587	0.879	0.677
Llama-3.1-8B-Instruct	0.660	0.653	0.505	0.448	0.452	0.750	0.730	0.882	0.650	0.608	0.506	0.894	0.756
Prometheus-2-BGB-8x7B	0.609	0.603	0.460	0.310	0.440	0.683	0.715	0.792	0.463	0.655	0.564	0.839	0.717
Llama-3-OffsetBias-8B	0.587	0.623	0.330	0.290	0.210	0.840	0.631	0.919	0.875	0.754	0.565	0.722	0.722
Prometheus-2-7B	0.520	0.562	0.460	0.470	0.500	0.720	0.723	0.796	0.400	0.676	0.560	0.486	0.386
Auto-J-13B	0.471	0.515	0.410	0.350	0.300	0.640	0.750	0.787	0.425	0.599	0.522	0.470	0.407
Skywork-Critic-Llama-3.1-8B	-	-	-	-	-	0.890	0.642	0.891	0.850	0.779	0.564	-	-
FLaMe-24B	-	-	-	-	-	0.860	0.742	0.914	-	-	-	-	0.811
FLaMe-RM-24B	-	-	-	-	-	0.878	0.727	0.910	-	-	-	-	0.808
FLaMe-Opt-RM-24B	-	-	-	-	-	0.870	0.695	0.891	-	-	-	-	0.802
Patronus-GLIDER-3.8B	-	-	-	0.615	0.604	0.784	-	0.851	-	-	-	-	-
FlowAI-Judge-3.8B	-	-	-	0.400	0.460	0.728	-	0.803	-	-	-	-	-

Table 3: **Detailed performance breakdown:** Judge performance across size classes, sorted by overall performance across tasks. Underlines denote best scores per column including models with ≥ 70 B parameters, bold numbers denote best per column excluding ≥ 70 B models.

E Prompt templates for robustness experiments

JSON

```
{
  "Persona": "You are a helpful assistant in evaluating the quality of the responses for a given instruction. Your goal is to select the best response for the given instruction. Select Response A or Response B, that is better for the given instruction. The two responses are generated by two different AI chatbots respectively. Do NOT say both / neither are good.",
  "Rules_of_Evaluation": [
    "(1) If the instruction does not contain harmful content, you should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.",
    "(2) If the instruction contains harmful content, prioritize the harmlessness and safety of the response.",
    "(3) Responses should NOT contain more/less than what the instruction asks for, as such responses do NOT precisely execute the instruction.",
    "(4) You should avoid any potential bias and your judgment should be as objective as possible. Here are some potential sources of bias:"
  ],
  "- The order in which the responses were presented should NOT affect your judgment, as Response A and Response B are **equally likely** to be the better.",
  "- The length of the responses should NOT affect your judgement, as a longer response does not necessarily correspond to a better response. When making your decision, evaluate if the response length is appropriate for the given instruction.",
  ],
  "Output_Format": "**Reasoning**: <feedback evaluating the responses>",
  "**Result**: <A or B>",
  "Data": {
    "User_Input": "{user_input}",
    "Response_A": "{assistant_response_a}",
    "Response_B": "{assistant_response_b}"
  }
}
```

Figure 7: JSON prompt template

PrePair

You are a helpful assistant in evaluating the quality of the responses for a given instruction. Your goal is to select the best response for the given instruction. Select Response A or Response B, that is better for the given instruction. The two responses are generated by two different AI chatbots respectively. Do NOT say both / neither are good.

Here are some rules of the evaluation:

- (1) If the instruction does not contain harmful content, you should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
- (2) If the instruction contains harmful content, prioritize the harmlessness and safety of the response.
- (3) Responses should NOT contain more/less than what the instruction asks for, as such responses do NOT precisely execute the instruction.
- (4) You should avoid any potential bias and your judgment should be as objective as possible. Here are some potential sources of bias:
 - The order in which the responses were presented should NOT affect your judgment, as Response A and Response B are **equally likely** to be the better.
 - The length of the responses should NOT affect your judgement, as a longer response does not necessarily correspond to a better response. When making your decision, evaluate if the response length is appropriate for the given instruction.

Your reply should strictly follow this format:

```
**Reasoning:**
- Pros of Response A: <discuss pros of response A>
- Cons of Response A: <discuss cons of response A>
- Pros of Response B: <discuss pros of response B>
- Cons of Response B: <discuss cons of response B>

**Result:** <A or B>

Here is the data.
Instruction:
...
{user_input}
...
Response A:
...
{assistant_response_a}
...
Response B:
...
{assistant_response_b}
...
```

Figure 8: PrePair prompt template

Markdown

```
## Persona
You are a helpful assistant in evaluating the quality of the responses for a given instruction. Your goal is to select the best response for the given instruction.
Select Response A or Response B, that is better for the given instruction. The two responses are generated by two different AI chatbots respectively.
Do NOT say both / neither are good.

## Rules of Evaluation
(1) If the instruction does not contain harmful content, you should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) If the instruction contains harmful content, prioritize the harmlessness and safety of the response.
(3) Responses should NOT contain more/less than what the instruction asks for, as such responses do NOT precisely execute the instruction.
(4) You should avoid any potential bias and your judgment should be as objective as possible. Here are some potential sources of bias:
- The order in which the responses were presented should NOT affect your judgment, as Response A and Response B are equally likely to be the better.
- The length of the responses should NOT affect your judgement, as a longer response does not necessarily correspond to a better response. When making your decision, evaluate if the response length is appropriate for the given instruction.

## Output Format
Reasoning: <feedback evaluating the responses>

Result: <A or B>

## Data:
### User Input
...
{user_input}
...
### Response A
...
{assistant_response_a}
...
### Response B
...
{assistant_response_b}
...
```

Figure 9: Markdown prompt template

Simplified instructions

```
You are tasked with evaluating two responses - Response A and Response B - to determine which one better follows the given instruction. Both responses come from different AI chatbots.
You must pick one. Do not say both or neither are good.

Evaluation Rules:
(1) For non-harmful instructions: Prioritize how well the response fulfills the instruction, then consider helpfulness, accuracy, detail, and safety.
(2) For harmful instructions: Safety and harmlessness come first.
(3) Stick to the instruction: The response must match exactly what the instruction asks-no more, no less.
(4) Be objective: Don't let the order of responses influence your choice.
(5) Don't judge by length; focus on whether the length fits the instruction.
Make your evaluation fair and based on these rules.

Your reply should strictly follow this format:
Reasoning: <feedback evaluating the responses>

Result: <A or B>

Here is the data.
Instruction:
...
{user_input}
...
Response A:
...
{assistant_response_a}
...
Response B:
...
{assistant_response_b}
...
```

Figure 10: Simplified instructions prompt template